

Bachelier, Louis (1870–1946)

Formation Years

Louis Bachelier was born in Le Havre, France, on March 11, 1870. His father, a native of Bordeaux, moved to Le Havre after his marriage to the daughter of a notable citizen of Le Havre. He started a wine and spirits shop, and bought and exported wines from Bordeaux and Champagne. At the time, Le Havre was an important port. The Protestant bourgeoisie in the city, which dominated the local cotton and coffee markets, occupied the upper echelons of society. The young Louis was educated at a high school in Le Havre. He seems to have been a fairly good student, but he interrupted his studies after earning his high school diploma in 1889, when both of his parents died in the span of a few weeks. To provide for his youngest brother and his older sister, most likely, he took over his father's business, but he sold it after a few years. In 1892, he completed his military service as an infantryman and then moved to Paris, where his activities are unclear. What is clear, however, is that Bachelier focused on his interests in the stock market and undertook university studies at the University of Paris, where in 1895 he obtained his bachelor's degree in the mathematical sciences, without being a particularly distinguished student. After earning his degree, he continued to attend the lectures of the Faculty, including courses in mathematical physics taught by Poincaré and Boussinesq.

Although we cannot be absolutely certain, it is likely that in 1894, Bachelier attended lectures in probability theory given by Poincaré, which were published in 1896 and were based on the remarkable treatise that Joseph Bertrand published in 1888. His attendance at these lectures, his reading of treatises by Bertrand and Poincaré, and his interest in the stock market probably inspired his thesis, "theory of speculation", which was defended by Bachelier [1] in Paris on March 29, 1900, before a jury composed of Appell, Boussinesq, and Poincaré. On the report by Henri Poincaré, he was conferred the rank of Doctor of Mathematics with an "honorable" designation, that is, a designation insufficient for him to obtain employment in higher education, which was extremely limited at the time.

Let us say a few words about this extraordinary thesis. The problem investigated by Bachelier is described in less than a page. The stock market is subject to innumerable random influences, and so it is unreasonable to expect a mathematically precise forecast of stock prices. However, we can try to establish the law of the changes in stock prices over a fixed period of time. The determination of this law was the subject of Bachelier's thesis. The thesis was not particularly original. Since the early nineteenth century, people had applied probability theory to study exchange rates. In France, in particular, we can cite the work of Biquilley (around 1800) or Jules Regnault (around 1850). In his thesis, Bachelier [1] intended to revisit this issue from several viewpoints taken from physics and probability theory, as these subjects were taught in Europe, including Paris, around 1900. He adapted these viewpoints to aid his investigation. The first method he used is the method adopted by Einstein, five years later, to determine the law of Brownian motion in a physical context. It consists of studying the integral equation that governs the probability that the change in price is y at time t , under two natural assumptions: the change in price during two separate time intervals is independent and the expectation of the change in price is zero. The resulting equation is a homogeneous version of the diffusion equation, now known as the *Kolmogorov* (or *Chapman–Kolmogorov*) equation, in which Bachelier boldly asserts that the appropriate solution is given by a centered Gaussian law with variance proportional to time t . He proved a statement already proposed, without justification, by Regnault in 1860 that the expectation of the absolute change in price after time t is proportional to the square root of t .

But this first method, which would eventually be used in the 1930s by physicists and probabilists, did not seem to satisfy Bachelier, since he proposed a second method, which was further developed in the 1930s by the Moscow School: the approximation of the law of Brownian motion by an infinite sequence of coin flips, properly normalized. Since the change in price over a given period of time is the result of a very large number of independent random variables, it is not surprising that this change in price is Gaussian. But the extension of this approximation to a continuous-time version is not straightforward. Bachelier, who already knew the result he wanted to obtain, states and prepares the way to the first known version of a theorem, which in the current

language reads as follows: let $\{X_1, X_2, \dots, X_n, \dots\}$ be a sequence of independent random variables taking values 1 or -1 with probability $1/2$. If we let $S_n = X_1 + \dots + X_n$ and let $[x]$ denote the integer part of a real number x , then

$$\left(\frac{1}{\sqrt{n}} S_{[nt]}, t \geq 0 \right) \longrightarrow (B_t, t \geq 0) \quad (1)$$

in law as $n \longrightarrow \infty$, where $(B_t, t \geq 0)$ is a standard Brownian motion.

This second method, which is somewhat difficult to read and not very rigorous, naturally leads to the previous solution. But it is still not sufficient. Bachelier proposes a third method, the “radiation (or diffusion) of probability”. Bachelier, having attended the lectures of Poincaré and Boussinesq on the theory of heat, was aware of the “method of Laplace”, which gives the fundamental solution of the heat equation, a solution that has exactly the form given by the first (and second) methods used by Bachelier. Hence, there is a coincidence to be elucidated. We know that Laplace probably knew the reason for this coincidence. Lord Rayleigh had recently noticed this coincidence in his solution to the problem of “random phases”. It is likely that neither Bachelier nor Poincaré had read the work of Rayleigh. Anyway, Bachelier, in turn, explains this curious intersection between the theory of heat and the prices of annuities on the Paris stock exchange. This is his third method, which can be summarized as follows.

Consider the game of flipping a fair coin an infinite number of times and set $f(n, x) = \mathbb{P}(S_n = x)$. It has been known since at least the seventeenth century that

$$f(n+1, x) = \frac{1}{2}f(n, x-1) + \frac{1}{2}f(n, x+1) \quad (2)$$

Subtracting $f(n, x)$ from both the sides of the equation, we obtain

$$f(n+1, x) - f(n, x) = \frac{1}{2} \left(f(n, x+1) - 2f(n, x) + f(n, x-1) \right) \quad (3)$$

It then suffices to take the unit 1 in the preceding equation to be infinitely small to obtain the heat equation

$$\frac{\partial f}{\partial n} = \frac{1}{2} \frac{\partial^2 f}{\partial x^2} \quad (4)$$

whose solution is the law of a centered Gaussian random variable with variance n .

Theory of Speculation

At the stock market, probability radiates like heat. This “demonstrates” the role of Gaussian laws in problems related to the stock market, as acknowledged by Poincaré himself in his report: “A little reflection shows that the analogy is real and the comparison legitimate. The arguments of Fourier are applicable, with very little change, to this problem that is so different from the problem to which these arguments were originally applied.” And Poincaré regretted that Bachelier did not develop this point further, though this point would be developed in a masterly way by Kolmogorov in a famous article published in 1931 in the *Mathematische Annalen*. In fact, the first and third methods used by Bachelier are intrinsically linked: the Chapman–Kolmogorov equation for any regular Markov process is equivalent to a partial differential equation of parabolic type. In all regular Markovian schemes that are continuous, probability radiates like heat from a fire fanned by the thousand winds of chance. And further work, exploiting this real analogy, would transform not only the theory of Markov processes but also the century-old theory of Fourier equations and parabolic equations.

Now, having determined the law of price changes, all calculations of financial products involving time follow easily. But Bachelier did not stop there. He proposed a general theory of speculation integrating all stock market products that could be proposed to clients, whose (expected) value at maturity—and therefore whose price—can be calculated using general formulas resulting from theory. The most remarkable product that Bachelier priced was based on the maximum value of a stock during the period between its purchase and a maturity date (usually one month later). In this case, one must determine the law of the maximum of a stock price over some interval of time. This problem would be of concern to Norbert Wiener, the inventor of the mathematical theory of Brownian motion, in 1923. It involves knowing *a priori* the law of the price over an infinite time interval, but it was not known—either in 1923 or in 1900—how to easily calculate the integrals of functions of an infinite number of variables. Let us explain the reasoning used by Bachelier [1] as an example of his methods of analysis.

Bachelier proceeded in two different ways. The first way was based on the second method developed in Bachelier's thesis. It consists of discretizing time in steps of Δt , and introducing a change in price at each step of $\pm \Delta x$. Bachelier wanted to calculate the probability that before time $t = n\Delta t$, the game (or price) exceeds a given value $c = m\Delta x$. Let $n = m + 2p$. Bachelier proposed to first calculate the probability that the price c is reached for the first time at exactly time t . To this end, he uses the gambler's ruin argument: the probability is equal to $(m/n)C_n^p 2^{-n}$, which Bachelier obtained from the ballot formula of Bertrand, which he learned from Poincaré or Bertrand's work, or perhaps both. It suffices to then pass properly to the limit so that $\Delta x = \mathcal{O}(\sqrt{\Delta t})$. One then obtains the probability that the price exceeds c before t . Bachelier then noted that this probability is equal to twice the probability that the price exceeds c at time t .

The result is Bachelier's formula for the law of the maximum M_t of the price B_t over the interval $[0, t]$; that is,

$$\mathbb{P}(M_t > c) = 2\mathbb{P}(B_t > c) \quad (5)$$

It would have been difficult to proceed in a simpler fashion. Having obtained this formula, Bachelier had to justify it in a simple way to understand why it holds. Bachelier therefore added to his first calculation (which was somewhat confusing and difficult to follow) a "direct demonstration" without passing to the limit. He used the argument that "the price cannot pass the threshold c over a time interval of length t without having done so previously" and hence that

$$\mathbb{P}(B_t > c) = \mathbb{P}(M_t > c)\alpha \quad (6)$$

where α is the probability that the price c , having been attained before time t , is greater than c at time t . The latter probability is obviously $1/2$, due to symmetry of the sample paths that go above and that remain below c by time t . And Bachelier concludes: "It is remarkable that the multiple integral that expresses the probability $\mathbb{P}(M_t > c)$ does not seem amenable to ordinary methods of calculation, but can be determined by very simple probabilistic reasoning." It was, without doubt, the first example of the use of the reflection principle in probability theory. In two steps, a complicated calculation yields

a simple formula by using a very simple probabilistic (or combinatorial) argument.

Of course, Bachelier had to do his mathematics without a safety net. What could his safety net have been? The mathematical analysis available during his time could not deal with such strange objects and calculations. It was not until the following year, 1901, that Lebesgue introduced the integral based on the measure that Borel had just recently constructed. The Daniell integral, which Wiener used, dates to 1920 and it was not until the 1930s that European mathematicians realized that computing probabilities with respect to Brownian motion, or with respect to sequences of independent random variables, could be done using Lebesgue measure on the unit interval. Since Lebesgue's theory came to be viewed as one of the strongest pillars of analysis in the twentieth century, this approach gave probability theory a very strong analytic basis. We will have to wait much longer to place the stochastic calculus of Brownian motion and sample path arguments involving stopping times into a relatively uniform analytical framework. Anyway, Bachelier had little concern for either this new theory in analysis or the work of his contemporaries, whom he never cites. He refers to the work of Laplace, Bertrand, and Poincaré, who never cared about the Lebesgue integral, and so Bachelier always ignored its existence.

It seems that in 1900, Bachelier [1] saw very clearly how to model the continuous movement of stock prices and he established new computational techniques, derived notably from the classical techniques involving infinite sequences of fair coin flips. He provided an intermediate mathematical argument to explain a new class of functions that reflected the vagaries of the market, just as in the eighteenth century, when one used geometric reasoning and physical intuition to explain things.

After the Thesis

His Ph.D. thesis defended, Bachelier suddenly seemed to discover the immensity of a world in which randomness exists. The theory of the stock market allowed him to view the classical results of probability with a new eye, and it opened new viewpoints for him. Starting in 1901, Bachelier showed that the known results about infinite sequences of fair coin flips could all (or almost all) be obtained from stock

market theory and that one can derive new results that are more precise than anyone had previously suspected. In 1906, Bachelier proposes an almost general theory of “related probabilities”, that is to say, a theory about what would, 30 years later, be called *Markov processes*. This article by Bachelier was the starting point of a major study by Kolmogorov in 1931 that we already mentioned. All of Bachelier’s work was published with the distant but caring recommendation of Poincaré, so that by 1910, Bachelier, whose income remains unknown and was probably modest, is permitted to teach a “free course” in probability theory at the Sorbonne, without compensation. Shortly thereafter, he won a scholarship that allowed him to publish his *Calculus of Probability*, Volume I, Paris, Gauthier-Villars, 1912 (Volume II never appeared), which included all of his work since his thesis. This very surprising book was not widely circulated in France, and had no impact on the Paris stock market or on French mathematics, but it was one of the sources that motivated work in stochastic processes at the Moscow School in the 1930s. It also influenced work by the American School on sums of independent random variables in the 1950s, and at the same time, influenced new theories in mathematical finance that were developing in the United States. And, as things should rightly be, these theories traced back to France, where Bachelier’s name had become so well recognized that in 2000, the centennial anniversary of his work in “theory of speculation” was celebrated.

The First World War interrupted the work of Bachelier, who was summoned for military service in September 1914 as a simple soldier. When the war ended in December 1918, he was a sublieutenant in the Army Service Corps. He served far from the front, but he carried out his service with honor. As a result, in 1919, the Directorate of Higher Education in Paris believed it was necessary to appoint Bachelier to a university outside of Paris, since the war had decimated the ranks of young French mathematicians and there were many positions to be filled. After many difficulties, due to his marginalization in the French mathematical community and the incongruent nature of his research, Bachelier finally received tenure in 1927 (at the age of 57) as a professor at the University of Besançon, where he remained until his retirement in 1937. Throughout the postwar years,

Bachelier essentially did not publish any original work. He married in 1920, but his wife died a few months later. He was often ill and he seems to have been quite isolated.

In 1937, he moved with his sister to Saint-Malo in Brittany. During World War II, he moved to Saint-Servan, where he died in 1946. He seemed to be aware of the new theory of stochastic processes that was then developing in Paris and Moscow, and that was progressively spreading all over the world. He attempted to claim credit for the things that he had done, without any success. He regained his appetite for research, to the point that in 1941, at the age of 70, he submitted a note for publication to the Academy of Sciences in Paris on the “probability of maximum oscillations”, in which he demonstrated a fine mastery of the theory of Brownian motion, which was undertaken systematically by Paul Levy starting in 1938. Paul Levy, the principal French researcher of the theory of Brownian motion, recognized, albeit belatedly, the work of Bachelier, and his work provided a more rigorous foundation for Bachelier’s “theory of speculation”.

Reference

- [1] Bachelier, L. (1900). *Théorie de la spéculation*, Thèse Sciences mathématiques Paris. *Annales Scientifiques de l'Ecole Normale Supérieure* **17**, 21–86; *The Random Character of Stock Market Prices*, P. Cootner, ed, MIT Press, Cambridge, 1964, pp. 17–78.

Further Reading

- Courtault, J.M. & Kabanov, Y. (eds) (2002). *Louis Bachelier: Aux origines de la Finance Mathématique*, Presses Universitaires Franc-Comtoises, Besançon.
- Taqqu, M.S. (2001). Bachelier and his times: a conversation with Bernard Bru, *Finance and Stochastics* **5**(1), 3–32.

Related Articles

Black–Scholes Formula; Markov Processes; Martingales; Option Pricing; General Principles.

BERNARD BRU

Samuelson, Paul A.

Paul Anthony Samuelson (1915–) is Institute Professor Emeritus at the Massachusetts Institute of Technology where he has taught since 1940. He earned a BA from the University of Chicago in 1935 and his PhD in economics from Harvard University in 1941. He received the John Bates Clark Medal in 1947 and the National Medal of Science in 1996. In 1970, he became the first American to receive the Alfred Nobel Memorial Prize in Economic Sciences. His textbook, *Economics*, first published in 1948, and in its 18th edition, is the best-selling and arguably the most influential economics textbook of all time.

Paul Samuelson is the last great general economist—never again will any one person make such foundational contributions to so many distinct areas of economics. His prolific and profound theoretical contributions over seven decades of published research have been universal in scope, and his ramified influence on the whole of economics has led to foundational contributions in virtually every field of economics, including financial economics. Representing 27 years of scientific writing from 1937 to the middle of 1964, the first two volumes of his *Collected Scientific Papers* contain 129 articles and 1772 pages. These were followed by the publication of the 897-page third volume in 1972, which registers the succeeding seven years' product of 78 articles published when he was between the ages of 49 and 56 [18]. A mere five years later, at the age of 61, Samuelson had published another 86 papers, which fill the 944 pages of the fourth volume. A decade later, the fifth volume appeared with 108 articles and 1064 pages. A glance at his list of publications since 1986 assures us that a sixth and even seventh volume could be filled. That Samuelson paid no heed to the myth of debilitating age in science is particularly well-exemplified in his contributions to financial economics, with all but 6 of his more than 60 papers being published after he had reached the age of 50.

Samuelson's contribution to quantitative finance, as with mathematical economics generally, has been foundational and wide-ranging: these include reconciling the axioms of expected utility theory first with nonstochastic theories of choice [9] and then with the ubiquitous and practical mean–variance criterion of choice [16], exploring the foundations of diversification [13] and optimal portfolio selection when facing

fat-tailed, infinite-variance return distributions [14], and, over a span of nearly four decades, analyzing the systematic dependence on age of optimal portfolio strategies, in particular, optimal long-horizon investment strategies, and the improper use of the Law of Large Numbers to arrive at seemingly dominating strategies for the long run [10, 15, 17, 21–27]. In investigating the oft-told tale that investors become systematically more conservative as they get older, Samuelson shows that perfectly rational risk-averse investors with constant relative risk aversion will select the same fraction of risky stocks *versus* safe cash period by period, independently of age, provided that the investment opportunity set is unchanging. Having shown that greater investment conservatism is not an inevitable consequence of aging, he later [24] demonstrates conditions under which such behavior can be optimal: with mean-reverting changing opportunity sets, older investors will indeed be more conservative than in their younger days, provided that they are more risk averse than a growth-optimum, log-utility maximizer. To complete the rich set of age-dependent risk-taking behaviors, Samuelson shows that rational investors may actually become less conservative with age, if either they are less risk averse than log or if the opportunity set follows a trending, momentum-like dynamic process. He recently confided that in finance, this analysis is a favorite brainchild of his.

Published in the same issue of the *Industrial Management Review*, “Proof That Properly Anticipated Prices Fluctuate Randomly” and “Rational Theory of Warrant Pricing” are perhaps the two most influential Samuelson papers in quantitative finance. During the decade before their printed publication in 1965, Samuelson had set down, in an unpublished manuscript, many of the results in these papers and had communicated them in lectures at MIT, Yale, Carnegie, the American Philosophical Society, and elsewhere. In the early 1950s, he supervised a PhD thesis on put and call pricing [5].

The sociologist or historian of science would undoubtedly be able to develop a rich case study of alternative paths for circulating scientific ideas by exploring the impact of this oral publication of research in rational expectations, efficient markets, geometric Brownian motion, and warrant pricing in the period between 1956 and 1965.

Samuelson (1965a) and Eugene Fama independently provide the foundation of the Efficient Market

theory that developed into one of the most important concepts in modern financial economics. As indicated by its title, the principal conclusion of the paper is that in well-informed and competitive speculative markets, the intertemporal changes in prices will be essentially random. Samuelson has described the reaction (presumably his own as well as that of others) to this conclusion as one of “initial shock—and then, upon reflection, that it is obvious”. The argument is as follows: the time series of changes in most economic variables gross national product (GNP, inflation, unemployment, earnings, and even the weather) exhibit cyclical or serial dependencies. Furthermore, in a rational and well-informed capital market, it is reasonable to presume that the prices of common stocks, bonds, and commodity futures depend upon such economic variables. Thus, the shock comes from the seemingly inconsistent conclusion that in such well-functioning markets the changes in speculative prices should exhibit no serial dependencies. However, once the problem is viewed from the perspective offered in the paper, this seeming inconsistency disappears and all becomes obvious.

Starting from the consideration that in a competitive market, if everyone *knew* that a speculative security was expected to rise in price by more (less) than the required or fair expected rate of return, it would *already* be bid up (down) to negate that possibility, Samuelson postulates that securities will be priced at each point in time so as to yield this fair expected rate of return. Using a backward-in-time induction argument, he proves that the changes in speculative prices around that fair return will form a *martingale*. And this follows no matter how much serial dependency there is in the underlying economic variables upon which such speculative prices are formed. In an informed market, therefore, current speculative prices will already reflect anticipated or forecastable future changes in the underlying economic variables that are relevant to the formation of prices, and this leaves only the unanticipated or unforecastable changes in these variables as the sole source of fluctuations in speculative prices.

Samuelson is careful to warn the reader against interpreting his mathematically derived theoretical conclusions about markets as empirical statements. Nevertheless, for 40 years, his model has been important to the understanding and interpretation of the empirical results observed in real-world markets. For

the most part in those ensuing years, his interpretation of the data is that organized markets where widely owned securities are traded are well approximated as *microefficient*, meaning that the relative pricing of individual securities within the same or very similar asset classes is such that active asset management applied to those similar securities (e.g., individual stock selection) does not earn greater risk-adjusted returns.

However, Samuelson is discriminating in his assessment of the efficient market hypothesis as it relates to real-world markets. He notes a list of the “few not-very-significant apparent exceptions” to microefficient markets [23, p. 5]. He also expresses belief that there are exceptionally talented people who can probably garner superior risk-corrected returns, and even names a few. He does not see them as offering a practical broad alternative investment prescription for active management since such talents are few and hard to identify. As Samuelson believes strongly in microinefficiency of the markets, he expresses doubt about *macromarket* efficiency: namely that indeed asset-value “bubbles” do occur.

There is no doubt that the mainstream of the professional investment community has moved significantly in the direction of Paul Samuelson’s position during the 35 years since he issued his challenge to that community to demonstrate widespread superior performance [20]. Indexing as either a core investment strategy or a significant component of institutional portfolios is ubiquitous, and even among those institutional investors who believe they can deliver superior performance, performance is typically measured incrementally relative to an index benchmark and the expected performance increment to the benchmark is generally small compared to the expected return on the benchmark itself. It is therefore with no little irony that as investment practice has moved in this direction, for the last 15 years, academic research has moved in the opposite direction, strongly questioning even the microinefficiency case for the efficient market hypothesis. The conceptual basis of these challenges comes from theories of asymmetric information and institutional rigidities that limit the arbitrage mechanisms that enforce microinefficiency and of cognitive dissonance and other systematic behavioral dysfunctions among individual investors that purport to distort market prices away from rationally determined asset prices in identified ways. A substantial quantity of empirical

evidence has been assembled, but there is considerable controversy over whether it does indeed make a strong case to reject market microefficiency in the Samuelsonian sense. What is not controversial at all is that Paul Samuelson's efficient market hypothesis has had a deep and profound influence on finance research and practice for more than 40 years and all indications are that it will continue to do so well into the future.

If one were to describe the 1960s as "the decade of capital asset pricing and market efficiency" in view of the important research gains in quantitative finance during then, one need hardly say more than "the Black-Scholes option pricing model" to justify describing the 1970s as "the decade of option and derivative security pricing." Samuelson was ahead of the field in recognizing the arcane topic of option pricing as a rich area for problem choice and solution. By at least the early 1950s, Samuelson had shown that the assumption of an absolute random walk or arithmetic Brownian motion for stock price changes leads to absurd prices for long-lived options, and this was done before his rediscovery of Bachelier's pioneering work [1] in which this very assumption is made. He introduced the alternative process of a "geometric" Brownian motion in which the log of price changes follows a Brownian motion, possibly with a drift. His paper on the rational theory of warrant pricing [12] resolves a number of apparent paradoxes that had plagued the existing mathematical theory of option pricing from the time of Bachelier. In the process (with the aid of a mathematical appendix provided by H. P. McKean, Jr), Samuelson also derives much of what has become the basic mathematical structure of option pricing theory today.

Bachelier [1] considered options that could only be exercised on the expiration date. In modern times, the standard terms for options and warrants permit the option holder to exercise on or before the expiration date. Samuelson coined the terms *European* option to refer to the former and *American* option to refer to the latter. As he tells the story, to get a practitioner's perspective in preparation for his research, he went to New York to meet with a well-known put and call dealer (there were no traded options exchanges until 1973) who happened to be Swiss. Upon his identifying himself and explaining what he had in mind, Samuelson was quickly told, "You are wasting your time—it takes a *European* mind to understand options." Later on, when writing

his paper, Samuelson thus chose the term *European* for the relatively simple(-minded)-to-value option contract that can only be exercised at expiration and *American* for the considerably more(-complex)-to-value option contract that could be exercised early, any time on or before its expiration date.

Although real-world options are almost always of the American type, published analyses of option pricing prior to his 1965 paper focused exclusively on the evaluation of European options and therefore did not include the extra value to the option from the right to exercise early.

The most striking comparison to make between the Black-Scholes option pricing theory and Samuelson's *rational theory* [12] is the formula for the option price. The Samuelson partial differential equation for the option price is the same as the corresponding equation for the Black-Scholes option price *if* one sets the Samuelson parameter for the expected return on the underlying stock equal to the riskless interest rate minus the dividend yield and sets the Samuelson parameter for the expected return on the option equal to the riskless interest rate. It should, however, be underscored that the mathematical equivalence between the two formulas with the redefinition of parameters is purely a formal one. The Samuelson model simply posits the expected returns for the stock and option. By employing a dynamic hedging or replicating portfolio strategy, the Black-Scholes analysis derives the option price without the need to know either the expected return on the stock or the required expected return on the option. Therefore, the fact that the Black-Scholes option price satisfies the Samuelson formula implies neither that the expected returns on the stock and option are equal nor that they are equal to the riskless rate of interest. Furthermore, it should also be noted that Black-Scholes pricing of options does not require knowledge of investors' preferences and endowments as is required, for example, in the sequel Samuelson and Merton [28] warrant pricing paper. The "rational theory" put forward in 1965 is thus clearly a "miss" with respect to the Black-Scholes development. However, as this analysis shows, it is just as clearly a "near miss". See [6, 19] for a formal comparison of the two models.

Extensive reviews of Paul Samuelson's remarkable set of contributions to quantitative finance can be found in [2–4, 7, 8].

References

- [1] Bachelier, L. (1900, 1966). Theory de la Speculation, Gauthier-Villars, Paris, in *The Random Character of Stock Market Prices*, P. Cootner, ed, MIT Press, Cambridge.
- [2] Bernstein, P.L. (2005). *Capital Ideas: The Improbable Origins of Modern Wall Street*, John Wiley & Sons, Hoboken.
- [3] Carr, P. (2008). The father of financial engineering, *Bloomberg Markets* **17**, 172–176.
- [4] Fischer, S. (1987). Samuelson, Paul Anthony, *The New Palgrave: A Dictionary of Economics*, MacMillan Publishing, Vol. 4, pp. 234–241.
- [5] Krueger, R. (1956). *Put and Call Options: A Theoretical and Market Analysis*, Doctoral dissertation, MIT, Cambridge, MA.
- [6] Merton, R.C. (1972). Continuous-time speculative processes: appendix to P. A. Samuelson's 'mathematics of speculative price', in *Mathematical Topics in Economic Theory and Computation*, R.H., Day & S.M. Robinson, eds, Philadelphia Society for Industrial and Applied Mathematics, pp. 1–42, reprinted in *SIAM Review* 15, 1973.
- [7] Merton, R.C. (1983). Financial economics, in *Paul Samuelson and Modern Economic Theory*, E.C. Brown & R.M. Solow, eds, McGraw Hill, New York.
- [8] Merton, R.C. (2006). Paul Samuelson and financial economics, in *Samuelsonian Economics and the Twenty-First Century*, M. Szenberg, L. Ramrattan & A. Gottesman, Oxford University Press, Oxford, Reprinted in *American Economist* 50, no. 2 (Fall 2006).
- [9] Samuelson, P.A. (1952). Probability, utility, and the independence axiom, *Econometrica* **20**, 670–678, Collected Scientific Papers, I, Chap. 14.
- [10] Samuelson, P.A. (1963). Risk and uncertainty: a fallacy of large numbers, *Scientia* **57**, 1–6, Collected Scientific Papers, I, Chap. 16.
- [11] Samuelson, P.A. (1965). Proof that properly anticipated prices fluctuate randomly, *Industrial Management Review* **6**, 41–49, Collected Scientific Papers, III, Chap. 198.
- [12] Samuelson, P.A. (1965). Rational theory of warrant pricing, *Industrial Management Review* **6**, 13–39, Collected Scientific Papers, III, Chap. 199.
- [13] Samuelson, P.A. (1967). General proof that diversification pays, *Journal of Financial and Quantitative Analysis* **2**, 1–13, Collected Scientific Papers, III, Chap. 201.
- [14] Samuelson, P.A. (1967). Efficient portfolio selection for Pareto-Levy investments, *Journal of Financial and Quantitative Analysis* **2**, 107–122, Collected Scientific Papers, III, Chap. 202.
- [15] Samuelson, P.A. (1969). Lifetime portfolio selection by dynamic stochastic programming, *Review of Economics and Statistics* **51**, 239–246, Collected Scientific Papers, III, Chap. 204.
- [16] Samuelson, P.A. (1970). The fundamental approximation theorem of portfolio analysis in terms of means, variances and higher moments, *Review of Economic Studies* **37**, 537–542, Collected Scientific Papers, III, Chap. 203.
- [17] Samuelson, P.A. (1971b). The 'Fallacy' of maximizing the geometric mean in long sequences of investing or gambling, *Proceedings of the National Academy of Sciences of United States of America* **68**, 2493–2496, Collected Scientific Papers, III, Chap. 207.
- [18] Samuelson, P.A. (1972). *The Collected Scientific Papers of Paul A. Samuelson*, R.C. Merton, ed, MIT Press, Cambridge, Vol. 3.
- [19] Samuelson, P.A. (1972). Mathematics of speculative price, in *Mathematical Topics in Economic Theory and Computation*, R.H. Day & S.M. Robinson, eds, Society for Industrial and Applied Mathematics, Philadelphia, pp. 1–42, reprinted in *SIAM Review* 15, 1973, Collected Scientific Papers, IV, Chap. 240.
- [20] Samuelson, P.A. (1974). Challenge to judgment, *Journal of Portfolio Management* **1**, 17–19, Collected Scientific Papers, IV, Chap. 243.
- [21] Samuelson, P.A. (1979). Why we should not make mean log of wealth big though years to act are long, *Journal of Banking and Finance* **3**, 305–307.
- [22] Samuelson, P.A. (1989). A case at last for age-phased reduction in equity, *Proceedings of the National Academy of Science of United States of America* **86**, 9048–9051.
- [23] Samuelson, P.A. (1989). The judgment of economic science on rational portfolio management: indexing, timing, and long-horizon effects, *Journal of Portfolio Management* Fall, **16**, 4–12.
- [24] Samuelson, P.A. (1991). Long-run risk tolerance when equity returns are mean regressing pseudoparadoxes and vindication of 'businessmen's risk', in *Money, Macroeconomics, and Economic Policy: Essays in Honor of James Tobin*, W.C. Brainard, W.D. Nordhaus & H.W. Watts, eds, The MIT Press, Cambridge, pp. 181–200.
- [25] Samuelson, P.A. (1992). At last a rational case for long horizon risk tolerance and for asset-allocation timing? in *Active Asset Allocation*, D.A. Robert & F.J. Fabozzi, eds, Probus Publishing, Chicago.
- [26] Samuelson, P.A. (1994). The long-term case of equities and how it can be oversold, *Journal of Portfolio Management* Fall, **21**, 15–24.
- [27] Samuelson, P.A. (1997). Proof by certainty equivalents that diversification-across-time does worse, risk-corrected, than diversification-throughout-time, *Journal of Risk and Uncertainty* **14**, 129–142.
- [28] Samuelson, P.A. & Merton, R.C. (1969). A complete model of warrant pricing that maximizes utility, *Industrial Management Review* **10**, 17–46, Collected Scientific Papers, III, Chap. 2000.

Further Reading

- Samuelson, P.A. (1966). *The Collected Scientific Papers of Paul A. Samuelson*, J.E. Stiglitz, ed, MIT Press, Cambridge, Vols. 1 and 2.
- Samuelson, P.A. (1971). Stochastic speculative price, *Proceedings of the National Academy of Sciences of the United States of America* **68**, 335–337, Collected Scientific Papers, III, Chap. 206.

- Samuelson, P.A. (1977). *The Collected Scientific Papers of Paul A. Samuelson*, H. Nagatani & K. Crowley, eds, MIT Press, Cambridge, Vol. 4.
- Samuelson, P.A. (1986). *The Collected Scientific Papers of Paul A. Samuelson*, K. Crowley, ed, MIT Press, Cambridge, Vol. 5.

ROBERT C. MERTON

Black, Fischer

The central focus of the career of Fischer Black (1938–1995) was on teasing out the implications of the capital asset pricing model (CAPM) for the changing institutional framework of financial markets of his day. He became famous for the Black–Scholes options formula [14], an achievement that is now widely recognized as having opened the door to modern quantitative finance and financial engineering. Fischer was the first quant, but a very special kind of quant because of his taste for the big picture [16].

Regarding that big picture, as early as 1970, he sketched a vision of the future that has by now largely come true:

Thus a long term corporate bond could actually be sold to three separate persons. One would supply the money for the bond; one would bear the interest rate risk; and one would bear the risk of default. The last two would not have to put up any capital for the bonds, although they might have to post some sort of collateral.

Today we recognize the last two instruments as an interest rate swap and a credit default swap, the two instruments that have been the central focus of financial engineering ever since.

All of the technology involved in this engineering can be traced back to roots in the original Black–Scholes option pricing formula [14]. Black himself came up with a formula through CAPM, by thinking about the exposure to systemic risk that was involved in an option, and how that exposure changes as the price of the underlying changes. Today the formula is more commonly derived using the Ito formula and the option replication idea introduced by Merton [17]. For a long time, Black himself was unsure about the social utility of equity options. If all they do is to allow people to achieve the same risk exposure they could achieve by holding equity outright with leverage, then what is the point?

The Black–Scholes formula and the hedging methodology behind it subsequently became a central pillar in the pricing of contingent claims of all kinds and in doing so gave rise to many innovations that contributed to making the world more like his 1970 vision. Black and Cox [9] represents an early attempt to use the option pricing technology to price default risk. Black [4] similarly uses the option pricing technology to price currency risk. Perhaps, Black's most

important use of the tools was in his work on interest rate derivatives, in the famous Black–Derman–Toy term structure model [10].

Black got his start in finance after already earning his PhD in applied mathematics (Harvard, 1964) when he learned about CAPM from Treynor [18], his colleague at the business consulting firm Arthur D. Little, Inc. Fischer had never taken a single course in economics or finance, nor did he ever do so subsequently. Nevertheless, the field was underdeveloped at the time, and Fischer managed to set himself up as a financial consultant and to parlay his success in that capacity into a career in academia (University of Chicago 1971–1975, Massachusetts Institute of Technology 1975–1984), and then into a partnership at the Wall Street investment firm of Goldman Sachs (1984–1995). There can be no doubt that his early success with the options pricing formula opened these doors. The more important point is how, in each of these settings, Fischer used the opportunity he had been given to help promote his vision of a CAPM future for the financial side of the economy.

CAPM is only about a world of debt and equity, and the debt in that world is both short term and risk free. In such a world, everyone holds the fully diversified market portfolio of equity and then adjusts risk exposure by borrowing or lending in the market for risk-free debt. As equity values fluctuate, outstanding debt also fluctuates, as people adjust their portfolios to maintain desired risk exposure. One implication of CAPM, therefore, is that there should be a market for passively managed index mutual funds [15]. Another implication is that the regulatory apparatus surrounding banking, both lending and deposit taking, should be drastically relaxed to facilitate dynamic adjustment of risk exposure [3]. And yet a third implication is that there might be a role for an automatic risk rebalancing instrument, essentially what is known today as *portfolio insurance* [6, 13].

Even while Black was working on remaking the world in the image of CAPM, he was also expanding the image of the original CAPM to include a world without a riskless asset in his famous zero-beta model [1] and to include a world with multiple currencies in his controversial universal hedging model [2, 7] that subsequently formed the analytical core of the Black–Litterman model of global asset allocation [11, 12].

These and other contributions to quantitative finance made Fischer Black famous, but according

to him, his most important work was the two books he wrote that extended the image of CAPM to the real economy, including the theory of money and business cycles [5, 8]. The fluctuation of aggregate output, he reasoned, was nothing more than the fluctuating yield on the national stock of capital. Just as risk is the price we pay for higher expected yield, business fluctuation is also the price we pay for higher expected rates of economic growth.

The rise of modern finance in the last third of twentieth century transformed the financial infrastructure within which businesses and households interact. A system of banking institutions was replaced by a system of capital markets, as financial engineering developed ways to turn loans into bonds. This revolution in institutions has also brought with it a revolution in our thinking about how the economy works, including the role of government regulation and stabilization policy. Crises in the old banking system gave rise to the old macroeconomics. Crises in the new capital markets system will give rise to a new macroeconomics, possibly built on the foundations laid by Fischer Black.

References

- [1] Black, F. (1972). Capital market equilibrium with restricted borrowing, *Journal of Business* **45**, 444–455.
- [2] Black, F. (1974). International capital market equilibrium with investment barriers, *Journal of Financial Economics* **1**, 337–352.
- [3] Black, F. (1975). Bank funds management in an efficient market, *Journal of Financial Economics* **2**, 323–339.
- [4] Black, F. (1976). The pricing of commodity contracts, *Journal of Financial Economics* **3**, 167–179.
- [5] Black, F. (1987). *Business Cycles and Equilibrium*, Basil Blackwell, Cambridge, MA.
- [6] Black, F. (1988). Individual investment and consumption under uncertainty, in *Portfolio Insurance, A Guide to Dynamic Hedging*, D.L. Luskin, ed, John Wiley & Sons, New York, pp. 207–225.
- [7] Black, F. (1990). Equilibrium exchange rate hedging, *Journal of Finance* **45**, 899–907.
- [8] Black, F. (1995). *Exploring General Equilibrium*, MIT Press, Cambridge, MA.
- [9] Black, F. & Cox, J.C. (1976). Valuing corporate securities: some effects of bond indenture provisions, *Journal of Finance* **31**, 351–368.
- [10] Black, F., Derman, E. & Toy, W.T. (1990). A one-factor model of interest rates and its application to treasury bond options, *Financial Analysts Journal* **46**, 33–39.
- [11] Black, F. & Litterman, R. (1991). Asset allocation: combining investor views with market equilibrium, *Journal of Fixed Income* **1**, 7–18.
- [12] Black, F. & Litterman, R. (1992). Global portfolio optimization, *Financial Analysts Journal* **48**, 28–43.
- [13] Black, F. & Perold, A.F. (1992). Theory of constant proportion portfolio insurance, *Journal of Economic Dynamics and Control* **16**, 403–426.
- [14] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.
- [15] Black, F. & Scholes, M. (1974). From theory to a new financial product, *Journal of Finance* **19**, 399–412.
- [16] Mehrling, P.G. (2005). *Fischer Black and the Revolutionary Idea of Finance*, John Wiley & Sons, Hoboken, New Jersey.
- [17] Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.
- [18] Treynor, J.L. (1962). Toward a theory of market value of risky assets, in *Asset Pricing and Portfolio Performance*, R.A. Korajczyk, ed, Risk Books, London, pp. 15–22.

Related Articles

Black–Scholes Formula; Black–Litterman Approach; Option Pricing Theory: Historical Perspectives; Merton, Robert C.; Modern Portfolio Theory; Term Structure Models; Sharpe, William F.

PERRY MEHRLING

Mandelbrot, Benoit



Benoit B. Mandelbrot, Sterling Professor Emeritus of Mathematical Sciences at Yale University and IBM Fellow Emeritus at the IBM Research Center, best known as the “father of fractal geometry”, is a Polish-born French-American multidisciplinary scientist with numerous contributions to different fields of knowledge including mathematics, statistics, hydrology, physics, engineering, physiology, economics and, last but not least, quantitative finance. In this short text we will focus on Mandelbrot’s contributions to the study of financial markets.

Benoit Mandelbrot was born in Warsaw, Poland, on November 20, 1924 in a family of scholars from Lithuania. In 1936 Mandelbrot’s family moved to Paris, where he was influenced by his mathematician uncle Szolem Mandelbrojt (1899–1983). He entered the Ecole Polytechnique in 1944. Among his professors at Polytechnique was Paul Levy, whose pioneering work on stochastic processes influenced Mandelbrot.

After two years in Caltech and after obtaining a doctoral degree in mathematics from University of Paris in 1952, he started his scientific career at the Centre National de la Recherche Scientifique in Paris, before moving on various scientific appointments which included those at Ecole Polytechnique, Université de Lille, the University of Geneva MIT, Princeton, University of Chicago, and finally the IBM Thomas J. Watson Research Center in Yorktown Heights, New York and Yale University where he spent the longer part of his career.

A central thread in his scientific career is the “ardent pursuit of the concept of roughness” which resulted in a rich theoretical apparatus—fractal and multifractal geometry—whose aim is to describe and represent the order hidden in apparently wildly

disordered and random phenomena ranging from the geometry of coastlines to the variation of foreign exchange rates. In his own words

The roughness of clusters in the physics of disorder, of turbulent flows, of exotic noises, of chaotic dynamical systems, of the distribution of galaxies, of coastlines, of stock price charts, and of mathematical constructions,—these have typified the topics I studied.

He formalized the notion of ‘fractal process’—and later, that of multifractal [13]—which provided a tool for quantifying the “degree of irregularity” of various random phenomena in mathematics, physics, and economics.

Benoit Mandelbrot’s numerous awards include the 1993 *Wolf Prize for Physics* and the 2003 *Japan Prize for Science and Technology*, the 1985 *F. Barnard Medal for Meritorious Service to Science* (“*Magna est Veritas*”) of the US National Academy of Sciences, the 1986 *Franklin Medal for Signal and Emergent Service in Science* of the Franklin Institute of Philadelphia, the 1988 *Charles Proteus Steinmetz Medal* of IEEE, the 2004 *Prize of Financial Times/Deutschland*, and a *Humboldt Preis* from the Alexander von Humboldt Stiftung.

From Mild to Wild Randomness: The Noah Effect

Mandelbrot developed an early interest in the stochastic modeling of financial markets. Familiar with the work of Louis Bachelier (see **Bachelier, Louis (1870–1946)**), Mandelbrot published a series of pioneering studies [6–8, 21] on the tail behavior of the distribution of price variations, where he advocated the use of heavy-tailed distributions and scale-invariant Lévy processes for modeling price fluctuations. The discovery of the heavy-tailed nature of price movements led him to coin the term “wild randomness” for describing market behavior, as opposed to the “mild randomness” represented by Bachelier’s Brownian model, which later became the standard approach embodied in the Black–Scholes model. Mandelbrot likened the sudden bursts of volatility in financial markets to the “Noah effect”, by analogy with the flood which destroys the world in Noah’s biblical story:

In science, all important ideas need names and stories to fix them in the memory. It occurred to

me that the market's first wild trait, abrupt change or discontinuity, is prefigured in the tale of Noah. As Genesis relates, in Noah's six-hundredth year God ordered the Great Flood to purify a wicked world. [...] The flood came and went, catastrophic but transient. Market crashes are like that : at times, even a great bank or brokerage house can seem like a little boat in a big storm.

Long-range Dependence: The Joseph Effect

Another early insight of Mandelbrot's studies of financial and economic data was the presence of long-range dependence [9–11] in market fluctuations:

The market's second wild trait—almost cycles—is prefigured in the story of Joseph. The Pharaoh dreamed that seven fat cattle were feeding in the meadows, when seven lean kine rose out of the Nile and ate them. [...] Joseph, a Hebrew slave, called the dreams prophetic : Seven years of famine would follow seven years of prosperity. [...] Of course, this is not a regular or predictable pattern. But the appearance of one is strong. Behind it is the influence of long-range dependence in an otherwise random process or, put another way, a long-term memory through which the past continues to influence the random fluctuations of the present. I called these two distinct forms of wild behavior the Noah effect and the Joseph effect. They are two aspects of one reality.

Building on his earlier work Mandelbrot [22, 23] on long-range dependence in hydrology and fractional Brownian motion, he proposed the use of fractional processes for modeling long-range dependence and scaling properties of economic quantities (*see Long Range Dependence*).

Multifractal Models and Stochastic Time Changes

In a series of papers [2, 4, 20] with Adlai Fisher and Laurent Calvet, Mandelbrot studied the scaling properties of the US/DEM foreign exchange rate at frequencies ranging from a few minutes to weeks and, building on earlier work by Clark [3] and Mandelbrot [12, 13], introduced a new family of stochastic models, where the (log) price of an asset is represented by a time-changed fractional Brownian motion, where the time change, representing market

activity, is given by a multifractal (*see Multifractals*) increasing process (*see Mixture of Distribution Hypothesis; Time Change*) [5, 15]:

The key step is to introduce an auxiliary quantity called *trading time*. The term is self-explanatory and embodies two observations. While price changes over fixed clock time intervals are long-tailed, price changes between successive transactions stay near-Gaussian over sometimes long period between discontinuities. Following variations in the trading volume, the time interval between successive transactions vary greatly. This suggests that trading time is related to volume.

The topic of multifractal modeling in finance was further developed in [1, 17–19]; a nontechnical account is given in [16].

Mandelbrot's work in quantitative finance has been generally 20 years ahead of its time: many of his ideas proposed in the 1960s—such as long-range dependence, volatility clustering, and heavy tails—became mainstream in financial modeling in the 1990s. If this is anything of a pattern, his more recent work in the field might deserve a closer look. Perhaps, one of the most important insights of his work on financial modeling is to *closely examine* the empirical features of data before axiomatizing and writing down complex equations, a timeless piece of advice which can be a useful guide for quantitative modeling in finance.

Mandelbrot's work in finance is summarized in the books [14, 15] and a popular account of this work is given in the book [5].

References

- [1] Barral, J. & Mandelbrot, B. (2002). Multifractal products of cylindrical pulses, *Probability Theory and Related Fields* **124**, 409–430.
- [2] Calvet, L., Fisher, A. & Mandelbrot, B. (1997). *Large Deviations and the Distribution of Price Changes*. Cowles Foundation Discussion Papers: 1165.
- [3] Clark, P.K. (1973). A subordinated stochastic process model with finite variance for speculative prices, *Econometrica* **41**(1), 135–155.
- [4] Fisher, A., Calvet, L.M. & Mandelbrot, B. (1997). *Multifractality of the Deutschmark/US Dollar exchange rates*. Cowles Foundation Discussion Papers: 1166.
- [5] Hudson, R.L. (2004). *The (Mis)behavior of Prices: A Fractal View of Risk, Ruin, and Reward*, Basic Books, New York, & Profile Books, London, pp. xxvi + 329.
- [6] Mandelbrot, B. (1962). Sur certains prix spéculatifs: faits empiriques et modèle basé sur les processus stables

- additifs de Paul Lévy, *Comptes Rendus (Paris)* **254**, 3968–3970.
- [7] Mandelbrot, B. (1963). The variation of certain speculative prices, *The Journal of Business of the University of Chicago* **36**, 394–419.
- [8] Mandelbrot, B. (1963). New methods in statistical economics, *The Journal of Political Economy* **71**, 421–440.
- [9] Mandelbrot, B. (1971). Analysis of long-run dependence in economics: the *R/S* technique, *Econometrica* **39**, (July Supplement), 68–69.
- [10] Mandelbrot, B. (1971). When can price be arbitrated efficiently? A limit to the validity of the random-walk and martingale models, *Review of Economics and Statistics* **53**, 225–236.
- [11] Mandelbrot, B. (1972). Statistical methodology for non-periodic cycles: from the covariance to *R/S* analysis, *Annals of Economic and Social Measurement* **1**, 257–288.
- [12] Mandelbrot, B. (1973). Comments on “A subordinated stochastic process model with finite variance for speculative prices.” by Peter K. Clark, *Econometrica* **41**, 157–160.
- [13] Mandelbrot, B. (1974). Intermittent turbulence in self-similar cascades; divergence of high moments and dimension of the carrier, *Journal of Fluid Mechanics* **62**, 331–358.
- [14] Mandelbrot, B. (1997). *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*, Springer, New York, pp. x + 551.
- [15] Mandelbrot, B. (1997). *Fractales, hasard et finance (1959–1997)*, Flammarion (Collection Champs), Paris, p. 246.
- [16] Mandelbrot, B. (1999). *A Multifractal Walk down Wall Street*, *Scientific American*, February 1999, pp. 50–53.
- [17] Mandelbrot, B. (2001). Scaling in financial prices, I: tails and dependence, *Quantitative Finance* **1**, 113–123.
- [18] Mandelbrot, B. (2001). Scaling in financial prices, IV: multifractal concentration, *Quantitative Finance* **1**, 641–649.
- [19] Mandelbrot, B. (2001). Stochastic volatility, power-laws and long memory, *Quantitative Finance* **1**, 558–559.
- [20] Mandelbrot B., Fisher A. & Calvet, L. (1997). *The Multifractal Model of Asset Returns*. Cowles Foundation Discussion Papers: 1164.
- [21] Mandelbrot, B. & Taylor, H.M. (1967). On the distribution of stock price differences, *Operations Research* **15**, 1057–1062.
- [22] Mandelbrot, B. & Van Ness, J.W. (1968). Fractional Brownian motions, fractional noises and applications, *SIAM Review* **10**, 422–437.
- [23] Mandelbrot, B. & Wallis, J.R. (1968). Noah, Joseph and operational hydrology, *Water Resources Research* **4**, 909–918.

Further Reading

- Mandelbrot, B. (1966). Forecasts of future prices, unbiased markets and “martingale” models, *The Journal of Business of the University of Chicago* **39**, 242–255.
- Mandelbrot, B. (1982). *The Fractal Geometry of Nature*.
- Mandelbrot, B. (2003). Heavy tails in finance for independent or multifractal price increments, in *Handbook on Heavy Tailed Distributions in Finance*, T.R. Svetlozar, ed., Handbooks in Finance, 30, Elsevier, pp. 1–34, Vol. 1.

Related Articles

Exponential Lévy Models; Fractional Brownian Motion; Heavy Tails; Lévy Processes; Long Range Dependence; Mixture of Distribution Hypothesis; Stylized Properties of Asset Returns.

RAMA CONT

Sharpe, William F.

William Forsyth Sharpe (born on June 16, 1934) is one of the leading contributors to financial economics and shared the Nobel Memorial Prize in Economic Sciences in 1990 with **Harry Markowitz** and Merton Miller. His most important contribution is the **capital asset pricing model** (CAPM), which provided an equilibrium-based relationship between the expected return on an asset and its risk as measured by its covariance with market portfolio. Similar ideas were developed by John Lintner, Jack Treynor (*see Treynor, Lawrence Jack*), and Jan Mossin around the same time. Sharpe has made other important contributions to the field of financial economics but, given the space limitations, we only describe two of his contributions: the CAPM and the Sharpe ratio.

It is instructive to trace the approach used by Sharpe in developing the CAPM. His starting point was Markowitz's model of portfolio selection, which showed how rational investors would select optimal portfolios. If investors only care about the expected return and the variance of their portfolios, then the optimal weights can be obtained by quadratic programming. The inputs to the optimization are the expected returns on the individual securities and their covariance matrix. In 1963, Sharpe [1] showed how to simplify the computations required under the Markowitz approach. He assumed that each security's return was generated by two random factors: one common to all securities and a second factor that was uncorrelated across securities. This assumption leads to a simple diagonal covariance matrix. Although the initial motivation for this simplifying assumption was to reduce the computational time, it would turn out to have deep economic significance.

These economic ideas were developed in Sharpe's [2] *Journal of Finance* paper. He assumed that all investors would select mean-variance-efficient portfolios. He also assumed that investors had homogeneous beliefs and that investors could borrow and lend at the same riskless rate. As Tobin had shown, this implied two fund separations where the investor would divide his money between the risk-free asset and an efficient portfolio of risky assets. Sharpe highlighted the importance of the notion of equilibrium in this context. This efficient portfolio of risky assets in equilibrium can be identified with the

market portfolio. Sharpe's next step was to derive a relationship between the expected return on any risky asset and the expected return on the market. As a matter of curiosity, the CAPM relationship does not appear in the body of the paper but rather as the final equation in footnote 23 on page 438.

The CAPM relationship in modern notation is

$$E[R_j] - r_f = \beta_j (E[R_m] - r_f) \quad (1)$$

where R_j is the return on security j , R_m is the return on the market portfolio of all risky assets, r_f is the return on the risk-free security, and

$$\beta_j = \frac{\text{Cov}(R_j, R_m)}{\text{Var}(R_m)} \quad (2)$$

is the *beta* of security j . The CAPM asserts that the excess expected return on a risky security is equal to the security's beta times the excess expected return on the market. Note that this is a single period model and that it is formulated in terms of *ex ante* expectations. Note also that formula (2) provides an explicit expression for the risk of a security in terms of its covariance with the market and the variance with the market.

The CAPM has become widely used in both investment finance and corporate finance. It can be used as a tool in portfolio selection and also in the measurement of investment performance of portfolio managers. The CAPM is also useful in capital budgeting applications since it gives a formula for the required expected return on an investment. For this reason, the CAPM is often used in rate hearings in some jurisdictions for regulated entities such as utility companies or insurance companies.

The insights from the CAPM also played an important role in subsequent theoretical advances, but owing to space constraint we only mention one. The original derivation of the classic Black–Scholes option formula was based on the CAPM. Black assumed that the return on the stock and the return on its associated warrant both obeyed the CAPM. Hence he was able to obtain expressions for the expected return on both of these securities and he used this in deriving the Black–Scholes equation for the warrant price.

The second contribution that we discuss is the Sharpe ratio. In the case of a portfolio p with expected return $E[R_p]$ and standard deviation σ_p , the

Sharpe ratio is

$$\frac{E[R_p] - r_f}{\sigma_p} \quad (3)$$

Sharpe [3] introduced this formula in 1966. It represents the excess expected return on the portfolio normalized by the portfolio's standard deviation and thus provides a compact measure of the reward to variability. The Sharpe ratio is also known as the market price of risk. Sharpe used this ratio to evaluate the performance of mutual funds, and it is now widely used as a measure of portfolio performance.

In continuous time finance, the instantaneous Sharpe ratio, γ_t , plays a key role in the transformation of a Brownian motion under the real-world measure P to a Brownian motion under the risk neutral measure Q . Suppose W_t is a Brownian motion under P and $\tilde{W}_t$ is a Brownian motion under Q , then we have, from the Girsanov theorem under suitable conditions, on γ

$$d\tilde{W}_t = dW_t + \gamma_t dt \quad (4)$$

It is interesting to see that the Sharpe ratio figures so prominently in this fundamental relationship in modern mathematical finance.

Bill Sharpe has made several other notable contributions to the development of the finance field. His papers have profoundly influenced investment science and portfolio management. He developed the first binomial tree model (see **Binomial Tree**) for option pricing, the gradient method for asset

allocation optimization and returns-based **style analysis** for evaluating the style and performance of investment funds. Sharpe has helped translate these theoretical ideas into practical applications. These applications include the creation of index funds and several aspects of retirement portfolio planning. He has written a number of influential textbooks, including *Investments*, used throughout the world. It is clear that Sharpe's ideas have been of great significance in the subsequent advances in the discipline of finance.

References

- [1] Sharpe, W.F. (1963). A simplified model for portfolio analysis, *Management Science* **9**(2), 277–293.
- [2] Sharpe, W.F. (1964). Capital asset prices—a theory of market equilibrium under conditions of risk, *The Journal of Finance*, **XIX**(3), 425–442.
- [3] Sharpe, W.F. (1966). Mutual fund performance, *Journal of Business* **39**, 119–138.

Further Reading

Sharpe, W.F., Alexander, G.J. & Bailey, J. (1999). *Investments*, Prentice-Hall.

Related Articles

Capital Asset Pricing Model; Style Analysis; Binomial Tree.

PHELIM BOYLE

Markowitz, Harry

♣ Harry Max Markowitz, born in Chicago in 1927, said in his 1990 Nobel Prize acceptance speech that, as a child, he was unaware of the Great Depression, which caused a generation of investors and noninvestors the world over to mistrust the markets. However, it was a slim, 15-page paper published by Markowitz as a young man that would eventually transform the way people viewed the relationship between risk and return, and that overhauled the way the investment community constructed diversified portfolios of securities.

Markowitz was working on his dissertation in economics at the University of Chicago when his now-famous “Portfolio Selection” paper appeared in the March 1952 issue of the *Journal of Finance* [1]. He was 25. He went on to win the Nobel Prize in Economic Sciences in 1990 for providing the cornerstone to what came to be known as *modern portfolio theory* (**Modern Portfolio Theory**).

Markowitz shared the Nobel Prize with Merton H. Miller and William F. Sharpe (**Sharpe, William F.**), who were recognized, respectively, for their work on how firms’ capital structure and dividend policy affect their stock price, and the development of the capital asset pricing model, which presents a way to measure the riskiness of a stock relative to the performance of the stock market as a whole. Together, the three redefined the way investors thought about the investment process, and created the field of financial economics. Markowitz, whose work built on earlier work on diversification by Yale University’s James Tobin, who received a Nobel Prize in 1981, was teaching at Baruch College at the City University of New York when he won the Nobel at the age of 63.

Markowitz received a bachelor of philosophy in 1947 and a PhD in economics in 1955, both from the University of Chicago. Years later he said that when he decided to study economics, his philosophical interests drew him toward the “economics of uncertainty”. At Chicago, he studied with Milton Friedman, Jacob Marschak, Leonard Savage, and Tjalling Koopmans, and became a student member of the famed Cowles Commission for Research in Economics (which moved to Yale University in 1955 and was renamed the Cowles Foundation).

The now-landmark 1952 “Portfolio Selection” paper skipped over the problem of selecting individual stocks and focused instead on how a manager or investor selects a portfolio best suited to the individual’s risk and return preferences. Pre-Markowitz, diversification was considered important, but there was no framework to determine how diversified a portfolio was or how an investor could create a well-diversified portfolio.

Keeping in mind that “diversification is both observed and sensible,” the paper began from the premise that investors consider expected return a “desirable thing” and risk an “undesirable thing”. Markowitz’s first insight was to look at a portfolio’s risk as the variance of its returns. This offered a way to quantify investment risk that previously had not existed. He then perceived that a portfolio’s riskiness depended not just on the expected returns and variances of the individual assets but also on the correlations between the assets in the portfolio. For Markowitz, the wisdom of diversification was not simply a matter of holding a large number of different securities, but of holding securities whose value did not rise and fall in tandem with one another. “It is necessary to avoid investing in securities with high covariances among themselves,” he stated in the paper. Investing in companies in different industries, for instance, increased a portfolio’s diversification and, paradoxically, improved the portfolio’s expected returns by reducing its variance.

Markowitz’s paper laid out a mathematical theory for deriving the set of optimal portfolios based on their risk-return characteristics. Markowitz showed how mean-variance analysis could be used to find a set of securities whose risk-return combinations were deemed “efficient”. Markowitz referred to this as the expected returns–variance of returns rule (E-V rule). The range of possible risk–return combinations yielded what Markowitz described as efficient and inefficient portfolios, an idea he based on Koopmans’ notion that there are efficient and inefficient allocations of resources [3]. Koopmans, at the time, was one of Markowitz’s professors. Markowitz’s notion of efficient portfolios was subsequently called the *efficient frontier*. “Not only does the E-V hypothesis imply diversification, it implies the ‘right kind’ of diversification for the ‘right reason,’” Markowitz wrote. The optimal portfolio was the one that would provide the minimum risk for a

given expected return, or the highest expected return for a given level of risk. An investor would select the portfolio whose risk-return characteristics he preferred.

It has been said many times over the years that Markowitz's portfolio theory provided, at long last, the math behind the adage "Don't put all your eggs in one basket." In 1988, Sharpe said of Markowitz's portfolio selection concept: "I liked the parsimony, the beauty, of it. . . . I loved the mathematics. It was simple but elegant. It had all the aesthetic qualities that a model builder likes" [5].

Back in 1952, Markowitz already knew the practical value of the E-V rule he had crafted. It functioned, his paper noted, both "as a hypothesis to explain well-established investment behavior and as a maxim to guide one's own action." However, Markowitz's insight was deeper. The E-V rule enabled the investment management profession to distinguish between investment and speculative behavior, which helped fuel the gradual institutionalization of the investment management profession. In the wake of Markowitz's ideas, investment managers could strive to build portfolios that were not simply groupings of speculative stocks but well-diversified sets of securities designed to meet the risk-return expectations of investors pursuing clear investment goals.

Markowitz's ideas gained traction slowly, but within a decade investment managers were turning to Markowitz's theory of portfolio selection (**Modern Portfolio Theory**) to help them determine how to select portfolios of diversified securities. This occurred as institutional investors in the United States were casting around for ways to structure portfolios that relied more on analytics and less on relationships with brokers and bankers. In the intervening years, Markowitz expanded his groundbreaking work. In 1956, he published the Critical Line Algorithm, which explained how to compute the efficient frontier for portfolios with large numbers of securities subject to constraints. In 1959, he published *Portfolio Selection: Efficient Diversification of Investments*, which bored further into the subject and explored the relationship between his mean-variance analysis and the fundamental theories of action under uncertainty of John von Neumann and Oskar Morgenstern, and of Leonard J. Savage [2].

However, while Markowitz is most widely known for his work in portfolio theory, he has said that

he values another prize he received more than the Nobel: the von Neumann Prize in operations research theory. That prize, he said, recognized the three main research areas that have defined his career. Markowitz received the von Neumann prize in 1989 from the Operations Research Society of America and the Institute of Management Sciences (now combined as INFORMS) for his work on portfolio theory, sparse matrix techniques and the high-level simulation language called *SIMSCRIPT programming language*.

After Chicago, Markowitz went to the RAND Corp. in Santa Monica, CA, where he worked with a group of economists on linear programming techniques. In the mid-1950s, he developed sparse matrices, a technique to solve large mathematical optimization problems. Toward the end of the decade, he went to General Electric to build models of manufacturing plants in the company's manufacturing services department. After returning to RAND in 1961, he and his team developed a high-level programming language for simulations called *SIMSCRIPT* to support Air Force projects that involved simulation models. The language was published in 1962. The same year, Markowitz and former colleague Herb Karr formed CACI, the California Analysis Center Inc. The firm later changed its name to Consolidated Analysis Centers Inc. and became a publicly traded company that provided IT services to the government and intelligence community. It is now called *CACI International*.

Markowitz's career has ranged across academia, research, and business. He worked in the money management industry as president of Arbitrage Management Company from 1969 to 1972. From 1974 until 1983, Markowitz was at IBM's T.J. Watson Research Center in Yorktown Heights, NY. He has taught at the University of California at Los Angeles, Baruch College and, since 1994, at the University of California at San Diego. He continues to teach at UC-San Diego and is an academic consultant to Index Fund Advisors, a financial services firm that provides low-cost index funds to investors.

In the fall of 2008 and subsequent winter, Markowitz's landmark portfolio theory came under harsh criticism in the lay press as all asset classes declined together. Markowitz, however, argued that the credit crisis and ensuing losses highlighted the benefits of diversification and exposed the risks in

not understanding, or in misunderstanding, the correlations between assets in a portfolio. “Portfolio theory was not invalidated, it was validated,” he noted in a 2009 interview with Index Fund Advisors [4]. He has said numerous times over the years that there are no “shortcuts” to understanding the trade-off between risk and return. “US portfolio theorists do not talk about risk control,” he said in that interview. “It sounds like you can control risk. You can’t.” “But diversification,” he continued, “is the next best thing.”

References

- [1] Markowitz, H.M. (1952). Portfolio selection, *Journal of Finance* 7, 77–91.
- [2] Markowitz, H.M. (1959). *Portfolio Selection: Efficient Diversification of Investments*, John Wiley & Sons, New York.
- [3] Markowitz, H.M. (2002). *An Interview with Harry Markowitz by Jeffrey R. Yost*, Charles Babbage Institute, University of Minnesota, Minneapolis, MN.
- [4] Markowitz, H.M. (2009). *An Interview with Harry M. Markowitz by Mark Hebner*, Index Fund Advisors, Irvine, CA.
- [5] Sharpe, W.F. (1988). *Revisiting the Capital Asset Pricing Model, an interview by Jonathan Burton*. Dow Jones Asset Manager, May/June, 20–28.

Related Articles

Modern Portfolio Theory; Risk–Return Analysis; Sharpe, William F.

NINA MEHTA

Merton, Robert C.

Robert C. Merton is the John and Natty McArthur University Professor at Harvard Business School. In 1966, he earned a BS in engineering mathematics from Columbia University where he published his first publication “The ‘Motionless’ Motion of Swift’s Flying Island” in the *Journal of the History of Ideas* [4]. He then went on to pursue graduate studies in applied mathematics at the California Institute of Technology, leaving the institution with an MS in 1967. He obtained a PhD in economics in 1970 from the Massachusetts Institute of Technology where he worked under the Nobel laureate Paul A. Samuelson (see **Samuelson, Paul A.**). His dissertation was entitled “Analytical Optimal Control Theory as Applied to Stochastic and Non-stochastic Economics.” Prior to joining Harvard in 1988, Merton served on the finance faculty of Massachusetts Institute of Technology.

In 1997, Merton shared the Nobel Prize in Economic Sciences with Myron Scholes “for a new method to determine the value of derivatives”. Merton taught himself stochastic dynamic programming and Ito calculus during graduate school at Massachusetts Institute of Technology and subsequently introduced Ito calculus (see **Stochastic Integrals**) into finance and economics. Continuous-time stochastic calculus had become a cornerstone in mathematical finance, and more than anyone Merton is responsible in making manifest the mathematical tool’s power in financial modeling and applications. Merton had also produced highly regarded work on dynamic models of optimal life-time consumption and portfolio selection, equilibrium asset pricing, contingent-claim analysis, and financial systems. Merton’s monograph “Continuous-time finance” [8] is a classic introduction to these topics.

Merton proposed an intertemporal capital asset pricing model (ICAPM) [6] (see **Capital Asset Pricing Model**), a model empirically more attractive than the single-period capital asset pricing model (CAPM) (see **Capital Asset Pricing Model**). Assuming continuous-time stochastic processes with continuous-decision-making and trading, Merton showed that mean–variance portfolio choice is optimal at each moment of time. It explained when and how the CAPM could hold in a dynamic setting. As an extension, Merton looked at the

case when the set of investment opportunities is stochastic and evolves over time. Investors hold a portfolio to hedge against shifts in the opportunity set of security returns. This implies that investors are compensated in the expected return for bearing the risk of shifts in the opportunity set of security returns, in addition to bearing market risk. Because of this additional compensation in expected return, in equilibrium, expected returns on risky assets may differ from the risk-less expected return even when they have no market risk. Through this work, we obtain an empirically more useful version of CAPM that allows for multiple risk factors. Merton’s ICAPM predated many subsequently published multifactor models like the arbitrage pricing theory [11] (see **Arbitrage Pricing Theory**).

Merton’s work in the 1970s laid the foundation for modern derivative pricing theory (see **Option Pricing: General Principles**). His paper “Theory of Rational Option Pricing” [5] is one of the two classic papers on derivative pricing that led to the Black–Scholes–Merton option pricing theory (see **Black–Scholes Formula**). Merton’s essential contribution was his hedging (see **Hedging**) argument for option pricing based on no arbitrage; he showed that one can use the prescribed dynamic trading strategy under Black–Scholes [1] to offset the risk exposure of an option and obtain a perfect hedge under the continuous trading limit. In other words, he discovered how to construct a “synthetic option” using continual revision of a “self-financing” portfolio involving the underlying asset and riskless borrowing to replicate the expiration-date payoff of the option. And no arbitrage dictates that the cost of constructing this synthetic option must give the price of the option even if it does not exist. This seminal paper also extended the Black–Scholes model to allow for predictably changing interest rates, dividend payments on the underlying asset, changing exercise price, and early exercise under American options. Merton also produced “perhaps the first closed-form formula for an exotic option”. [12] Merton’s approach to derivative securities provided the intellectual basis for the rise of the profession of financial engineering.

The Merton model (see **Structural Default Risk Models**) refers to an increasingly popular structural credit risk model introduced by Merton [7] in the early 1970s. Drawing on the insight that the payoff

structure of the leveraged equity of a firm is identical to that of a call option (*see Call Options*) on the market value of the assets of the whole firm, Merton proposed that the leveraged equity of a firm could be valued as if it were a call option on the assets of the whole firm. The isomorphic (same payoff structure) price relation between the leveraged equity of a firm and a call option allows one to apply the Black–Scholes–Merton contingent-claim pricing model to value the equities [7]. The value for the corporate debt could then be obtained by subtracting the value of the option-type structure that the leveraged equity represents from the total market value of the assets. Merton's methodology offered a way to obtain valuation functions for the equity and debt of a firm, a measure of the risk of the debt, as well as all the Greeks of contingent-claim pricing. The Merton model provided a useful basis for valuing and assessing corporate debt, its risk, and the sensitivity of debt value to various parameters (e.g., the delta gives the sensitivity of either debt value or equity value to change in asset value). Commercial versions of the Merton model include the KMV model and the Jarrow–Turnbull model.

Since the 1990s, Merton collaborated with Zvi Bodie, Professor of Finance at Boston University to develop a new line of research on the financial system [2, 9, 10]. They adopted a functional perspective, “similar in spirit to the functional approach in sociology pioneered by Robert K. Merton (1957)” [3, 9]. By focusing on the underlying functions of financial systems, the functional perspective takes functions rather than institutions and forms as the conceptual anchor in its analysis of financial institutional change over time and contemporaneous institutional differences across borders. The functional perspective is also useful for predicting and guiding financial institutional change. The existing approaches of neoclassical, institutional, and behavioral theories in economics are taken as complementary in the functional approach to understand financial systems.

Merton had made significant contributions to finance across a broad spectrum and they are too numerous to mention exhaustively. His other works include those on Markowitz–Sharpe-type models with investors with homogeneous beliefs but with incomplete information about securities, the use of

jump-diffusion models (*see Jump-diffusion Models*) in option pricing, valuation of market forecasts, pension reforms, and employee stock option (*see Employee Stock Options*).

In addition to his academic duties, Merton has also been partner of the now defunct hedge fund Long Term Capital Management (*see Long-Term Capital Management*) and is currently Chief Scientific Officer at the Trinsum Group.

References

- [1] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**(3), 637–659.
- [2] Crane, D., Froot, K., Mason, S., Perold, A., Merton, R.C., Bodie, Z., Sirri, E. & Tufano, P. (1995). *The Global Financial System: A Functional Perspective*, Harvard Business School Press, Boston, MA.
- [3] Merton, R.K. (1957). *Social Theory and Social Structure*, revised and enlarged edition, The Free Press, Glencoe, IL.
- [4] Merton, R.C. (1966). The “Motionless” Motion of Swift's flying island, *Journal of the History of Ideas* **27**, 275–277.
- [5] Merton, R.C. (1973). Theory of rational option theory, *Bell Journal of Economics and Management Science* **4**(1), 141–183.
- [6] Merton, R.C. (1973). An intertemporal capital asset pricing model, *Econometrica* **41**(5), 867–887.
- [7] Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**(2), 449–470.
- [8] Merton, R.C. (1990). *Continuous-Time Finance*, Blackwell, Malden, MA.
- [9] Merton, R.C. & Bodie, Z. (1995). A conceptual framework for analyzing the financial system. Chapter 1 in *The Global Financial System: A Functional Perspective*, D. Crane, K. Froot, S. Mason, A. Perold, R. Merton, Z. Bodie, E. Sirri, & P. Tufano, eds, Harvard Business School Press, Boston, MA, pp. 3–31.
- [10] Merton, R.C. & Bodie, Z. (2005). Design of financial systems: towards a synthesis of function and structure, *Journal of Investment Management* **3**(1), 1–23.
- [11] Ross, S. (1976). The arbitrage theory of capital asset pricing, *Journal of Economic Theory* **13**(3), 341–360.
- [12] Rubinstein, M. (2006). *A History of the Theory of Investments*, John Wiley & Sons, Hoboken, NJ, p. 240.

Further Reading

Merton, R.C. (1990). *Continuous-Time Finance*, Blackwell, Malden, MA.

Related Articles

Black, Fischer; Black–Scholes Formula; Jump-diffusion Models; Long-Term Capital Manage-

ment; Merton Problem; Option Pricing: General Principles; Option Pricing Theory: Historical Perspectives; Partial Differential Equations; Samuelson, Paul A.; Structural Default Risk Models; Thorp, Edward.

ALEX HAMILTON CHAN

Arbitrage: Historical Perspectives

The concept of arbitrage has acquired a precise, technical meaning in quantitative finance (see **Arbitrage Pricing Theory**; **Arbitrage Strategy**; **Arbitrage Bounds**). In theoretical pricing of derivative securities, an arbitrage is a riskless trading strategy that generates a positive profit with no net investment of funds. This definition can be loosened to allow the positive profit to be nonnegative, with no possible future state having a negative outcome and at least one state with a positive outcome. Pricing formulas for specific contingent claims are derived by assuming an absence of arbitrage opportunities. Generalizing this notion of arbitrage, the fundamental theorem of asset pricing provides that an absence of arbitrage opportunities implies the existence of an equivalent martingale measure (see **Fundamental Theorem of Asset Pricing**; **Equivalent Martingale Measures**). Combining absence of arbitrage with a linear model of asset returns, the arbitrage pricing theory decomposes the expected return of a financial asset into a linear function of various economic risk factors, including market indices. Sensitivity of expected return to changes in each factor is represented by a factor-specific beta coefficient. Significantly, while riskless arbitrage imposes restrictions on prices observed at a given point in time, the arbitrage pricing theory seeks to explain expected returns, which involve prices observed at different points in time.

In contrast to the technical definitions of arbitrage used in quantitative finance, colloquial usage of arbitrage in modern financial markets refers to a range of trading strategies, including municipal bond arbitrage; merger arbitrage; and convertible bond arbitrage. Correctly executed, these strategies involve trades that are low risk relative to the expected return but do have possible outcomes where profits can be negative. Similarly, uncovered interest arbitrage seeks to exploit differences between foreign and domestic interest rates leaving the risk of currency fluctuations unhedged. These notions of risky arbitrage can be contrasted with covered interest arbitrage, which corresponds to the definition of arbitrage used in quantitative finance of a riskless trading strategy that generates a positive profit with no net investment of funds. Cash-and-carry arbitrages related to

financial derivatives provide other examples of arbitrages relevant to the quantitative finance usage. Among the general public, confusion about the nature of arbitrage permitted Bernard Madoff to use the illusion of arbitrage profit opportunities to attract “hedge fund investments” into the gigantic Ponzi scheme that collapsed in late 2008. Tracing the historical roots of arbitrage trading provides some insight into the various definitions of arbitrage in modern usage.

Arbitrage in Ancient Times

Records about business practices in antiquity are scarce and incomplete. Available evidence is primarily from the Middle East and suggests that mercantile trade in ancient markets was extensive and provided a number of avenues for risky arbitrage. Potential opportunities were tempered by the lack of liquidity in markets; the difficulties of obtaining information and moving goods over distances; and, inherent political and economic risks. Trading institutions and available securities were relatively simple. Circa 1760 BC, the Code of Hammurabi dealt extensively with matters of trade and finance. Sumerian cuneiform tablets from that era indicate a rudimentary form of bill of exchange transaction was in use where a payment (disbursement) would be made in one location in the local unit of account, for example, barley, in exchange for disbursement (payment) at a later date in another location of an agreed upon amount of that local currency, for example, lead [6]. The date was typically determined by the accepted transport time between the locations. Two weeks to a month was a commonly observed time between the payment and repayment. The specific payment location was often a temple.

Ancient merchants developed novel and complex solutions to address the difficulties and risks in executing various arbitrage transactions. Because the two payments involved in the ancient bill of exchange were separated by distance and time, a network of agents, often bound together by family or tribal ties, was required to disburse and receive funds or goods in the different locations. Members of the caravan or ship transport were often involved in taking goods on consignment for sale in a different location where the cost of the goods would be repaid [6, p.15–6]. The merchant arbitrageur would offset the cost of purchasing goods given on consignment with payments from

other merchants seeking to avoid the risks of carrying significant sums of money over long distance, making a local payment in exchange for a disbursement of the local currency in a different location. The basic cash-and-carry arbitrage is complicated by the presence of different payment locations and currency units. The significant risk of delivery failure or nonpayment was controlled through the close-knit organizational structure of the merchant networks [7]. These same networks provided information on changing prices in different regions that could be used in geographical goods arbitrage.

The gradual introduction of standardized coinage starting around the 650 BC expanded available arbitrage opportunities to include geographical arbitrage of physical coins to exploit differing exchange ratios [6, p.19–20]. For example, during the era of the Athenian empire (480–404 BC), Persia maintained a bimetallic coinage system where silver was undervalued relative to gold. The resulting export of silver coins from Persia to Greece and elsewhere in the Mediterranean is an early instance of a type of arbitrage activity that became a mainstay of the arbitrageur in later years. This type of arbitrage trading was confined to money changers with the special skills and tools to measure the bullion value of coins. In addition to the costs and risks of transportation, the arbitrage was restricted by the seigniorage and minting charges levied in the different political jurisdictions. Because coinage was exchanged by weight and trading by bills of exchange was rudimentary, there were no arbitrageurs specializing solely in “arbitrating of exchange rates”. Rather, arbitrage opportunities arose from the trading activities of networks of merchants and money changers. These opportunities included uncovered interest arbitrage between areas with low interest rates, such as Jewish Palestine, and those with high rates, such as Babylonia [6, p.18–19].

Evolution of the Bill of Exchange

Though the precise origin of the practice is unknown, “arbitration of exchange” first developed during the Middle Ages. Around the time of the First Crusade, Genoa had emerged as a major sea power and important trading center. The Genoa fairs had become sufficiently important economic and financial events that attracted traders from around the Mediterranean.

To deal with the problems of reconciling transactions using different coinages and units of account, a forum for arbitrating exchange rates was introduced. On the third day of each fair, a representative body composed of recognized merchant bankers would assemble and determine the exchange rates that would prevail for that fair. The process involved each banker suggesting an exchange rate and, after some discussion, a voting process would determine the exchange rates that would apply at that fair. Similar practices were adopted at other important fairs later in the Middle Ages. At Lyon, for example, Florentine, Genoese, and Lucca bankers would meet separately to determine rates, with the average of these group rates becoming the official rate. These rates would then apply to bill transactions and other business conducted at the fair. Rates typically stayed constant between fairs in a particular location providing the opportunity for arbitrage of exchange rates across fairs in different locations.

From ancient beginnings involving commodity transactions of merchants, the bill of exchange evolved during the Middle Ages to address the difficulties of using specie or bullion to conduct foreign exchange transactions in different geographical locations. In general, a bill of exchange contract involved four persons and two payments. The bill is created when a “deliverer” exchanges domestic cash money for a bill issued by a “taker”. The issued bill of exchange is drawn on a correspondent or agent of the taker who is situated abroad. The correspondent, the “payer”, is required to pay a stated amount of foreign cash money to the “payee”, to whom the bill is made payable. Consider the precise text of an actual bill of exchange from the early seventeenth century that appeared just prior to the introduction of negotiability [28, p.123]:

March 14, 1611

In London for £69.15.7 at 33.9

At half usance pay by this first of exchange to Francesco Rois Serra sixty-nine pounds, fifteen shillings, and seven pence sterling at thirty-three shillings and nine pence groat per £ sterling, value [received] from Master Francesco Pinto de Britto, and put it into our account, God be with you.

Giovanni Calandrini and
Filippo Burlamachi

Accepted

[On the back:] To Balthasar Andrea in Antwerp
First 117.15.0 [pounds groat]

The essential features of the bill of exchange all appear here: the four separate parties; the final payment being made in a different location from the original payment; and the element of currency exchange. “Usance” is the period of time, set by custom, before a bill of exchange could be redeemed at its destination. For example, usance was 3 months between Italy and London and 4 weeks between Holland and London. The practice of issuing bills at usance, as opposed to specifying any number of days to maturity, did not disappear until the nineteenth century [34, p.7].

Commercial and financial activities in the Middle Ages were profoundly impacted by Church doctrine and arbitrage trading was no exception. Exchange rates determined for a given fair would have to be roughly consistent with triangular arbitrage to avoid Church sanctions. In addition, the Church usury prohibition impacted the payment of interest on money loans. Because foreign exchange transactions were licit under canon law, it was possible to disguise the payment of interest in a combination of bill of exchange transactions referred to as *dry exchange* or *fictitious exchange* [13, p.380–381], [17, 26]. The associated exchange and re-exchange of bills was a risky set of transactions that could be covertly used to invest money balances or to borrow funds to finance the contractual obligations. The expansion of bill trading for financial purposes combined with the variation in the exchange rates obtained at fairs in different locations provided the opportunity of geographical arbitrage of exchange rates using bills of exchange. It was this financial practice of exploiting differences in bill exchange rates between financial centers that evolved into the “arbitration of exchange” identified by la Porte [22], Savary [24], and Postelwayte [30] in the eighteenth century.

The bill of exchange contract evolved over time to meet the requirements of merchant bankers. As monetary units became based on coinage with specific bullion content, the relationship between exchange rates in different geographical locations for bills of exchange, coinage, and physical bullion became the mainstay of traders involved in “arbitration of exchange”. Until the development of the “inland” bill in early seventeenth century in England, all bills of exchange involved some form of foreign exchange trading, and hence the name *bill of exchange*. Contractual features of the bill of exchange, such as negotiability and priority of claim, evolved over time

producing a number of different contractual variations [9, 15, 26]. The market for bills of exchange also went through a number of different stages. At the largest and most strategic medieval fairs, financial activities, especially settlement and creation of bills of exchange, came to dominate the trading in goods [27]. By the sixteenth century, bourses such as the Antwerp Exchange were replacing the fairs as the key international venues for bill trading.

Arbitrage in Coinage and Bullion

Arbitrage trading in coins and bullion can be traced to ancient times. Reflecting the importance of the activity to ordinary merchants in the Middle Ages, methods of determining the bullion content of coins from assay results, and rates of exchange between coins once bullion content had been determined, formed a substantial part of important commercial arithmetics, such as the *Triparty* (1484) of Nicolas Chuquet [2]. The complications involved in trading without a standardized unit of account were imposing. There were a sizable number of political jurisdictions that minted coins, each with distinct characteristics and weights [14]. Different metals and combinations of metals were used to mint coinage. The value of silver coins, the type of coins most commonly used for ordinary transactions, was constantly changing because of debasement and “clipping”. Over time, significant changes in the relative supply of gold and silver, especially due to inflows from the New World, altered the relative values of bullion. As a result, merchants in a particular political jurisdiction were reluctant to accept foreign coinage at the par value set by the originating jurisdiction. It was common practice for foreign coinage to be assayed and a value set by the mint conducting the assay. Over time, this led to considerable market pressures to develop a unit of account that would alleviate the expensive and time-consuming practice of determining coinage value.

An important step in the development of such a standardized unit of account occurred in 1284 when the Doge of Venice began minting the gold ducat: a coin weighing about 3.5 g and struck in 0.986 gold. While ducats did circulate, the primary function was as a trade coin. Over time, the ducat was adopted as a standard for gold coins in other countries, including other Italian city states, Spain,

Austria, the German city states, France, Switzerland, and England. Holland first issued a ducat in 1487 and, as a consequence of the global trading power of Holland in the sixteenth and seventeenth centuries, the ducat became the primary trade coin for the world. Unlike similar coins such as the florin and guinea, the ducat specifications of about 3.5 g of 0.986 gold did not change over time. The use of mint parities for specific coins and market prices for others did result in the gold–silver exchange ratio differing across jurisdictions. For example, in 1688, the Amsterdam gold–silver ratio for the silver rixdollar mint price and gold ducat market price was 14.93 and, in London, the mint price ratio was 15.58 for the silver shilling and gold guinea [25, p.475]. Given transport and other costs of moving bullion, such gold/silver price ratio differences were not usually sufficient to generate significant bullion flows. However, combined in trading with bills of exchange, substantial bullion flows did occur from arbitrage trading.

Details of a May 1686 arbitrage by a London goldsmith involving bills of exchange and gold coins are provided by Quinn [25, p.479]. The arbitrage illustrates how the markets for gold, silver, and bills of exchange interacted. At that time, silver was the primary monetary metal used for transactions though gold coins were available. Prior to 1663, when the English Mint introduced milling of coins with serrated edges to prevent clipping, all English coins were “hammered” [20]. The minting technology of hammering coins was little changed from Roman times. The process produced imperfect coins, not milled at the edges, which were only approximately equal in size, weight, and imprint making altered coins difficult to identify [29, ch.4]. Such coins were susceptible to clipping, resulting in circulating silver coins that were usually under the nominal Mint weight. Despite a number of legislative attempts at remedying the situation, around 1686, the bulk of the circulating coins in England were still hammered silver. The Mint would buy silver and gold by weight in exchange for milled silver shilling coins at a set price per ounce. When the market price of silver rose sufficiently above the mint price, English goldsmiths would melt the milled silver coin issued by the Mint, though it was technically illegal to do so.

In addition to mint prices for silver and gold, there were also market prices for gold and silver. Around 1686, the Mint would issue guineas in exchange

for silver shillings at a fixed price (£1.075 = 21s. 6d./oz.). In Amsterdam, the market price for a Dutch gold ducat was 17.5 schellingen (*S*). Observing that the ducat contained 0.1091 ounces of recoverable gold and the guinea 0.2471 ounces, it follows that 36.87 *S* could be obtained for £1 if gold was used to effect the exchange. Or, put differently, 1 ducat would produce £0.4746. Because transportation of coins and bullion was expensive, there was a sizable band within which rates on bills of exchange could fluctuate without producing bullion flows. If the (*S*/£) bill exchange rate rose above the rate of exchange for gold plus transport costs, merchants in Amsterdam seeking funds in London would prefer to send gold rather than buy bills of exchange on London. Merchants in London seeking funds in Amsterdam would buy bills on Amsterdam to benefit from the favorable exchange. Similarly, if the bill exchange rate fell below the rate of exchange for silver plus transport costs, merchants in London would gain by exporting silver to Amsterdam rather than buying a bill on Amsterdam.

To reconstruct the 1686 goldsmith arbitrage, observe that the exchange rate for a 4-week bill in London on Amsterdam at the time of the arbitrage was 37.8 (*S*/£). Obtaining gold ducats in Holland for £0.4746 and allowing for transport costs of 1.5% and transport time of 1 week produces gold in London for £0.4676. Using this gold to purchase a bill of exchange on Amsterdam produces 17.6715 *S* in Amsterdam 5 weeks after the trade is initiated, an arbitrage profit of 0.1715 *S*. Even if the gold can be borrowed in Amsterdam and repaid in silver, the trade is not riskless owing to the transport risk and the possible movement in bill rates before the bill is purchased in London. These costs would be mitigated significantly for a London firm also operating in the bill and bullion market of Amsterdam, as was the case with a number of London goldsmiths. The strength of the pound sterling in the bill market from 1685–1688 generated gold inflows to England from this trade higher than any other four-year period in the seventeenth century [25, p.478]. The subsequent weakening of the pound in the bill market from 1689 until the great recoinage in 1696 led to arbitrage trades switching from producing gold inflows to substantial outflows of silver from melted coins and clipping.

Bill of Exchange Arbitrage

The roots of “arbitration of exchange” can be traced to the transactions of medieval merchant bankers seeking to profit from discrepancies in bill exchange rates across geographical locations [27, 28]. For example, if sterling bills on London were cheaper in Paris than in Bruges, then medieval bankers would profit by selling sterling in Bruges and buying in Paris. The effect of such transactions was to keep all exchange rates roughly in parity with the triangular arbitrage condition. Temporary discrepancies did occur but such trading provided a mechanism of adjustment. The arbitrages were risky even when done entirely with bills of exchange. Owing to the slowness of communications, market conditions could change before bills of exchange reached their destination and the re-exchange could be completed. As late as the sixteenth century, only the Italian merchant bankers, the Fuggers of Augsburg, and a few other houses with correspondents in all banking centers were able to engage actively in arbitrage [28, p.137]. It is not until the eighteenth century that markets for bills were sufficiently developed to permit arbitration of exchange to become standard practice of merchants deciding on the most profitable method of remitting or drawing funds offshore.

The transactions in arbitration of exchange by medieval bankers are complicated by the absence of offsetting cash flows in the locations where bills are bought and sold. In the example above, the purchase of a bill in Paris would require funds, which are generated by the bill sale in Bruges. The profits are realized in London. Merchant bankers would be able to temporarily mitigate the associated geographical fund imbalances with internally generated capital, but re-exchanges or movements of bullion were necessary if imbalances persisted. To be consistent with the spirit of the self-financing element of modern riskless arbitrage, the example of medieval banker arbitrage among Paris, Bruges, and London can be extended to two issuing locations and two payment centers. It is possible for the same location to be used as both the issuing and payment location but that will not be assumed. Let the two issuing locations be, say, Antwerp and Hamburg, with the two payment locations being London and Venice. The basic strategy involves making offsetting bill transactions in the two issuing locations

and then matching the settlements in the payment centers.

In the following example, $\$G$ is the domestic currency in Hamburg and $\$A$ is the domestic currency in Antwerp, the forward exchange rate imbedded in the bill transaction is denoted as F_1 for Ducats/ $\$A$; F_2 for Ducats/ $\$G$; F_3 for $\pounds/\$G$; and, F_4 for $\pounds/\$A$.

In Hamburg

Acquire $\$G Q_G$ using a bill which agrees to pay ($\$G Q_G F_2$) in Venice at time T	Deliver the $\$G Q_G$ on another bill which agrees to be repaid ($\$G Q_G F_3$) in London at time T
--	---

In Antwerp

Acquire $\$A Q_A$ using a bill which agrees to pay ($\$A Q_A F_4$) in London at time T	Deliver the $\$A Q_A$ on another bill which agrees to be repaid ($\$A Q_A F_1$) in Venice at time T
--	---

At $t = 0$, the cash flows from all the bill transactions at $t = 0$ offset. If the size of the borrowings in the two issuing centers is calculated to produce the same maturity value, in terms of the domestic currencies of the two payment centers, then the profit on the transaction depends on the relative values of the payment center currencies in the issuing centers. If there is sufficient liquidity in the Hamburg and Antwerp bill markets, the banker can generate triangular arbitrage trades designed to profit from discrepancies in bid/offer rates arising in different geographical locations.

To see the precise connection to triangular arbitrage, consider the profit function from the trading strategy. At time T in Venice, the cash flows would provide $(\$A Q_A F_1) - (\$G Q_G F_2)$. And, in London, the cash flows would provide $(\$G Q_G F_3) - (\$A Q_A F_4)$. For the intermediary operating in both locations, the resulting profit (π) on the trade would be the sum of the two cash flows:

$$\begin{aligned}
 \pi(T) &= (\$A Q_A F_1 - \$G Q_G F_2) \\
 &\quad + (\$G Q_G F_3 - \$A Q_A F_4) \\
 &= \$A Q_A (F_1 - F_4) + \$G Q_G (F_3 - F_2)
 \end{aligned} \tag{1}$$

Constructing the principal values of the two transactions to be of equal value now permits the substitution of $Q_G = Q_A (\$G/\$A)$, where $(\$G/\$A) = F_0$ is the prevailing exchange rate between $\$G$ and $\$A$:

$$\begin{aligned}\pi(T) &= \$A Q_A [(F_1 - F_0 F_2) - (F_4 - F_0 F_3)] \\ &= \$A Q_A \left[\left(\frac{\text{Ducats}}{\$A} - \frac{\$G \text{ Ducats}}{\$A G} \right) \right. \\ &\quad \left. - \left(\frac{\pounds}{\$A} - \frac{\$G \pounds}{\$A \$G} \right) \right] \quad (2)\end{aligned}$$

The two values in brackets will be zero if triangular arbitrage holds for both currencies. If the direct and indirect exchange rates for one of the currencies are not consistent with triangular arbitrage, then the banker can obtain a self-financing arbitrage profit.

Arbitration of Exchange

By the eighteenth century, the bill market in key financial centers such as Amsterdam, London, Hamburg, and Paris had developed to the point where merchants as well as bankers could engage in arbitration of exchange to determine the most profitable method of remitting funds to or drawing funds from offshore locations. From a relatively brief treatment in early seventeenth century sources, for example, [13], merchants' manuals detailing technical aspects of bill trading were available by the beginning of the eighteenth century. The English work by Justice, *A General Treatise on Money and Exchanges* [9], an expanded translation of an earlier treatise in French by M. Ricard, details the workings of bill transactions, recognizing subtle characteristics in the bill contract. However, as a reflection of the rudimentary state of the English bill market in the early eighteenth century, Justice did not approve of "drawing bills upon one country payable in another" due to the "difference in the Laws of Exchange, in different countries" giving rise to "a great many inconveniences" [9, p.28]. As the eighteenth century progressed, there was substantial growth in the breadth and depth of the bill market supported by increases in speed of communication between key financial centers with London emerging as the focal point [16, 31]. This progress was reflected in the increasingly sophisticated treatment of arbitration of exchange in merchants' manuals.

Merchants' manuals of the eighteenth and nineteenth centuries typically present arbitration of exchange from the perspective of a merchant engaged in transferring funds. In some sources, self-financing arbitrage opportunities created by combining remitting and drawing opportunities are identified. Discussions of the practice invariably involve calculations of the "arbitrated rates". Earlier manuals such as the one by Le Moine [11] only provide a few basic calculations aimed to illustrate the transactions involved. The expanded treatment in Postlewayt [24] provides a number of worked calculations. In one example, exchange rates at London are given as London–Paris 31 3/4 pence sterling for 1 French crown; London–Amsterdam as 240 pence sterling for 414 groats. Worked calculations are given for the problem "What is the proportional arbitrated price between Amsterdam and Paris?" Considerable effort is given to show the arithmetic involved in determining this arbitrated rate as 54 123/160 groat for 1 crown. Using this calculated arbitrated exchange rate and the already known actual London–Paris rate, Postlewayt then proceeds to determine the arbitrated rate for London–Amsterdam using these exchange rates for Paris–London and Paris–Amsterdam finding that it equals 240 pence sterling for 414 groats.

Having shown how to determine arbitrated rates, Postlewayt provides worked examples of appropriate arbitrage trades when the actual exchange rate is above or below the arbitrated rate. For example, when the arbitrated Amsterdam–Paris rate is above the actual rate, calculations are provided to demonstrate that drawing sterling in London by selling a bill on Paris, using the funds to buy a bill on Amsterdam and then exchanging the guilders/groats received in Amsterdam at the actual rate to cover the crown liability in Paris will produce a self-financing arbitrage profit. Similarly, when the arbitrated Amsterdam–Paris rate is below the actual rate, the trades in the arbitrage involve drawing sterling in London by selling a bill on Amsterdam, using the funds to buy a bill on Paris and then exchanging at the actual Amsterdam–Paris exchange rate the crowns received in Paris to cover the guilder liability. This is similar to the risky medieval banker arbitrage where the rate on re-exchange is uncertain. Though the actual rate is assumed to be known, in practice, this rate could change over the time period it takes to settle the relevant bill transactions. However, the degree of risk

facing the medieval banker was mitigated by the 18th century due to the considerably increased speed of communication between centers and subsequent developments in the bill contract, such as negotiability and priority of claim.

Earlier writers on arbitration of exchange, such as Postlewayt, accurately portrayed the concept but did not adequately detail all costs involved in the transactions. By the nineteenth century, merchants' manuals such as [34] accurately described the range of adjustments required for the actual execution of the trades. Taking the perspective of a London merchant with sterling seeking to create a fund of francs in Paris, a difference is recognized between two methods of determining the direct rate of exchange: buying a bill in the London market for payment in Paris; or having correspondents in Paris issue for francs a bill for sterling payment in London. In comparing with the arbitrated rates, the more advantageous direct rate is used. In determining direct rates, 3-month bill exchange rates are used even though the trade is of shorter duration. These rates are then adjusted to "short" rates to account for the interest factor. Arbitrated rates are calculated and, in comparing with direct rates, an additional brokerage charge (plus postage) is deducted from the indirect trade due to the extra transaction involved, for example, a London merchant buys a bill for payment in Frankfurt, which is then sold in Paris. No commissions are charged as it is assumed that the trade is done "between branches of the same house, or on joint account" [34, p.98].

Arbitrage in Securities and Commodities

Arbitrage involving bills of exchange survives in modern times in the foreign exchange swap trades of international banks. Though this arbitrage is of central historical importance, it attracts less attention now than a range of arbitrage activities involving securities and commodities that benefited from the financial and derivative security market developments of the nineteenth century. Interexchange and geographical arbitrages were facilitated by developments in communication. The invention of the telegraph in 1844 permitted geographical arbitrage in stocks and shares between London and the provincial stock exchanges by the 1850s. This trade was referred to as *shunting*. In 1866, Europe and America were linked by cable, significantly enhancing the

speed at which price discrepancies across international markets could be identified. Telegraph technology allowed the introduction of the stock market ticker in 1867. Opportunity for arbitrating differences in the prices of securities across markets was further aided by expansion of the number and variety of stocks and shares, many of which were interlisted on different regional and international exchanges. (Where applicable, the nineteenth century convention of referring to fixed-income securities as *stocks* and common stocks as *shares* will be used.) For example, after 1873 arbitrating the share price of Rio Tinto between the London and Paris stock exchanges was a popular trade.

Cohn [3, p.3] attributes "the enormous increase in business on the London Stock Exchange within the last few years" to the development of "Arbitrage transactions between London and Continental Bourses". In addition to various government bond issues, available securities liquid enough for arbitrage trading included numerous railway securities that appeared around the middle of the century. For example, both Haupt [8] and Cohn [3] specifically identify over a dozen securities traded in Amsterdam that were sufficiently liquid to be available for arbitrage with London. Included on both lists are securities as diverse as the Illinois and Erie Railway shares and the Austrian government silver loan. Securities of mines and banks increased in importance as the century progressed. The expansion in railway securities, particularly during the US consolidations of the 1860s, led to the introduction of traded contingencies associated with these securities such as rights issues, warrant options, and convertible securities. Weinstein [33] identifies this development as the beginning of arbitrage in equivalent securities, which, in modern times, encompasses convertible bond arbitrage and municipal bond arbitrage. However, early eighteenth century English and French subscription shares do have a similar claim [32]. Increased liquidity in the share market provided increased opportunities for option trading in stocks and shares.

Also during the nineteenth century, trading in "time bargains" evolved with the commencement of trading in such contracts for agricultural commodities on the Chicago Board of Trade in 1851. While initially structured as forward contracts, adoption of the General Rules of the Board of Trade in 1865 laid a foundation for trading of modern

futures contracts. Securities and contracts with contingencies have a history stretching to ancient times when trading was often done using samples and merchandise contracts had to allow for time to delivery and the possibility that the sample was not representative of the delivered goods. Such contingencies were embedded in merchandise contracts and were not suited to arbitrage trading. The securitization of such contingencies into forward contracts that are adaptable to cash-and-carry arbitrage trading can be traced to the introduction of “to arrive” contracts on the Antwerp bourse during the sixteenth century [19, ch.9]. Options trading was a natural development on the trade in time bargains, where buyers could either take delivery or could pay a fixed fee in lieu of delivery. In effect, such forward contracts were bundled with an option contract having the premium paid at delivery.

Unlike arbitration of exchange using bills of exchange, which was widely used and understood by the eighteenth century, arbitrage trades involving options—also known as *privileges* and *premiums*—were not. Available sources on such trades conducted in Amsterdam, Joseph de la Vega [21, ch.3] and Isaac da Pinto [19, p.366–377] were written by observers who were not the actual traders, so only crude details of the arbitrage trades are provided. Conversion arbitrages for put and call options, which involves knowledge of put–call parity, are described by both de la Vega and da Pinto. Despite this, prior to the mid-nineteenth century, options trading was a relatively esoteric activity confined to a specialized group of traders. Having attracted passing mention by Cohn [3], Castelli [1, p.2] identifies “the great want of a popular treatise” on options as the reason for undertaking a detailed treatment of mostly speculative option trading strategies. In a brief treatment, Castelli uses put–call parity in an arbitrage trade combining a short position in “Turks 5%” in Constantinople with a written put and purchased call in London. The trade is executed to take advantage of “enormous contangoes collected at Constantinople” [1, p.74–77].

Etymology and Historical Usage

The Oxford International Dictionary [12] defines arbitrage as: “the traffic in bills of exchange drawn on sundry places, and bought or sold in sight of the

daily quotations of rates in several markets. Also, the similar traffic in stock.” The initial usage is given as 1881. Reference is also directed to “arbitration of exchange” where the definition is “the determination of the rate of exchange to be obtained between two countries or currencies, when the operation is conducted through a third or several intermediate ones, in order to ascertain the most advantageous method of drawing or remitting bills.” The singular position given to “arbitration of exchange” trading using bills of exchange recognizes the practical importance of these securities in arbitrage activities up to that time. The Oxford International Dictionary definition does not recognize the specific concepts of arbitrage, such as triangular currency arbitrage or interexchange arbitrage, or that such arbitrage trading applies to coinage, bullion, commodities, and shares as well as to trading bills of exchange. There is also no recognition that doing arbitrage with bills of exchange introduces two additional elements not relevant to triangular arbitrage for manual foreign exchange transactions: time and location.

The word “arbitrage” is derived from a Latin root (*arbitrari*, to give judgment; *arbitrio*, arbitration) with variants appearing in the Romance languages. Consider the modern Italian variants: *arbitraggio* is the term for arbitrage; *arbitrato* is arbitration or umpiring; and, *arbitrarer* is to arbitrate. Similarly, for modern French variants, *arbitrage* is arbitration; *arbitrer* is to arbitrate a quarrel or to umpire; and *arbitre* is an arbitrator or umpire. Recognizing that the “arbitration of prices” concept underlying arbitrage predates Roman times, the historical origin where the word arbitrage or a close variant was first used in relation to arbitrating differences in prices is unknown. A possible candidate involves arbitration of exchange rates for different currencies observed at the medieval fairs, around the time of the First Crusade (1100). The dominance of Italian bankers in this era indicates the first usage was the close variant, *arbitrio*, with the French “*arbitrage*” coming into usage during the eighteenth century. Religious and social restrictions effectively barred public discussion of the execution and profitability of such banking activities during the Middle Ages, though account books of the merchant banks do remain as evidence that there was significant arbitrage trading.

As late as the seventeenth century, important English sources on the Law Merchant such as Gerard Malynes, *Lex Mercatoria* [13], make no reference to arbitrage trading strategies in bills of exchange. In contrast, a similar text in Italian, *Il Negotiante* (1638) by Giovanni Peri [18], a seventeenth century Italian merchant, has a detailed discussion on exchange dealings. Peri states that profit is the objective of all trade and that the “activity directed to this end is subject to chance, which mocks at every calculation. Yet there is still ample space for reasonable calculation in which the possibility of adverse fortunes is never left out of account” [5, p.327]. This mental activity engaged in the service of business is called *arbitrio*. Peri identifies a connection between speculation on future exchange rate movements and the *arbitrio* concept of arbitrage: “the profits from exchange dealings originate in price differences and not in time” with profits turning to losses if re-exchange is unfavorable [18, p.150]. For Peri, the connection between speculation and arbitrage applies to commodities and specie, as well as bills of exchange.

The first published usage of “arbitrage” in discussing the relationship between exchange rates and the most profitable locations for issuing and settling a bill of exchange appears in French in, *La Science des Négocians et Teneurs de Livres* [22, p.452]. From the brief reference in a glossary of terms by de la Porte, a number of French sources, including the section *Traité des arbitrages* by Mondoteguy in Le Moine, *Le Negoce d'Amsterdam* [11] and Savary, *Dictionnaire Universel de Commerce* (1730, 2nd ed.) [30], developed a more detailed presentation of arbitrage transactions involving bills of exchange. An important eighteenth century English source, *The Universal Dictionary of Trade and Commerce* [24], is an expanded translation of Savary where the French word “arbitrage” is translated into English as “arbitration”. This is consistent with the linguistic convention of referring to arbitration instead of arbitrage found in the earlier English source, *The Merchant's Public Counting House* [23]. This led to the common English use of the terms “simple arbitrations”, “compound arbitrations”, and “arbitrated rates”. The practice of using arbitration instead of arbitrage continues into nineteenth century works by Patrick Kelly, *The Universal Cambist* [10] and William Tate, *The Modern Cambist* [34]. The latter book went into six editions.

Following the usage of “arbitrage” in German and Dutch works in the 1860s, common usage of “arbitrageur” in English appears with Ottomar Haupt, *The London Arbitrageur* [8], though reference is still made to “arbitration of exchange” as the activity of the arbitrageur. Haupt produced similar works in German and French that used “arbitrage” to describe the calculation of parity relationships. A pamphlet by Maurice Cohn, *The Stock Exchange Arbitrageur* [3] describes “arbitrage transactions” between bourses but also uses “arbitration” to refer to calculated parity relationships. Charles Castelli's *The Theory of “Options” in Stocks and Shares* [1] concludes with a section on “combination of options with arbitrage operations” where arbitrage has exclusive use and no mention is made of “arbitration” of prices or rates across different locations. Following *Arbitrage in Bullion, Coins, Bills, Stocks, Shares and Options* by Henry Deutsch [4], “arbitration of exchange” is no longer commonly used.

References

- [1] Castelli, C. (1877). *The Theory of “Options” in Stocks and Shares*, F. Mathieson, London.
- [2] Chuquet, N. (1484, 1985). Triparty, in *Nicolas Chuquet, Renaissance Mathematician*, G. Flegg, C. Hay & B. Moss, eds, D. Reidel Publishing, Boston.
- [3] Cohn, M. (1874). *The London Stock Exchange in Relation with the Foreign Bourses. The Stock Exchange Arbitrageur*, Effingham Wilson, London.
- [4] Deutsch, H. (1904, 1933). *Arbitrage in Bullion, Coins, Bills, Stocks, Shares and Options*, 3rd Edition, Effingham Wilson, London.
- [5] Ehrenberg, R. (1928). *Capital and Finance in the Age of the Renaissance*, translated from the German by Lucas, H. Jonathan Cape, London.
- [6] Einzig, P. (1964). *The History of Foreign Exchange*, 2nd Edition, Macmillan, London.
- [7] Greif, A. (1989). Reputation and coalitions in medieval trade: evidence on the Maghribi Traders, *Journal of Economic History* **49**, 857–882.
- [8] Haupt, O. (1870). *The London Arbitrageur; or, the English Money Market in connexion with foreign Bourses. A Collection of Notes and Formulae for the Arbitration of Bills, Stocks, Shares, Bullion and Coins, with all the Important Foreign Countries*, Trubner and Co., London.
- [9] Justice, A. (1707). *A General Treatise on Monies and Exchanges; in which those of all Trading Nations are Describ'd and Consider'd*, S. and J. Sprint, London.
- [10] Kelly, P. (1811, 1835). *The Universal Cambist and Commercial Instructor; Being a General Treatise on Exchange including the Monies, Coins, Weights and*

- Measures, of all Trading Nations and Colonies*, 2nd Edition, Lackington, Allan and Co., London, 2 Vols.
- [11] Le Moine de l'Espine, J. (1710). *Le Negoce d'Amsterdam ... Augmenté d'un Traité des arbitrages & des changes sur les principales villes de l'Europe* (by Jacques Mondoteguy), Chez Pierre Brunel, Amsterdam.
- [12] Little, W., Fowler, H. & Coulson, J. (1933, 1958). *Oxford International Dictionary of the English Language*, Leland Publishing, Toronto, revised and edited by C. Onions, 1958.
- [13] Malynes, G. (1622, 1979). *Consuetudo, vel Lex Mercatoria or The Ancient Law Merchant*, Adam Islip, London. reprinted (1979) by Theatrum Orbis Terrarum, Amsterdam.
- [14] McCusker, J. (1978). *Money and Exchange in Europe and America, 1600–1775*, University of North Carolina Press, Chapel Hill NC.
- [15] Munro, J. (2000). English 'Backwardness' and financial innovations in commerce with the low countries, 14th to 16th centuries, in *International Trade in the Low Countries (14th –16th Centuries)*, P. Stabel, B. Blondé, A. Greve, eds, Garant, Leuven-Apeldoorn, pp. 105–167.
- [16] Neal, L. & Quinn, S. (2001). Networks of information, markets, and institutions in the rise of London as a financial centre, 1660–1720, *Financial History Review* 8, 7–26.
- [17] Noonan, J. (1957). *The Scholastic Analysis of Usury*, Harvard University Press, Cambridge, MA.
- [18] Peri, G. (1638, 1707). *Il Negotiante*, Giacomo Hertz, Venice. (last revised edition 1707).
- [19] Poitras, G. (2000). *The Early History of Financial Economics, 1478–1776*, Edward Elgar, Cheltenham, U.K.
- [20] Poitras, G. (2004). William Lowndes, 1652–1724, in *Biographical Dictionary of British Economists*, R. Donald, ed., Thoemmes Press, Bristol, UK, pp. 699–702.
- [21] Poitras, G. (2006). *Pioneers of Financial Economics: Contributions Prior to Irving Fisher*, Edward Elgar, Cheltenham, UK, Vol. I.
- [22] la Porte, M. (1704). *La Science des Négocians et Teneurs de Livres*, Chez Guillaume Chevelier, Paris.
- [23] Postlethwayt, M. (1750). *The Merchant's Public Counting House*, John and Paul Napton, London.
- [24] Postlethwayt, M. (1751, 1774). *The Universal Dictionary of Trade and Commerce*, 4th Edition, John and Paul Napton, London.
- [25] Quinn, S. (1996). Gold, silver and the glorious revolution: arbitrage between bills of exchange and bullion, *Economic History Review* 49, 473–490.
- [26] de Roover, R. (1944). What is dry exchange? A contribution to the study of english mercantilism, *Journal of Political Economy* 52, 250–266.
- [27] de Roover, R. (1948). *Banking and Credit in Medieval Bruges*, Harvard University Press, Cambridge, MA.
- [28] de Roover, R. (1949). *Gresham on Foreign Exchange*, Harvard University Press, Cambridge, MA.
- [29] Sargent, T. & Velde, F. (2002). *The Big Problem of Small Change*, Princeton University Press, Princeton, NJ.
- [30] Savary des Bruslons, J. (1730). *Dictionnaire Universel de Commerce*, Chez Jacques Etienne, Paris, Vol. 3.
- [31] Schubert, E. (1989). Arbitrage in the foreign exchange markets of London and Amsterdam during the 18th Century, *Explorations in Economic History* 26, 1–20.
- [32] Shea, G. (2007). Understanding financial derivatives during the south sea bubble: the case of the south sea subscription shares, *Oxford Economic Papers* 59 (Special Issue), 73–104.
- [33] Weinstein, M. (1931). *Arbitrage in Securities*, Harper & Bros, New York.
- [34] William, T. (1820, 1848). *The Modern Cambist: Forming a Manual of Foreign Exchanges, in the Different Operations of Bills of Exchange and Bullion*, 6th Edition, Effingham Wilson, London.

GEOFFREY POITRAS

Utility Theory: Historical Perspectives

The first recorded mention of a concave utility function in the context of risk and uncertainty is in a manuscript of Daniel Bernoulli [4] in 1738, though credit should also be given to Gabriel Cramer, who, according to Bernoulli himself, developed a remarkably similar theory in 1728. Bernoulli proposes a resolution of a paradox posed in 1713 by his cousin Nicholas Bernoulli. Known as the *St. Petersburg paradox*, it challenges the idea that rational agents value random outcomes by their expected returns. Specifically, a game is envisioned in which a fair coin is tossed repeatedly and the payoff equals 2^n ducats if the first *heads* appeared on the n th toss. The expected value of the payoff can be computed as

$$\begin{aligned} \frac{1}{2} \times 2 + \frac{1}{4} \times 4 + \frac{1}{8} \times 8 + \cdots \\ + \frac{1}{2^n} \times 2^n + \cdots = +\infty \end{aligned} \quad (1)$$

but, clearly, no one would pay an infinite, or even a large finite, amount of money for a chance to play such a game. Daniel Bernoulli suggests that the *satisfaction* or *utility* $U(w)$ from a payoff of size w should not be proportional to w (as mandated by the then prevailing valuation by expectation), but should exhibit diminishing marginal returns; in contemporary language, the derivative U' of the function U should be decreasing (*see Utility Function*). Proposing a logarithmic function as a suitable U , Bernoulli suggests that the *value* of the game to the agent should be calculated as the expected utility

$$\begin{aligned} \frac{1}{2} \times \log(2) + \frac{1}{4} \times \log(4) + \frac{1}{8} \times \log(8) + \cdots \\ + \frac{1}{2^n} \times \log(2^n) + \cdots = \log(4) \end{aligned} \quad (2)$$

Bernoulli's theory was poorly accepted by his contemporaries. It was only a hundred years later that Herman Gossen [11] used Bernoulli's idea of diminishing marginal utility of wealth to formulate his "Laws of Economic Activity". Gossen's "Second law"—the idea that the ratio of exchange values of two goods must equal the ratio of marginal utilities

of the traders—presaged, but did not directly influence, what will become known in economics as the "Marginalist revolution" led by William Jevons [13], Carl Menger [17], and Leon Walras [26].

Axiomatization

The work of Gossen notwithstanding, another century passed before the scientific community took an interest in Bernoulli's ideas (with some notable exceptions such as Alfred Marshall [16] or Francis Edgeworth's entry on probability [8] in the celebrated 1911 edition of *Encyclopedia Britannica*). In 1936, Franz Alt published the first axiomatic treatment of decision making in which he deduces the existence of an implied utility function solely on the basis of a simple set of plausible axioms. Eight years later, Oskar Morgenstern and John von Neumann published the widely influential "Theory of Games and Economic Behavior" [25]. Along with other contributions—the most important representative being a mathematically rigorous foundation of game theory—they develop, at great length, a theory similar to Alt's. Both Alt's and the von Neumann–Morgenstern axiomatizations study a preference relation on the collection of all lotteries (probability distributions on finite sets of outcomes) and show that one lottery is preferred to the other if and only if the expected utility of the former is larger than the expected utility of the latter. The major conceptual leap accomplished by Alt, von Neumann, and Morgenstern was to show that the *behavior* of a rational agent necessarily coincides with the behavior of an agent who values uncertain payoffs using an expected utility.

The Subjectivist Revolution and the State-preference Approach

All of the aforementioned derivations of the expected-utility hypothesis assumed the existence of a physical (objective) probability over the set of possible outcomes of the random payoff. An approach in which both the probability distribution and the utility function are determined jointly from simple behavioral axioms has been proposed by Leonard Savage [23], who was inspired by the work of Frank Ramsey [21] and Bruno de Finetti [5, 6].

One of the major features of the expected-utility theory is the separation between the utility function and the resolution of uncertainty, in that equal payoffs in different states of the world yield the same utilities. It has been argued that, while sometimes useful, such a separation is not necessary. An approach in which the utility of a payoff depends not only on its monetary value but also on the state of the world has been proposed. Such an approach has been popularized through the work of Kenneth Arrow [2] (see **Arrow, Kenneth**) and Gerard Debreu [7], largely because of its versatility and compatibility with general-equilibrium theory where the payoffs are not necessarily monetary. Further successful applications have been made by Roy Radner [20] and many others.

Empirical Paradoxes and Prospect Theory

With the early statistical evidence being mostly anecdotal, many empirical studies have found significant inconsistencies between the observed behavior and the axioms of utility theory. The most influential of these early studies were performed by George Shackle [24], Maurice Allais [1], and Daniel Ellsberg [9]. In 1979, Daniel Kahneman and Amos Tversky [14] proposed “prospect theory” as a psychologically more plausible alternative to the expected utility theory.

Utility in Financial Theory

The general notion of a numerical value associated with a risky payoff was introduced to finance by Harry Markowitz [15] (see **Markowitz, Harry**) through his influential “portfolio theory”.

Markowitz’s work made transparent the need for a precise measurement and quantitative understanding of the levels of “risk aversion” (degree of concavity of the utility function) in financial theory. Even though a similar concept had been studied by Milton Friedman and Leonard Savage [10] before that, the major contribution to this endeavor was made by John Pratt [19] and Kenneth Arrow [3].

With the advent of stochastic calculus (developed by Kiyosi Itô [12], see **Itô, Kiyosi (1915–2008)**), the mathematical tools for continuous-time financial modeling became available. Paul Samuelson [22] (see **Samuelson, Paul A.**) introduced geometric

Brownian motion as a model for stock evolution, and it was not long before it was combined with expected utility theory in the work of Robert Merton [18] (see **Merton, Robert C.**).

References

- [1] Allais, M. (1953). La psychologie de l’homme rationnel devant le risque: critique des postulats et axiomes de l’école Américaine, *Econometrica* **21**(4), 503–546. Translated and reprinted in Allais and Hagen, 1979.
- [2] Arrow, K.J. (1953). Le Rôle des valeurs boursières pour la Répartition la meilleure des risques, *Econométrie*, Colloques Internationaux du Centre National de la Recherche Scientifique, Paris **11**, 41–47; Published in English as (1964). The role of securities in the optimal allocation of risk-bearing, *Review of Economic Studies* **31**(2), 91–96.
- [3] Arrow, K.J. (1965). *Aspects of the Theory of Risk-Bearing*, Yrjö Jahnsson Foundation, Helsinki.
- [4] Bernoulli, D. (1738). Exposition of a new theory on the measurement of risk, *Econometrica* **22**(1), 23–36. Translation from the Latin by Dr. Louise Sommer of work first published 1738.
- [5] de Finetti, B. (1931). Sul significato soggettivo della probabilità, *Fundamenta Mathematicae* **17**, 298–329.
- [6] de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives, *Annales de l’Institut Henri Poincaré* **7**(1), 1–68.
- [7] Debreu, G. (1959). *Theory of Value—An Axiomatic Analysis of Economic Equilibrium*, Cowles Foundation Monograph # 17, Yale University Press.
- [8] Edgeworth, F.Y. (1911). *Probability and Expectation*, Encyclopedia Britannica.
- [9] Ellsberg, D. (1961). Risk, ambiguity and the Savage axioms, *Quarterly Journal of Economics* **75**, 643–69.
- [10] Friedman, M. & Savage, L.P. (1952). The expected-utility hypothesis and the measurability of utility, *Journal of Political Economy* **60**, 463–474.
- [11] Gossen, H.H. (1854). *The Laws of Human Relations and the Rules of Human Action Derived Therefrom*, MIT Press, Cambridge, 1983. Translated from 1854 original by Rudolph C. Blitz with an introductory essay by Nicholas Georgescu-Roegen.
- [12] Itô, K. (1942). On stochastic processes. I. (Infinitely divisible laws of probability), *Japan. Journal of Mathematics* **18**, 261–301.
- [13] Jevons, W.S. (1871). *The Theory of Political Economy*. History of Economic Thought Books, McMaster University Archive for the History of Economic Thought.
- [14] Kahneman, D. & Tversky, A. (1979). Prospect theory: an analysis of decision under risk, *Econometrica* **47**(2), 263–292.
- [15] Markowitz, H. (1952). Portfolio selection, *Journal of Finance* **7**(1), 77–91.

-
- [16] Marshal, A. (1895). *Principles of Economics*, 3rd Edition, 1st Edition 1890, Macmillan, London, New York.
 - [17] Menger, C. (1871). *Principles of Economics*, 1981 edition of 1971 Translation, New York University Press, New York.
 - [18] Merton, R.C. (1969). Lifetime portfolio selection under uncertainty: the continuous-time case, *The Review of Economics and Statistics* **51**, 247–257.
 - [19] Pratt, J. (1964). Risk aversion in the small and in the large, *Econometrica* **32**(1), 122–136.
 - [20] Radner, R. (1972). Existence of equilibrium of plans, prices, and price expectations in a sequence of markets, *Econometrica* **40**(2), 289–303.
 - [21] Ramsey, F.P. (1931). The foundations of mathematics and other logical essays, in *Truth and Probability*, R.B. Braithwaite, ed, Kegan, Paul, Trench, Trubner & Co., Harcourt, Brace and Company, London, New York, Chapter VII, pp. 156–198.
 - [22] Samuelson, P.A. (1965). Rational theory of Warrant Pricing, *Industrial Management Review* **6**(2), 13–31.
 - [23] Savage, L.J. (1954). *The Foundations of Statistics*, John Wiley & Sons Inc., New York.
 - [24] Shackle, G.L.S. (1949). *Expectations in Economics*, Gibson Press.
 - [25] von Neumann, J. & Morgenstern, O. (2007). *Theory of Games and Economic Behavior*, Anniversary Edition. 1st Edition, 1944, Princeton University Press, Princeton, NJ.
 - [26] Walras, L. (1874). *Eléments d'économie Politique Pure*, 4th Edition, L. Corbaz, Lausanne.

Related Articles

Behavioral Portfolio Selection; Expected Utility Maximization; Merton Problem; Risk Aversion; Risk–Return Analysis.

GORDAN ŽITKOVIĆ

Itô, Kiyosi (1915–2008)

Kiyosi Itô was born in 1915, approximately 60 years after the Meiji Restoration. Responding to the appearance of the “Black Ships” in Yokohama harbor and Commodore Perry’s demand that they open their doors, the Japanese overthrew the Tokugawa shogunate and in 1868 “restored” the emperor Meiji to power. The Meiji Restoration initiated a period of rapid change during which Japan made a concerted and remarkably successful effort to transform itself from an isolated, feudal society into a modern state that was ready to play a major role in the world.

During the first phase of this period, they sent their best and brightest abroad to acquire and bring back to Japan the ideas and techniques that had been previously blocked entry by the shogunate’s closed door policy. However, by 1935, the year that Itô entered Tokyo University, the Japanese transformation process had already moved to a second phase, one in which the best and brightest were kept at home to study, assimilate, and eventually disseminate the vast store of information which had been imported during the first phase. Thus, Itô and his peers were expected to choose a topic that they would first teach themselves and then teach their compatriots. For those of us who had the benefit of step-by-step guidance from knowledgeable teachers, it is difficult to imagine how Itô and his fellow students managed, and we can only marvel at the fact that they did.

The topic which Itô chose was that of stochastic processes. At the time, the field of stochastic processes had only recently emerged and was still in its infancy. N. Wiener (1923) had constructed Brownian motion, A.N. Kolmogorov (1933) and Wm. Feller (1936) had laid the analytic foundations on which the theory of diffusions would be built, and P. Lévy (1937) had given a pathspace interpretation of infinitely divisible laws. However, in comparison to well-established fields such as complex analysis, stochastic processes still looked more like a haphazard collection of examples than a unified field.

Having studied mechanics, Itô from the outset was drawn to Lévy’s pathspace perspective with its emphasis on paths and dynamics, and he set as his goal the reconciliation of Kolmogorov and Feller’s analytic treatment with Lévy’s pathspace picture. To carry out his program, he first had to thoroughly

understand Lévy, and, as anyone who has attempted to read Lévy in the original knows, this in itself a daunting task. Indeed, I have my doubts that, even now, many of us would know what Lévy did had Itô not explained it to us. Be that as it may, Itô’s first published paper (1941) was devoted to a reworking (incorporating important ideas due to J.L. Doob) of Lévy’s theory of homogeneous, independent increment processes.

Undoubtedly as a dividend of the time and effort which he spent unraveling Lévy’s ideas, shortly after completing this paper Itô had a wonderful insight of his own. To explain his insight, imagine that the space $M_1(\mathbb{R})$ of probability measures on $\mathbb{R}$ has a differentiable structure in which the underlying dynamics is given by convolution. Then, if $t \in [0, \infty) \mapsto \mu_t \in M_1(\mathbb{R})$ is a “smooth curve” which starts at the unit point mass δ_0 , its “tangent” at time 0, it should be given by the limit

$$\lim_{n \rightarrow \infty} \left(\mu \frac{1}{n} \right)^{\star n}$$

where $\star$ denotes convolution and therefore $\nu^{\star n}$ is the n -fold convolution power of $\nu \in M_1(\mathbb{R})$. What Itô realized is that, if this limit exists, it must be an infinitely divisible law. Applied to $\mu_t = P(t, x, \cdot)$, where $(t, x) \in [0, \infty) \times \mathbb{R} \mapsto P(t, x, \cdot) \in M_1(\mathbb{R})$ is the transition probability function for a Markov process, this key observation lead Itô to view Kolmogorov’s forward equation as describing the flow of a vector field on $M_1(\mathbb{R})$. In addition, because infinitely divisible laws play in the geometry of $M_1(\mathbb{R})$ the role^a that straight lines play in Euclidean space, he saw that one should be able to “integrate” Kolmogorov’s equation by piecing together infinitely divisible laws, just as one integrates a vector field in Euclidean space by piecing together straight lines.

Profound as the preceding idea is, Itô went a step further. Again under Lévy’s influence, he wanted to transfer his idea to a pathspace setting. Reasoning that if the transition function can be obtained by concatenating infinitely divisible laws, then the paths of the associated stochastic processes must be obtainable to concatenating paths coming from Lévy’s independent increment processes and that one should be able to encode this concatenation procedure in some sort of “differential equation” for the resulting paths. The implementation of this program required him to develop what is now called the “Itô calculus”.

It was during the period when he was working out the details of his calculus that he realized, at least in the special case when paths are continuous, there is a formula which plays role in his calculus that the chain rule plays in Newton's. This formula, which appeared for the first time in a footnote, is what we now call *Itô's formula*. Humble as its origins may have been, it has become one of the three or four most famous mathematics formulae of the twentieth century. Itô's formula is not only a boon of unquestioned and inestimable value to mathematicians but also has become an indispensable tool in the world of mathematically oriented finance.

Itô had these ideas in the early 1940s, around the time when Japan attacked Pearl Harbor and its population had to face the consequent horrors. In view of the circumstances, it is not surprising that few inside Japan, and nobody outside of Japan, knew what Itô was doing for nearly a decade. Itô did publish an outline of his program in a journal of mimeographed notes (1942) at Osaka University, but he says that only his friend G. Maruyama really read what he had written. Thus, it was not until 1950, when he sent the manuscript for a monograph to Doob who arranged that it be published by the A.M.S. as a Memoir, that Itô's work began to receive the attention which it deserved. Full appreciation of Itô's ideas by the mathematical community came only after first Doob and then H.P. McKean applied martingale theory to greatly simplify some of Itô's more technical arguments.

Despite its less than auspicious beginning, the story has a happy ending. Itô spent many years traveling the world: he has three daughters, one living in Japan, one in Denmark, and one in America. He is, in large part, responsible for the position of Japan as a major force in probability theory, and he has disciples all over the planet. His accomplishments are widely recognized: he is a member of the Japanese Academy of Sciences and the National Academy of Sciences; and he is the recipient of, among others, the Kyoto, Wolf, and Gauss Prizes. When I think of Itô's career and the rocky road that he had to travel, I recall what Jack Schwartz told a topology class I was attending about Jean Leray's invention of spectral sequences. At the time, Leray was a prisoner

in a German prison camp for French intellectuals, each of whom attempted to explain to the others something about which he was thinking. With the objective of not discussing anything that might be useful to the enemy, Leray chose to talk about algebraic topology rather than his own work on partial differential equations, and for this purpose, he introduced spectral sequences as a pedagogic tool. After relating this anecdote, Schwartz leaned back against the blackboard and spent several minutes musing about the advantages of doing research in ideal working conditions.

Kiyosi Itô died at the age of 93 on November 10, 2008. He is survived by his three daughters. A week before his death, he received the Cultural Medal from the Japanese emperor. The end of an era is fast approaching.

End Notes

^aNote that when $t \rightsquigarrow \mu_t$ is the flow of infinitely divisible law μ in the sense that $\mu_1 = \mu$ and $\mu_{s+t} = \mu_s \star \mu_t$, $\mu = (\mu_{(1/n)})^{\star n}$ for all $n \geq 1$, which is the convolution analog of $f(1) = n^{-1} f(n)$ for a linear function on $\mathbb{R}$.

References

- [1] Stroock, D. & Varadhan S.R.S. (eds) (1986). *Selected Papers: K. Itô*, Springer-Verlag.
- [2] Stroock, D. (2003). *Markov Processes from K. Itô's Perspective*, Annals of Mathematical Studies, Vol. 155, Princeton University Press.
- [3] Stroock, D. (2007). *The Japanese Journal of Mathematical Studies* 2(1).

Further Reading

A selection of Itô's papers as well as an essay about his life can be found in [1]. The first half of the book [2] provides a lengthy exposition of Itô's ideas about Markov processes. Reference [3] is devoted to articles, by several mathematicians, about Itô and his work. In addition, thumbnail biographies can be found on the web at www-groups.dcs.st-and.ac.uk/history/Biographies/Ito.html and www.math.uah.edu/stat/biographies/Ito.xhtml

DANIEL W. STROOCK

Thorp, Edward

Edward O. Thorp is a mathematician who has made seminal contributions to games of chance and investment science. He invented original strategies for the game of blackjack that revolutionized the game. Together with Sheen Kassouf, he showed how warrants could be hedged using a short position in the underlying stocks and described and implemented arbitrage portfolios of stocks and warrants. Thorp made other important contributions to the development of option pricing and to investment theory and practice. He has had a very successful record as an investment manager. This note contains a brief account of some of his major contributions.

Thorp studied physics as an undergraduate and obtained his PhD in mathematics from the University of California at Los Angeles in 1958. The title of his dissertation was *Compact Linear Operators in Normed Spaces*, and he has published several papers on functional analysis. He taught at UCLA, MIT, and New Mexico State University and was professor of mathematics and finance at the University of California at Irvine.

Thorp's interest in devising scientific systems for playing games of chance began when he was a graduate student in the late 1950s. He invented a system for playing roulette and also became interested in blackjack and devised strategies based on card counting systems. While at MIT, he collaborated with Claude Shannon, and together they developed strategies for improving the odds at roulette and blackjack. One of their inventions was a wearable computer that was the size of modern-day cell phone. In 1962, Thorp [3] published *Beat the Dealer: A Winning Strategy for the Game of Twenty One*. This book had a profound impact on the game of blackjack as gamblers tried to implement his methods, and casinos responded with various countermeasures that were sometimes less than gentle.

In June 1965, Thorp's interest in warrants was piqued by reading Sydney Fried's *RHM Warrant Survey*. He was motivated by the intellectual challenge of warrant valuation and by the prospect of making money using these instruments. He developed his initial ideas on warrant pricing and investing during the summer of 1965. Sheen Kassouf who was, like Thorp, a new faculty member at the University of California's newly established campus at Irvine, was

also interested in warrants because of his own investing. Kassouf had analyzed market data to determine the key variables that affected warrant prices. On the basis of his analysis, Kassouf developed an empirical formula for a warrant's price in terms of these variables.

In September 1965, Thorp and Kassouf discovered their mutual interest in warrant pricing and began their collaboration. In 1967, they published their book, *Beat the Market*, in which they proposed a method for hedging warrants using the underlying stock and developed a formula for the hedge ratio [5]. Their insights on warrant pricing were used^a by Black and Scholes in their landmark 1973 paper on option pricing.

Thorp and Kassouf were aware that the conventional valuation method was based on projecting the warrant's expected terminal payoff and discounting back to current time. This approach involved two troublesome parameters: the expected return on the warrant and the appropriate discount rate. Black and Scholes in their seminal paper would show that the values of both these parameters had to coincide with the riskless rate. There is strong evidence^b that Thorp independently discovered this solution in 1967 and used it in his personal investment strategies. Thorp^c makes it quite clear that the credit rightfully belongs to Black and Scholes.

Black Scholes was a watershed. It was only after seeing their proof that I was certain that this was the formula—and they justifiably get all the credit. They did two things that are required. They proved the formula(I didn't) and they published it (I didn't).

Thorp made a number of other contributions to the development of option theory and modern finance and his ideas laid the foundations for further advances. As one illustration based on my own experience, I will mention Thorp's essential contribution to a paper that David Emanuel and I published in 1980 [2]. Our paper examined the distribution of a hedged portfolio of a stock and option that was rebalanced after a short interval. The key equation on which our paper rests was first developed by Thorp in (1976) [4].

Throughout his career, Edward Thorp has applied mathematical tools to develop highly original solutions to difficult problems and he has demonstrated a unique ability to implement these solution in a practical way.

End Notes

^aBlack and Scholes state, “One of the concepts we use in developing our model was expressed by Thorp and Kassouf.”

^bFor a more detailed discussion of this issue, see Boyle and Boyle [1] Chapter Five.

^cEmail to the author dated July 26, 2000.

References

[1] Boyle, P.P. & Boyle, F.P. (2001). *Derivatives: the Tools that Changed Finance*, Risk Books, UK.

- [2] Boyle, P.P. & Emanuel, D. (1980). Discretely adjusted option hedges, *Journal of Financial Economics* **8**(3), 259–282.
- [3] Thorp, E.O. (1962). *Beat the Dealer: A Winning Strategy for the Game of Twenty-One*, Random House, New York.
- [4] Thorp, E.O. (1976). Common stock volatilities in option formulas, *Proceedings, Seminar on the Analysis of Security Prices*, Center for Research in Security Prices, Graduate School of Business, University of Chicago, Vol. 21, 1, May 13–14, pp. 235–276.
- [5] Thorp, E.O. & Kassouf, S. (1967). *Beat the Market: A Scientific Stock Market System*, Random House, New York.

PHELIM BOYLE

Option Pricing Theory: Historical Perspectives

This article traces the history of the option pricing theory from the turn of the twentieth century to the present. This history documents and clarifies the origins of the key contributions (authors and papers) to the theory of option pricing and hedging. Contributions with respect to the empirical understanding of the theories are not discussed, except implicitly, because the usefulness and longevity of any model is based on its empirical validity.

It is widely agreed that the modern theory of option pricing began in 1973 with the publication of the Black–Scholes–Merton model [12, 104]. Except for the early years (pre-1973), this history is restricted to papers that use the no arbitrage and complete markets technology to price options. Equilibrium option pricing models are not discussed herein. In particular, this excludes the consideration of option pricing in incomplete markets. An outline for this article is as follows. The following section discusses the early years of option pricing (pre-1973). The remaining sections deal with 1973 to the present: the section “Equity Derivatives” discusses the Black–Scholes–Merton model; the section “Interest Rate Derivatives” concerns the Heath–Jarrow–Morton model; and the section “Credit Derivatives” corresponds to credit risk derivative pricing models.

Early Option Pricing Literature (Pre-1973)

Interestingly, many of the basic insights of option pricing originated in the early years, that is, pre-1973. It all began at the turn of the century in 1900 with Bachelier’s [4] derivation of an option pricing formula in his doctoral dissertation on the theory of speculation at France’s Sorbonne University. Although remarkably close to the Black–Scholes–Merton model, Bachelier’s formula was flawed because he used normally distributed stock prices that violated limited liability. More than half a century later, Paul Samuelson read Bachelier’s dissertation, recognized this flaw, and fixed it by using geometric Brownian motion instead in his

work on warrant pricing [117]. Samuelson derived valuation formulas for both European and American options, coining these terms in the process.

Samuelson’s derivation was almost identical to that used nearly a decade later to derive the Black–Scholes–Merton formula, except that instead of invoking the no arbitrage principle to derive the valuation formula, Samuelson postulated the condition that the discounted option’s payoffs follow a martingale (see [117], p. 19). Furthermore, it is also interesting to note that, in the appendix to this article, Samuelson and McKean determined the price of an American option by observing the correspondence between an American option’s valuation and the free boundary problem for the heat equation.

A few years later, instead of invoking the postulate that discounted option payoffs follow a martingale, Samuelson and Merton [118] derived this condition as an implication of a utility maximizing investor’s behavior. In this article, they also showed that the option’s price could be viewed as its discounted expected value, where instead of using the actual probabilities to compute the expectation, one employs utility or risk-adjusted probabilities (see expression (20) on page 26). These risk-adjusted probabilities are now known as “risk-neutral” or “equivalent martingale” probabilities. Contrary to a widely held belief, the use of “equivalent martingale probabilities” in option pricing theory predated the paper by Cox and Ross [36] by nearly 10 years (Merton (footnote 5 p. 218, [107]) points out that Samuelson knew this fact as early as 1953).

Unfortunately, these early option pricing formulas depended on the expected return on the stock, or equivalently, the stock’s risk premium. This dependency made the formulas difficult to estimate and to use. The reason for this difficulty is that the empirical finance literature has documented that the stock’s risk premium is nonstationary. It varies across time according to both changing tastes and changing economic fundamentals. This nonstationarity makes both the modeling of risk premium and their estimation problematic. Indeed, at present, there is still no generally accepted model for an asset’s risk premium that is consistent with historical data (see [32], Part IV for a review).

Perhaps the most important criticism of this early approach to option pricing is that it did not invoke the riskless hedging argument in conjunction with the no-arbitrage principle to price an option. (The first use of

riskless hedging with no arbitrage to prove a pricing relationship between financial securities can be found in [110].) And, as such, these valuation formulas provided no insights into how to hedge an option using the underlying stock and riskless borrowing. It can be argued that the idea of hedging an option is the single most important insight of modern option pricing theory. The use of the no arbitrage hedging argument to price an option can be traced to the seminal papers by Black and Scholes [12] and Merton [104], although the no arbitrage hedging argument itself has been attributed to Merton (see [79] in this regard).

Equity Derivatives

Fischer Black, Myron Scholes, and Robert Merton pioneered the modern theory of option pricing with the publication of the Black–Scholes–Merton option pricing model [12, 104] in 1973. The original Black–Scholes–Merton model is based on five assumptions: (i) competitive markets, (ii) frictionless markets, (iii) geometric Brownian motion, (iv) deterministic interest rates, and (v) no credit risk. For the purposes of this section, the defining characteristics of this model are the assumptions of deterministic interest rates and no credit risk.

The original derivation followed an economic hedging argument. The hedging argument involves holding simultaneous and offsetting positions in a stock and option that generates an instantaneous riskless position. This, in turn, implies a partial differential equation (pde.) for the option's value that is subject to a set of boundary conditions. The solution under geometric Brownian motion is the Black–Scholes formula.

It was not until six years later that the martingale pricing technology was introduced by Harrison and Kreps [65] and Harrison and Pliska [66, 67], providing an alternative derivation of the Black–Scholes–Merton model. These papers, and later refinements by Delbaen and Schachermayer [40, 41, 42], introduced the first and second fundamental theorems of asset pricing, thereby providing the rigorous foundations to option pricing theory.

Roughly speaking, the first fundamental theorem of asset pricing states that no arbitrage is equivalent to the existence of an equivalent martingale probability measure, that is, a probability measure that makes

the discounted stock price process a martingale. The second fundamental theorem of asset pricing states that the market is complete if and only if the equivalent martingale measure is unique. A complete market is one in which any derivative security's payoffs can be generated by a dynamic trading strategy in the stock and riskless asset. These two theorems enabled the full fledged use of stochastic calculus for option pricing theory. A review and summary of these results can be found in [43].

At the beginning, this alternative and more formal approach to option pricing theory was viewed as only of tangential interest. Indeed, all existing option pricing theorems could be derived without this technology and only using the more intuitive economic hedging argument. It was not until the Heath–Jarrow–Morton (HJM) model [70] was developed—circulating as a working paper in 1987—that this impression changed. The HJM model was the first significant application that could not be derived without the use of the martingale pricing technology. More discussion relating to the HJM model is contained in the section “Interest Rate Derivatives”.

Extensions

The original Black–Scholes–Merton model is based on the following five assumptions: (i) competitive markets, (ii) frictionless markets, (iii) geometric Brownian motion, (iv) deterministic interest rates, and (v) no credit risk. The first two assumptions—competitive and frictionless markets—are the mainstay of finance. Competitive markets means that all traders act as price takers, believing their trades have no impact on the market price. Frictionless markets imply that there are no transaction costs nor trade restrictions, for example, no short sale constraints. Geometric Brownian motion implies that the stock price is lognormally distributed with a constant volatility. Deterministic interest rates are self-explanatory. No credit risk means that the investors (all counterparties) who trade financial securities will not default on their obligations.

Extensions of the Black–Scholes–Merton model that relaxed assumptions (i)–(iii) quickly flourished. Significant papers relaxing the geometric Brownian motion assumption include those by Merton [106] and Cox and Ross [36], who studied jump and jump-diffusion processes. Merton's paper [106] also

included the insight that if unhedgeable jump risk is diversifiable, then it carries no risk premium. Under this assumption, one can value jump risk using the statistical probability measure, enabling the simple pricing of options in an incomplete market. This insight was subsequently invoked in the context of stochastic volatility option pricing and in the context of pricing credit risk derivatives.

Merton [104], Cox [34] and Cox and Ross [36] were among the first to study stochastic volatility option pricing in a complete market. Option pricing with stochastic volatility in incomplete markets was subsequently studied by Hull and White [73] and Heston [71]. More recent developments in this line of research use a HJM [70] type model with a term structure of forward volatilities (see [51, 52]). Stochastic volatility models are of considerable current interest in the pricing of volatility swaps, variance swaps, and options on variance swaps.

A new class of Levy processes was introduced by Madan and Milne [102] into option pricing and generalized by Carr *et al.* [20]. Levy processes have the nice property that their characteristic function is known, and it can be shown that an option's price can be represented in terms of the stock price's characteristic function. This leads to some alternative numerical procedures for computing option values using fast Fourier transforms (see [23]). For a survey of the use of Levy processes in option pricing, see [33].

The relaxation of the frictionless market assumption has received less attention in the literature. The inclusion of transaction costs into option pricing was originally studied by Leland [99], while Heath and Jarrow [69] studied the imposition of margin requirements. A more recent investigation into the impact of transaction costs on option pricing, using the martingale pricing technology, can be found in [26].

The relaxation of the competitive market assumption was first studied by Jarrow [77, 78] *via* the consideration of a large trader whose trades change the price. Jarrow's approach maintains the no arbitrage assumption, or in this context, a no market manipulation assumption (see also [5]).

In between a market with competitive traders and a market with a large trader is a market where traders have only a temporary impact on the market price. That is, purchase/sales change the price paid/received depending upon a given supply curve. Traders act as price takers with respect to the supply curve. Such a

price impact is called *liquidity risk*. Liquidity risk, of this type, can be considered as an endogenous transaction cost. This extension is studied in [26]. Liquidity risk is currently a hot research topic in option pricing theory.

The Black–Scholes–Merton model has been applied to foreign currency options (see [58]) and to all types of exotic options on both equities and foreign currencies. A complete reference for exotic options is [44].

Computations

The original derivation of the Black–Scholes–Merton model yields an option's value satisfying a pde, subject to a set of boundary conditions. For a European call or put option, under geometric Brownian motion, the pde has an analytic solution. For American options under geometric Brownian motion, analytic solutions are not available for puts independent of dividend payments on the underlying stock, and for American calls with dividends. For different stock price processes, analytic solutions are often not available as well, even for European options. In these cases, numerical solutions are needed. The first numerical approaches employed in this regard were finite difference methods (see [15, 16]).

Closely related, but containing more economic intuition, option prices can also be computed numerically by using a binomial approximation. The first users in this regard were Sharpe [122] chapter 16, and Rendleman and Barter [113]. Cox *et al.* [37] published the definitive paper documenting the binomial model and its convergence to the continuous time limit (see also [68]). A related paper on convergence of discrete time models to continuous time models is that by Duffie and Protter [48].

The binomial pricing model, as it is now known, is also an extremely useful pedagogical device for explaining option pricing theory. This is true because the binomial model uses only discrete time mathematics. As such, it is usually the first model presented in standard option pricing textbooks. It is interesting to note that both the first two textbooks on option pricing utilized the binomial model in this fashion (see [38] and [84]).

Another technique for computing option values is to use a series expansions (see [50, 83 and 123]). Series expansions are also useful for hedging exotic options that employ only static hedge positions with

plain vanilla options (see [38] chapter 7.2, [24, 63, and 116]).

As computing a European option's price is equivalent to computing an expectation, an alternative approach to either finite difference methods or the binomial model is Monte Carlo simulation. The paper that introduced this technique to option pricing is by Boyle [13]. This technique has become very popular because of its simplicity and its ability to handle high-dimensional problems (greater than three dimensions). This technique has also recently been extended to pricing American options. Important contributions in this regard are by Longstaff and Schwartz [101] and Broadie and Glasserman [18]. For a complete reference on Monte Carlo techniques, see [61].

Following the publication of Merton's original paper [104], which contained an analytic solution for a perpetual American put option, much energy has been expended in the search for analytic solutions for both American puts and calls with finite maturities. For the American call, with a finite number of known dividends, a solution was provided by Roll [115]. For American puts, breaking the maturity of the option into a finite number of discrete intervals, the compound option pricing technique is applicable, (see [60] and [93]). More recently, the decomposition of American options into a European option and an early exercise premium was discovered by Carr *et al.* [22], Kim [96], and Jacka [75].

These computational procedures are more generally applicable to all derivative pricing models, including those discussed in the next two sections.

Interest Rate Derivatives

Interest rate derivative pricing models provided the next major advance in option pricing theory. Recall that a defining characteristic of the Black–Scholes–Merton model is that it assumes deterministic interest rates. This assumption limits its usefulness in two ways. First, it cannot be used for long-dated contracts. Indeed, for long-dated contracts (greater than a year or two), interest rates cannot be approximated as being deterministic. Second, for short-dated contracts, if the underlying asset's price process is highly correlated with interest rate movements, then interest rate risk will affect hedging, and therefore valuation. The extreme cases, of course, are interest rate derivatives where the underlyings are the interest rates themselves.

During the late 1970s and 1980s, interest rates were large and volatile, relative to historical norms. New interest rate risk management tools were needed because the Black–Scholes–Merton model was not useful in this regard. In response, a class of interest rate pricing models were developed by Vasicek [124], Brennan and Schwartz [17], and Cox *et al.* (CIR) [35]. This class, called the *spot rate models*, had two limitations. First, they depended on the market price(s) of interest rate risk, or equivalently, the expected return on default free bonds. This dependence, just as with the option pricing models pre-Black–Scholes–Merton, made their implementation problematic. Second, these models could not easily match the initial yield curve. This calibration is essential for the accurate pricing and hedging of interest rate derivatives because any discrepancies in yield curve matching may indicate “false” arbitrage opportunities in the priced derivatives.

To address these problems, Ho and Lee [72] applied the binomial model to interest rate derivatives with a twist. Instead of imposing an evolution on the spot rate, they had the zero coupon bond price curve that evolved in a binomial tree. Motivated by this paper, Heath–Jarrow–Morton [70] generalized this idea in the context of a continuous time and multifactor model to price interest rate derivatives. The key step in the derivation of the HJM model was determined as the necessary and sufficient conditions for an arbitrage free evolution of the term structure of interest rates.

The defining characteristic of the HJM model is that there is a continuum of underlying assets, a term structure, whose correlated evolution needs to be considered when pricing and hedging options. For interest rate derivatives, this term structure is the term structure of interest rates. To be specific, it is the term structure of default free interest rates. But there are other term structures of relevance, including foreign interest rates, commodity futures prices, convenience yields on commodities, and equity forward volatilities. These alternative applications are discussed later in this section.

To simplify the mathematics, HJM focused on forward rates instead of zero-coupon bond prices. The martingale pricing technology was the tool used to obtain the desired conditions —the “HJM drift conditions”. Given the HJM drift conditions and the fact that the interest rate derivative market is

complete in the HJM model, standard techniques are then applied to price interest rate derivatives.

The HJM model is very general: all previous spot rate models are special cases. In fact, the labels Vasicek, extended Vasicek (or sometimes Hull and White [74]), and CIR are now exclusively used to identify subclasses of the HJM model. Subclasses are uniquely identified by a particular volatility structure for the evolution of forward rate curve. For example, the Ho and Lee model is now identified as a single factor HJM model, where the forward rate volatility is a constant across maturities. This can be shown to be the term structure evolution to which the Ho and Lee binomial model converges.

Adoption of the HJM model was slow at first, hampered mostly by computational concerns, but as these computational concerns dissipated, the modern era for pricing interest rate derivatives was born. As mentioned previously, the HJM model was very general. In its most unrestricted form, the evolution of the term structure of interest rates could be path dependent (non-Markov) and it could generate negative interest rates with positive probability. Research into the HJM model proceeded in two directions: (i) investigations into the abstract mathematical structure of HJM models and (ii) studying subclasses that had nice analytic and computational properties for applications.

With respect to the understanding of the mathematical structure of HJM models, three questions arose. First, what structures would guarantee interest rates that remained positive? Second, given an initial forward rate curve and its evolution, what is the class of forward rate curves that can be generated by all possible evolutions? Third, under what conditions is an HJM model a finite dimensional Markov process? The first question was answered by Flesaker and Hughston [55], Rogers [114], and Jin and Glasserman [91]. The second was solved by Bjork and Christensen [7] and Filipovic [56]. The third was studied by Cheyette [30], Caverhill [25], Jeffrey [92], Duffie and Kan [45], and Bjork and Svensson [9], among others.

The original HJM model had the term structure of interest rates generated by a finite number of Brownian motions. Extensions include (i) jump processes (see [8, 53 and 82]); (ii) stochastic volatilities (see [1, 31]); and (iii) random fields (see [64, 95]).

Subclasses

Subsequent research developed special cases of the HJM model that have nice analytic and computational properties for implementation. Perhaps the most useful class, for its analytic properties, is the affine model of Duffie and Kan [45] and Dai and Singleton [39]. The class of models is called *affine* because the spot rate can be written as an affine function of a given set of state variables. The affine class includes both the Vasicek and CIR models as mentioned earlier. This class of term structure evolutions have known characteristic functions for the spot rate, which enables numerical computations for various interest rate derivatives (see [47]). Extensions of the affine class include those by Filipovic [57], Chen *et al.* [28], and Cheng and Scaillet [29].

The original HJM paper showed that instantaneous forward rates being lognormally distributed is inconsistent with no arbitrage. Hence, geometric Brownian motion was excluded as an acceptable forward rate process. This was unfortunate because it implies that caplets, options on forward rates, will not satisfy Black's formula [10]. And historically, because of the industry's familiarity with the Black–Scholes formula (a close relative of Black's formula), Black's formula was used extensively to value caplets. This inconsistency between theory and practice lead to a search for a theoretical justification for using Black's formula with caplets.

This problem was resolved by Sandmann *et al.* [119], Miltersen *et al.* [109], and Brace *et al.* [14]. The solution was to use a simple interest rate, compounded discretely, for the London Interbank Offer Rate (LIBOR). Of course, simple rates better match practice. And it was shown that the evolution of a simple LIBOR could evolve as a geometric Brownian motion in an arbitrage free setting. Subsequently, the lognormal evolution has been extended to jump diffusions (see [62]), Levy processes (see [54]), and stochastic volatilities (see [1]).

Key to the use of the “LIBOR model”, as it has become known, is the forward price martingale measure. The forward price martingale measure is an equivalent probability measure that makes asset payoffs at some future date T martingales when discounted by the T maturity zero coupon bond price. The forward price martingale measure

was first discovered by Jarrow [76] and later independently discovered by Geman [59] (see [112] for a discussion of the LIBOR model and its history).

Applications

The HJM model has been extended to multiple term structures and applied to foreign currency derivatives [2], to equities and commodities [3], and to Treasury inflation protected bonds [89]. The HJM model has also been applied to term structures of futures prices (see [21], and [108]), term structures of convenience yields [111], term structures of credit risky bonds (discussed in the next section), and term structures of equity forward volatilities ([51, 52], and [121]). In fact, it can be shown that almost all option pricing applications can be viewed as special cases of a multiple term structure HJM model (see [88]). A summary of many of these applications can be found in [19].

Credit Derivatives

The previously discussed models excluded the consideration of default when trading financial securities. The first model for studying credit risk, called the *structural approach*, was introduced by Merton [105]. Credit risk, although always an important consideration in fixed income markets, dramatically expanded its market wide recognition with the introduction of trading in credit default swaps after the mid-1990s. The reason for this delayed importance was that it took until then for the interest rate derivative markets to mature sufficiently for sophisticated financial institutions to successfully manage/hedge equity, foreign currency and interest rate risk. This risk-controlling ability enabled firms to seek out arbitrage opportunities, and in the process, lever up on the remaining financial risks, which are credit/counterparty, liquidity, and operational risk. This greater risk exposure by financial institutions to both credit and liquidity risk (as evidenced by the events surrounding the failure of Long Term Capital Management) spurred the more rapid development of credit risk modeling.

As the first serious contribution to credit risk modeling, Merton's original model was purposely simple. Merton considered credit risk in the context

of a firm issuing only a single zero coupon bond. As such, risky debt could be decomposed into riskless debt plus a short put option on the assets of the firm. Shortly thereafter, extensions to address this simple liability structure were quickly discovered by Black and Cox [11] Jones *et al.* [94] and Leland [100] among others.

The structural approach to credit risk modeling has two well-known empirical shortcomings: (i) that default occurs smoothly, implying that bond prices do not jump at default and (ii) that the firm's assets are neither traded nor observable. The first shortcoming means that for short maturity bonds, credit spreads as implied by the structural model are smaller than those observed in practice. Extensions of the structural approach that address the absence of a jump at default include that by Zhou [125]. These extensions, however, did not overcome the second shortcoming.

Almost 20 years after Merton's original paper, Jarrow and Turnbull [85, 86] developed an alternative credit risk model that overcame the second shortcoming. As a corollary, this approach also overcame the first shortcoming. This alternative approach has become known as the *reduced form model*. Early important contributions to the reduced form model were by Lando [97], Madan and Unal [103], Jarrow *et al.* [80], and Duffie and Singleton [49].

As the credit derivative markets expanded, so did extensions to the reduced form model. To consider credit rating migration, Jarrow *et al.* [80] introduced a Markov chain model, where the states correspond to credit ratings. Next, there was the issue of default correlation for pricing credit derivatives on baskets (e.g., credit default obligations (CDOs)). This correlation was first handled with Cox processes (Lando [97]).

The use of Cox processes induces default correlations across firms through common state variables that drive the default intensities. But when conditioning on the state variables, defaults are assumed to be independent across firms. If this structure is true, then after conditioning, defaults are diversifiable in a large portfolio and require no additional risk premium. The implication is that the empirical and risk neutral default intensities are equal. This equality, of course, would considerably simplify direct estimation of the risk neutral default intensity [81].

This is not the only mechanism through which default correlations can be generated. Default contagion is also possible through competitive industry considerations. This type of default contagion is a type of “counterparty” risk, and it was first studied in the context of a reduced form model by Jarrow and Yu [90]. “Counterparty risk” in a reduced form model, an issue in and of itself, was previously studied by Jarrow and Turnbull [86, 87].

Finally, default correlation could be induced *via* information flows as well. Indeed, a default by one firm may cause other firm’s default intensities to increase as the market learns about the reasons for the realized default (see [120]). Finding a suitable correlation structure for implementation and estimation is still a topic of considerable interest.

An important contribution to the credit risk model literature was the integration of structural and reduced form models. These two credit risk models can be understood through the information sets used in their construction. Structural models use the management’s information set, while reduced form models use the market’s information set. Indeed, the manager has access to the firm’s asset values, while the market does not. The first paper making this connection was by Duffie and Lando [46] who viewed the market as having the management’s information set plus noise, due to the accounting process. An alternative view is that the market has a coarser partitioning of management’s information, that is, less of it. Both views are reasonable, but the mathematics is quite different. The second approach was first explored by Cetin *et al.* [27].

Credit risk modeling continues to be a hot area of research. Books on the current state of the art with respect to credit risk derivative pricing models are by Lando [98] and Bielecki and Rutkowski [6].

References

- [1] Andersen, L. & Brotherton-Ratcliffe, R. (2005). Extended LIBOR market models with stochastic volatility, *Journal of Computational Finance* **9**, 1–26.
- [2] Amin, K. & Jarrow, R. (1991). Pricing foreign currency options under stochastic interest rates, *Journal of International Money and Finance* **10**(3), 310–329.
- [3] Amin, K. & Jarrow, R. (1992). Pricing American options on risky assets in a stochastic interest rate economy, *Mathematical Finance* **2**(4), 217–237.
- [4] Bachelier, L. (1990). *Theorie de la Speculation*, Ph.D. Dissertation, L’Ecole Normale Supérieure. English translation in P. Cootner (ed.) (1964) *The Random Character of Stock Market Prices*, MIT Press, Cambridge, MA.
- [5] Bank, P. & Baum, D. (2004). Hedging and Portfolio optimization in illiquid Financial markets with a large trader, *Mathematical Finance* **14**(1), 1–18.
- [6] Bielecki, T. & Rutkowski, M. (2002). *Credit Risk: Modeling, Valuation, and Hedging*, Springer Verlag.
- [7] Bjork, T. & Christensen, B. (1999). Interest rate dynamics and consistent forward rate curves, *Mathematical Finance* **9**(4), 323–348.
- [8] Bjork, T., Di Masi, G., Kabanov, Y. & Runggaldier, W. (1997). Towards a general theory of bond markets, *Finance and Stochastics* **1**, 141–174.
- [9] Bjork, T. & Svensson, L. (2001). On the existence of finite dimensional realizations for nonLinear forward rate models, *Mathematical Finance* **11**(2), 205–243.
- [10] Black, F. (1976). The pricing of commodity contracts, *Journal of Financial Economics* **3**, 167–179.
- [11] Black, F. & Cox, J. (1976). Valuing corporate securities: some effects of bond indenture provisions, *Journal of Finance* **31**, 351–367.
- [12] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- [13] Boyle, P. (1977). Options: a Monte Carlo approach, *Journal of Financial Economics* **4**, 323–338.
- [14] Brace, A., Gatarek, D. & Musiela, M. (1997). The market model of interest rate dynamics, *Mathematical Finance* **7**(2), 127–147.
- [15] Brennan, M. & Schwartz, E. (1977). The valuation of American put options, *Journal of Finance* **32**, 449–462.
- [16] Brennan, M. & Schwartz, E. (1978). Finite difference methods and jump processes arising in the pricing of contingent claims: a synthesis, *Journal of Financial and Quantitative Analysis* **13**, 461–474.
- [17] Brennan, M. & Schwartz, E. (1979). A continuous time approach to the pricing of bonds, *Journal of Banking and Finance* **3**, 135–155.
- [18] Broadie, M. & Glasserman, P. (1997). Pricing American style securities by simulation, *Journal of Economic Dynamics and Control* **21**, 1323–1352.
- [19] Carmona, R. (2007). HJM: a unified approach to dynamic models for fixed income, credit and equity markets. *Paris-Princeton Lectures on Mathematical Finance 2004*, *Lecture Notes in Mathematics*, vol. 1919, Springer Verlag.
- [20] Carr, P., Geman, H., Madan, D. & Yor, M. (2003). Stochastic volatility for levy processes, *Mathematical Finance* **13**, 345–382.
- [21] Carr, P. & Jarrow, R. (1995). A discrete time synthesis of derivative security valuation using a term structure of futures prices, in *Handbooks in OR & MS*, R. Jarrow, V. Maksimoviz & W. Ziemba, eds, Elsevier Science B.V., Vol. 9, pp. 225–249.
- [22] Carr, P., Jarrow, R. & Myneni, R. (1992). Alternative characterizations of American put options, *Mathematical Finance* **2**(2), 87–106.

- [23] Carr, P. & Madan, D. (1998). Option valuation using the fast Fourier transform, *Journal of Computational Finance* **2**, 61–73.
- [24] Carr, P. & Madan, D. (1998). Toward a theory of volatility trading, in *Volatility*, R. Jarrow, ed., Risk Publications, pp. 417–427.
- [25] Caverhill, A. (1994). When is the spot rate Markovian?, *Mathematical Finance* **4**, 305–312.
- [26] Çetin, U., Jarrow, R. & Protter, P. (2004). Liquidity risk and arbitrage pricing theory, *Finance and Stochastics* **8**, 311–341.
- [27] Cetin, U., Jarrow, R., Protter, P. & Yildirim, Y. (2004). Modeling credit risk with partial information, *The Annals of Applied Probability* **14**(3), 1167–1178.
- [28] Chen, L., Filipovic, D. & Poor, H. (2004). Quadratic term structure models for risk free and defaultable rates, *Mathematical Finance* **14**(4), 515–536.
- [29] Cheng, P. & Scaillet, O. (2007). Linear-quadratic jump diffusion modeling, *Mathematical Finance* **17**(4), 575–598.
- [30] Cheyette, O. (1992). Term structure dynamics and mortgage valuation, *Journal of Fixed Income* **1**, 28–41.
- [31] Chiarella, C. & Kwon, O. (2000). A complete Markovian stochastic volatility model in the HJM framework, *Asia-Pacific Financial Markets* **7**, 293–304.
- [32] Cochrane, J. (2001). *Asset Pricing*, Princeton University Press.
- [33] Cont, R. & Tankov, P. (2004). *Financial Modeling with Jump Processes*, Chapman & Hall.
- [34] Cox, J. (1975). *Notes on Option Pricing I: Constant Elasticity of Variance Diffusions*, working paper, Stanford University.
- [35] Cox, J., Ingersoll, J. & Ross, S. (1985). A theory of the term structure of interest rates, *Econometrica* **53**, 385–407.
- [36] Cox, J. & Ross, S.A. (1976). The valuation of options for alternative stochastic processes, *Journal of Financial Economics* **3**(1/2), 145–166.
- [37] Cox, J., Ross, S. & Rubinstein, M. (1979). Option pricing: a simplified approach, *Journal of Financial Economics* **7**, 229–263.
- [38] Cox, J. & Rubinstein, M. (1985). *Option Markets*, Prentice Hall.
- [39] Dai, Q. & Singleton, K. (2000). Specification analysis of affine term structure models, *Journal of Finance* **55**, 1943–1978.
- [40] Delbaen, F. & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing, *Mathematische Annalen* **300**, 463–520.
- [41] Delbaen, F. & Schachermayer, W. (1995). The existence of absolutely continuous local Martingale measures, *Annals of Applied Probability* **5**, 926–945.
- [42] Delbaen, F. & Schachermayer, W. (1998). The fundamental theorem for unbounded stochastic processes, *Mathematische Annalen* **312**, 215–250.
- [43] Delbaen, F. & Schachermayer, W. (2006). *The Mathematics of Arbitrage*, Springer Verlag.
- [44] Detemple, J. (2006). *American Style Derivatives: Valuation and Computation*, *Financial Mathematics Series*, Chapman & Hall/CRC.
- [45] Duffie, D. & Kan, R. (1996). A yield factor model of interest rates, *Mathematical Finance* **6**, 379–406.
- [46] Duffie, D. & Lando, D. (2001). Term structure of credit spreads with incomplete accounting information, *Econometrica* **69**, 633–664.
- [47] Duffie, D., Pan, J. & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions, *Econometrica* **68**, 1343–1376.
- [48] Duffie, D. & Protter, P. (1992). From discrete to continuous time finance: weak convergence of the financial gain process, *Mathematical Finance* **2**(1), 1–15.
- [49] Duffie, D. & Singleton, K. (1999). Modeling term structures of defaultable bonds, *Review of Financial Studies* **12**(4), 687–720.
- [50] Dufresne, D. (2000). Laguerre series for Asian and other options, *Mathematical Finance* **10**(4), 407–428.
- [51] Dupire, B. (1992). Arbitrage pricing with stochastic volatility. *Proceedings of AFFI Conference*, Paris, June.
- [52] Dupire, B. (1996). *A Unified Theory of Volatility*. Paribas working paper.
- [53] Eberlein, E. & Raible, S. (1999). Term structure models driven by general Levy processes, *Mathematical Finance* **9**(1), 31–53.
- [54] Eberlein, E. & Ozkan, F. (2005). The Levy LIBOR model, *Finance and Stochastics* **9**, 327–348.
- [55] Flesaker, B. & Hughston, L. (1996). Positive interest, *Risk Magazine* **9**, 46–49.
- [56] Filipovic, D. (2001). *Consistency Problems for Heath Jarrow Morton Interest Rate Models*, *Springer Lecture Notes in Mathematics*, Vol. 1760, Springer Verlag.
- [57] Filipovic, D. (2002). Separable term structures and the maximal degree problem, *Mathematical Finance* **12**(4), 341–349.
- [58] Garman, M. & Kohlhagen, S. (1983). Foreign currency exchange values, *Journal of International Money and Finance* **2**, 231–237.
- [59] Geman, H. (1989). *The Importance of the Forward Neutral Probability in a Stochastic Approach of Interest Rates*, working paper, ESSEC.
- [60] Geske, R. (1979). The valuation of compound options, *Journal of Financial Economics* **7**, 63–81.
- [61] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*, Springer Verlag.
- [62] Glasserman, P. & Kou, S. (2003). The term structure of simple forward rates with jump risk, *Mathematical Finance* **13**(3), 383–410.
- [63] Green, R. & Jarrow, R. (1987). Spanning and completeness in markets with contingent claims, *Journal of Economic Theory* **41**(1), 202–210.
- [64] Goldstein, R. (2000). The term structure of interest rates as a random field, *Review of Financial Studies* **13**(2), 365–384.

- [65] Harrison, J. & Kreps, D. (1979). Martingales and arbitrage in multiperiod security markets, *Journal of Economic Theory* **20**, 381–408.
- [66] Harrison, J. & Pliska, S. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and Their Applications* **11**, 215–260.
- [67] Harrison, J. & Pliska, S. (1983). A stochastic calculus model of continuous trading: complete markets, *Stochastic Processes and Their Applications* **15**, 313–316.
- [68] He, H. (1990). Convergence of discrete time to continuous time contingent claims prices, *Review of Financial Studies* **3**, 523–546.
- [69] Heath, D. & Jarrow, R. (1987). Arbitrage, continuous trading and margin requirements, *Journal of Finance* **17**, 1129–1142.
- [70] Heath, D., Jarrow, R. & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation, *Econometrica* **60**(1), 77–105.
- [71] Heston, S. (1993). A closed form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**, 327–343.
- [72] Ho, T. & Lee, S. (1986). Term structure movements and pricing interest rate contingent claims, *Journal of Finance* **41**, 1011–1028.
- [73] Hull, J. & White, A. (1987). The pricing of options on assets with stochastic volatilities, *Journal of Finance* **42**, 271–301.
- [74] Hull, J. & White, A. (1990). Pricing interest rate derivative securities, *Review of Financial Studies* **3**, 573–592.
- [75] Jacka, S. (1991). Optimal stopping and the American put, *Mathematical Finance* **1**, 1–14.
- [76] Jarrow, R. (1987). The pricing of commodity options with stochastic interest rates, *Advances in Futures and Options Research* **2**, 15–28.
- [77] Jarrow, R. (1992). Market manipulation, bubbles, corners and short squeezes, *Journal of Financial and Quantitative Analysis* **27**(3), 311–336.
- [78] Jarrow, R. (1994). Derivative security markets, market manipulation and option pricing, *Journal of Financial and Quantitative Analysis* **29**(2), 241–261.
- [79] Jarrow, R. (1999). In honor of the Nobel Laureates Robert C. Merton and Myron S. Scholes: a partial differential equation that changed the world, *Journal of Economic Perspectives* **13**(4), 229–248.
- [80] Jarrow, R., Lando, D. & Turnbull, S. (1997). A Markov model for the term structure of credit risk spreads, *Review of Financial Studies* **10**(1), 481–523.
- [81] Jarrow, R., Lando, D. & Yu, F. (2005). Default risk and diversification: theory and empirical applications, *Mathematical Finance* **15**(1), 1–26.
- [82] Jarrow, R. & Madan, D. (1995). Option pricing using the term structure of interest rates to hedge systematic discontinuities in asset returns, *Mathematical Finance* **5**(4), 311–336.
- [83] Jarrow, R. & Rudd, A. (1982). Approximate option valuation for arbitrary stochastic processes, *Journal of Financial Economics* **10**, 347–369.
- [84] Jarrow, R. & Rudd, A. (1983). *Option Pricing*, Dow Jones Irwin.
- [85] Jarrow, R. & Turnbull, S. (1992). Credit risk: drawing the analogy, *Risk Magazine* **5**(9).
- [86] Jarrow, R. & Turnbull, S. (1995). Pricing derivatives on financial securities subject to credit risk, *Journal of Finance* **50**(1), 53–85.
- [87] Jarrow, R. & Turnbull, S. (1997). When swaps are dropped, *Risk Magazine* **10**(5), 70–75.
- [88] Jarrow, R. & Turnbull, S. (1998). A unified approach for pricing contingent claims on multiple term structures, *Review of Quantitative Finance and Accounting* **10**(1), 5–19.
- [89] Jarrow, R. & Yildirim, Y. (2003). Pricing treasury inflation protected securities and related derivatives using an HJM model, *Journal of Financial and Quantitative Analysis* **38**(2), 337–358.
- [90] Jarrow, R. & Yu, F. (2000). Counterparty risk and the pricing of defaultable securities, *Journal of Finance* **56**(5), 1765–1799.
- [91] Jin, Y. & Glasserman, P. (2001). Equilibrium positive interest rates: a unified view, *Review of Financial Studies* **14**, 187–214.
- [92] Jeffrey, A. (1995). Single factor heath Jarrow Morton term structure models based on Markov spot rate dynamics, *Journal of Financial and Quantitative Analysis* **30**, 619–642.
- [93] Johnson, H. (1983). An analytic approximation of the American put price, *Journal of Financial and Quantitative Analysis* **18**, 141–148.
- [94] Jones, E., Mason, S. & Rosenfeld, E. (1984). Contingent claims analysis of corporate capital structures: an empirical investigation, *Journal of Finance* **39**, 611–627.
- [95] Kennedy, D. (1994). The term structure of interest rates as a Gaussian random field, *Mathematical Finance* **4**, 247–258.
- [96] Kim, J. (1990). The analytic valuation of American options, *Review of Financial Studies* **3**, 547–572.
- [97] Lando, D. (1998). On Cox processes and credit risky securities, *Review of Derivatives Research* **2**, 99–120.
- [98] Lando, D. (2004). *Credit Risk Modeling: Theory and Applications*, Princeton University Press, Princeton.
- [99] Leland, H. (1985). Option pricing and replication with transaction costs, *Journal of Finance* **15**, 1283–1391.
- [100] Leland, H. (1994). Corporate debt value, bond covenants and optimal capital structure, *Journal of Finance* **49**, 1213–1252.
- [101] Longstaff, F. & Schwartz, E. (2001). Valuing American options by simulation: a simple least squares approach, *Review of Financial Studies* **14**, 113–147.

- [102] Madan, D. & Milne, F. (1991). Option pricing with variance gamma martingale components, *Mathematical Finance* **1**, 39–55.
- [103] Madan, D. & Unal, H. (1998). Pricing the risks of default, *Review of Derivatives Research* **2**, 121–160.
- [104] Merton, R.C. (1973). The theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.
- [105] Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.
- [106] Merton, R.C. (1976). Option pricing when underlying stock returns are discontinuous, *Journal of Financial Economics* **3**, 125–144.
- [107] Merton, R.C. (1990). *Continuous Time Finance*, Basil Blackwell, Cambridge, Massachusetts.
- [108] Miltersen, K., Nielsen, J. & Sandmann, K. (2006). New no-arbitrage conditions and the term structure of interest rate futures, *Annals of Finance* **2**, 303–325.
- [109] Miltersen, K., Sandmann, K. & Sondermann, D. (1997). Closed form solutions for term structure derivatives with log-normal interest rates, *Journal of Finance* **52**, 409–430.
- [110] Modigliani, F. & Miller, M. (1958). The cost of capital, corporation finance, and the theory of investment, *American Economic Review* **48**, 261–297.
- [111] Nakajima, K. & Maeda, A. (2007). Pricing commodity spread options with stochastic term structure of convenience yields and interest rates, *Asia Pacific Financial Markets* **14**, 157–184.
- [112] Rebonato, R. (2002). *Modern Pricing of Interest Rate Derivatives: The LIBOR Market Model and Beyond*, Princeton University Press.
- [113] Rendleman, R. & Bartter, B. (1979). Two state option pricing, *Journal of Finance* **34**, 1093–1110.
- [114] Rogers, L. (1994). The potential approach to the term structure of interest rates and foreign exchange rates, *Mathematical Finance* **7**, 157–176.
- [115] Roll, R. (1977). An analytic valuation formula for unprotected American call options on stocks with known dividends, *Journal of Financial Economics* **5**, 251–258.
- [116] Ross, S. (1976). Options and efficiency, *Quarterly Journal of Economics* **90**, 75–89.
- [117] Samuelson, P. (1965). Rational theory of warrant pricing, *Industrial Management Review* **6**, 13–39.
- [118] Samuelson, P. & Merton, R.C. (1969). A complete model of warrant pricing that maximizes utility, *Industrial Management Review* **10**(2), 17–46.
- [119] Sandmann, K., Sondermann, D. & Miltersen, K. (1995). Closed form term structure derivatives in a Heath Jarrow Morton model with log-normal annually compounded interest rates, *Proceedings of the Seventh Annual European Research Symposium*, Bonn, September 1994, Chicago Board of Trade, pp. 145–164.
- [120] Schonbucher, P. (2004). *Information Driven Default Contagion*, working paper, ETH Zurich.
- [121] Schweizer, M. & Wissel, J. (2008). Term structure of implied volatilities: absence of arbitrage and existence results, *Mathematical Finance* **18**(1), 77–114.
- [122] Sharpe, W. (1981). *Investments*, Prentice Hall, Englewood Cliffs.
- [123] Turnbull, S. & Wakeman, L. (1991). A quick algorithm for pricing European average options, *Journal of Financial and Quantitative Analysis* **26**, 377–389.
- [124] Vasicek, O. (1977). An equilibrium characterization of the term structure, *Journal of Financial Economics* **5**, 177–188.
- [125] Zhou, C. (2001). The term structure of credit spreads with jump risk, *Journal of Banking and Finance* **25**, 2015–2040.

ROBERT A. JARROW

Modern Portfolio Theory

Modern portfolio theory (MPT) is generally defined as the body of financial economics beginning with Markowitz' famous 1952 paper, "Portfolio Selection", and extending through the next several decades of research into what has variously been called *Financial Decision Making under Uncertainty*, *The Theory of Investments*, *The Theory of Financial Economics*, *Theory of Asset Selection and Capital–Market Equilibrium*, and *The Revolutionary Idea of Finance* [45, 53, 58, 82, 88, 98]. Usually this definition includes the Capital Asset Pricing Model (CAPM) and its various extensions. Markowitz once remarked to Marschak that the first "CAPM" should be attributed to Marschak because of his pioneering work in the field [56]; Marschak politely declined the honor.

The original CAPM, as we understand it today, was first developed by Treynor [91, 92], and subsequently independently derived in the works of Sharpe [84], Lintner [47], and Mossin [65]. With the exception of some commercially successful multifactor models that implement the approaches pioneered in [71, 72, 74, 75], most practitioners have little use for market models other than the CAPM, although (or, perhaps rather, *because of* the simplicity it derives from the fact that) its conclusions are based on extremely restrictive and unrealistic assumptions. Academics have spent much time and effort attempting to substantiate or refute the validity of the CAPM as a positive economic model. The best examples of such attempts are [13, 28]. Roll [70] effectively ended this debate, however, by demonstrating that, since the "market portfolio" is not measurable, the CAPM can never be empirically proven or disproven.

History of Modern Portfolio Theory

The history of MPT extends back farther than the history of CAPM, to Tobin [90], Markowitz [53], and Roy [78], all of whom consider the "price of risk". For more detailed treatments of MPT and pre-MPT financial economic thought, refer to [22, 69, 82]. The prehistory of MPT can be traced further yet, to Hicks [34] who includes the "price of risk" in his discussion of commodity futures and to Williams [95] who considers stock prices to

be determined by the present value of discounted future dividends. MPT prehistory can be traced even beyond to Bachelier [3], who was the first to describe arithmetic Brownian motion with the objective of determining the value of financial derivatives, all the way to Bernoulli [7], who originated the concept of risk aversion while working to solve the St. Petersburg Paradox. Bernoulli, in his derivation of logarithmic utility, suggested that people maximize "moral expectation"—what we call today *expected utility*; further, Bernoulli, like Markowitz [53] and Roy [78], advised risk-averse investors to diversify: "... it is advisable to divide goods which are exposed to some small danger into several portions rather than to risk them all together."

Notwithstanding this ancient history, MPT is inextricably connected to CAPM, which for the first time placed the investor's problem in the context of an economic equilibrium. This modern approach finds its origin in the work of Mossin [65], Lintner [47, 48], and Sharpe [84], and even earlier in Treynor [91, 92]. Accounts of these origins can be found in [8, 29, 85]. Treynor [92] built on the single-period discrete-time foundation of Markowitz [53, 54] and Tobin [90]. Similar CAPM models of this type were later published in [47, 48, 84]. Mossin [65] clarified Sharpe [84] by providing a more precise specification of the equilibrium conditions. Fama [26] reconciled the Sharpe and Lintner models; Lintner [49] incorporated heterogeneous beliefs; and Mayers [57] allowed for concentrated portfolios through trading restrictions on risky assets, transactions costs, and information asymmetries. Black [10] utilized the two-fund separation theorem to construct the zero-beta CAPM, by using a portfolio that is orthogonal to the market portfolio in place of a risk-free asset. Rubinstein [79] extended the model to higher moments and also (independently of Black) derived the CAPM without a riskless asset.

Discrete-time multiperiod models were the next step; these models generally extend the discrete-time single-period model into an intertemporal setting in which investors maximize the expected utility of lifetime consumption and bequests. Building upon the multiperiod lifetime consumption literature of Phelps [68], Mirrlees [63], Yaari [97], Levhari and Srinivasan [44], and Hahn [30], models of this type include those of Merton [59, 60], Samuelson [83], Hakansson [31, 32], Fama [27], Beja [4], Rubinstein [80, 81], Long [50, 51], Kraus and Litzenberger

[41], and culminate in the consumption CAPMs (CCAPMs) of Lucas [52] and Breeden [15].

The multiperiod approach was taken to its continuous-time limit in the intertemporal CAPM (“ICAPM”) of Merton [61]. In addition to the standard assumptions—limited liability of assets, no market frictions, individual trading does not affect prices, the market is in equilibrium, a perfect borrowing and lending market exists, and no nonnegativity constraints (relaxing the no short-sale rule employed by Tobin and Sharpe but not by Treynor and Lintner)—this model assumes that trading takes place continually through time, as opposed to at discrete points in time. Rather than assuming normally distributed security returns, the ICAPM assumes a log-normal distribution of prices and a geometric Brownian motion of security returns. Also, the constant rate of interest provided by the risk-free asset in the CAPM is replaced by a dynamically changing rate, which is certain in the next instant but uncertain in the future. Williams [96] extended this model by relaxing the homogeneous expectations assumption, and Duffie and Huang [23] confirmed that such a relaxation is consistent with the ICAPM. The continuous-time model was shown to be consistent with a single-beta CCAPM by Breeden [15]. Hellwig [33] and Duffie and Huang [24] construct continuous-time models that allow for informational asymmetries. The continuous-time model was further extended to include macroeconomic factors in [20]. Kyle [42] constructs an ICAPM to model insider trading.

These, and other CAPMs, including the international models of Black [12], Solnik [86], and Stulz [89], as well as the CAPMs of Ross [73, 76] and Stapleton and Subrahmanyam [87], are reviewed in [16, 17, 19, 62, 77]. Bergstrom [5] provides a survey of continuous-time models.

Extensions of the CAPM have also been developed for use, in particular, in industrial applications; for example, Cummins [21] reviews the models of Cooper [18], Biger and Kahane [9], Fairley [25], Kahane [39], Hill [35], Ang and Lai [2], and Turner [94], which are specific to the insurance industry. More recent work continues to extend the theory. Nielsen [66, 67], Allingham [1], and Berk [6] examine conditions for equilibrium in the CAPM. Current research, such as the collateral adjusted CCAPM of

Hindy and Huang [36] and the parsimonious conditional discrete-time CAPM and simplified infinite-date model of LeRoy [43], continues to build upon the model originated in [91]. Each is perhaps more realistic, if less elegant, than the original. And yet it is the single period, discrete-time CAPM that has become popular and endured, as all great models do, precisely *because* it is simple and unrealistic. It is realistic enough, apparently, to be coincident with the utility functions of great many agents.

A Perspective on CAPM

One of the puzzles that confronts the historian of CAPM is the changing attitude over time and across different scholarly communities toward the seminal work of Treynor [91, 92]. Contemporaries consistently cited the latter paper [11, 13, 37, 38], including also [84, 85]. However, in other papers, such as [16, 45, 55], these citations were not made. Histories and bibliographies continue to take note of Treynor’s contribution [8, 14, 58, 82], but not textbooks or the scholarly literature that builds on CAPM. Why not?

One reason is certainly that Treynor’s manuscript [92] was not actually published in a book until much later [40], although the paper did circulate widely in mimeograph form. Another is that Treynor never held a permanent academic post, and so did not have a community of students and academic colleagues to draw attention to his work. A third is that, although Treynor continued to write on financial topics, writings collected in [93], these writings were consistently addressed to practitioners, not to an academic audience.

Even more than these, perhaps the most important reason (paradoxically) is the enormous attention that was paid in subsequent years to refinement of MPT. Unlike Markowitz and Sharpe, Treynor came to CAPM from a concern about the firm’s capital budgeting problem, not the investor’s portfolio allocation problem. (This concern is clear in the 1961 draft, which builds explicitly on [64].) This was the same concern, of course, that motivated Lintner, and it is significant therefore that the CAPMs of Lintner and Sharpe were originally seen as different theories, rather than different formulations of the same theory.

Because the portfolio choice problem became such a dominant strand of academic research, it

was perhaps inevitable that retrospective accounts of CAPM would emphasize the line of development that passes from the individual investor's problem to the general equilibrium problem, which is to say the line that passes through Tobin and Markowitz to Sharpe. Lintner and Mossin come in for some attention, as academics who contributed not only their own version of CAPM but also produced a series of additional contributions to the academic literature. However, Treynor was not only interested in a different problem but also was, and remained, a practitioner.

Conclusion

In 1990, the world beyond financial economists was made aware of the importance of MPT, when Markowitz and Sharpe, along with Miller, were awarded the Nobel Prize in Economics for their roles in the development of MPT. In the presentation speech, Assar Lindbeck of the Royal Swedish Academy of Sciences said "Before the 1950s, there was hardly any theory whatsoever of financial markets. A first pioneering contribution in the field was made by Harry Markowitz, who developed a theory . . . [which] shows how the multidimensional problem of investing under conditions of uncertainty in a large number of assets . . . may be reduced to the issue of a trade-off between only two dimensions, namely the expected return and the variance of the return of the portfolio The next step in the analysis is to explain how these asset prices are determined. This was achieved by development of the so-called Capital Asset Pricing Model, or CAPM. It is for this contribution that William Sharpe has been awarded. The CAPM shows that the optimum risk portfolio of a financial investor depends only on the portfolio manager's prediction about the prospects of different assets, not on his own risk preferences The Capital Asset Pricing Model has become the backbone of modern price theory of financial markets" [46].

References

- [1] Allingham, M. (1991). Existence theorems in the capital asset pricing model, *Econometrica* **59**(4), 1169–1174.
- [2] Ang, J.S. & Lai, T.-Y. (1987). Insurance premium pricing and ratemaking in competitive insurance and capital asset markets, *The Journal of Risk and Insurance* **54**, 767–779.
- [3] Bachelier, L. (1900). Théorie de la spéculation, *Annales Scientifique de l'École Normale Supérieure* **17**, 3e serie, 21–86; Translated by Boness, A.J. and reprinted in Cootner, P.H. (ed.) (1964). *The Random Character of Stock Market Prices*, MIT Press, Cambridge. (Revised edition, first MIT Press Paperback Edition, July 1967). pp. 17–78; Also reprinted as Bachelier, L. (1995). *Théorie de la Spéculation & Théorie Mathématique du jeu*, (2 titres en 1 vol.) *Les Grands Classiques Gauthier-Villars*, Éditions Jacques Gabay, Paris, Part 1, pp. 21–86.
- [4] Beja, A. (1971). The structure of the cost of capital under uncertainty, *The Review of Economic Studies* **38**, 359–369.
- [5] Bergstrom, A.R. (1988). The history of continuous-time econometric models, *Econometric Theory* **4**(3), 365–383.
- [6] Berk, J.B. (1992). *The Necessary and Sufficient Conditions that Imply the CAPM*, working paper, Faculty of Commerce, University of British Columbia, Canada; Subsequently published as (1997). Necessary conditions for the CAPM, *Journal of Economic Theory* **73**, 245–257.
- [7] Bernoulli, D. (1738). *Exposition of a new theory on the measurement of risk*, *Papers of the Imperial Academy of Science*, Petersburg, Vol. II, pp. 175–192; Translated and reprinted in Sommer, L. (1954). *Econometrica* **22**(1), 23–36.
- [8] Bernstein, P.L. (1992). *Capital Ideas: The Improbable Origins of Modern Wall Street*, The Free Press, New York.
- [9] Biger, N. & Kahane, Y. (1978). Risk considerations in insurance ratemaking, *The Journal of Risk and Insurance* **45**, 121–132.
- [10] Black, F. (1972). Capital market equilibrium with restricted borrowing, *Journal of Business* **45**(3), 444–455.
- [11] Black, F. (1972). Equilibrium in the creation of investment goods under uncertainty, in *Studies in the Theory of Capital Markets*, M.C. Jensen, ed., Praeger, New York, pp. 249–265.
- [12] Black, F. (1974). International capital market equilibrium with investment barriers, *Journal of Financial Economics* **1**(4), 337–352.
- [13] Black, F., Jensen, M.C. & Scholes, M. (1972). The capital asset pricing model: some empirical tests, in *Studies in the Theory of Capital Markets*, M.C. Jensen, ed., Praeger, New York, pp. 79–121.
- [14] Brealey, R.A. & Edwards, H. (1991). *A Bibliography of Finance*, MIT Press, Cambridge.
- [15] Breeden, D.T. (1979). An intertemporal asset pricing model with stochastic consumption and investment opportunities, *Journal of Financial Economics* **7**(3), 265–296.
- [16] Breeden, D.T. (1987). Intertemporal portfolio theory and asset pricing, in *The New Palgrave Finance*, J. Eatwell, M. Milgate & P. Newman, eds, W.W. Norton, New York, pp. 180–193.

-
- [17] Brennan, M.J. (1987). Capital asset pricing model, in *The New Palgrave Finance*, J. Eatwell, M. Milgate & P. Newman, eds, W.W. Norton, New York, pp. 91–102.
- [18] Cooper, R.W. (1974). *Investment Return and Property-Liability Insurance Ratemaking*, Huebner Foundation, University of Pennsylvania, Philadelphia.
- [19] Copeland, T.E. & Weston, J.F. (1987). Asset pricing, in *The New Palgrave Finance*, J. Eatwell, M. Milgate & P. Newman, eds, W.W. Norton, New York, pp. 81–85.
- [20] Cox, J.C., Ingersoll Jr, J.E. & Ross, S.A. (1985). An intertemporal general equilibrium model of asset prices, *Econometrica* **53**(2), 363–384.
- [21] Cummins, J.D. (1990). Asset pricing models and insurance ratemaking, *ASTIN Bulletin* **20**(2), 125–166.
- [22] Dimson, E. & Mussavain, M. (2000). *Three Centuries of Asset Pricing*, Social Science Research Network Electronic Library, paper 000105402.pdf. January.
- [23] Duffie, D. & Huang, C.F. (1985). Implementing Arrow-Debreu equilibria by continuous trading of few long-lived securities, *Econometrica* **53**, 1337–1356; Also reprinted in edited by Schaefer, S. (2000). *Continuous-Time Finance*, Edward Elgar, London.
- [24] Duffie, D. & Huang, C.F. (1986). Multiperiod security markets with differential information: martingales and resolution times, *Journal of Mathematical Economics* **15**, 283–303.
- [25] Fairley, W. (1979). Investment income and profit margins in property-liability insurance: theory and empirical tests, *Bell Journal of Economics* **10**, 192–210.
- [26] Fama, E.F. (1968). Risk, return, and equilibrium: some clarifying comments, *Journal of Finance* **23**(1), 29–40.
- [27] Fama, E.F. (1970). Multiperiod consumption—investment decisions, *The American Economic Review* **60**, 163–174.
- [28] Fama, E.F. & MacBeth, J. (1973). Risk, return and equilibrium: empirical tests, *The Journal of Political Economy* **81**(3), 607–636.
- [29] French, C.W. (2003). The Treynor capital asset pricing model, *Journal of Investment Management* **1**(2), Second quarter, 60–72.
- [30] Hahn, F.H. (1970). Savings and uncertainty, *The Review of Economic Studies* **37**(1), 21–24.
- [31] Hakansson, N.H. (1969). Optimal investment and consumption strategies under risk, an uncertain lifetime, and insurance, *International Economic Review* **10**(3), 443–466.
- [32] Hakansson, N.H. (1970). Optimal investment and consumption strategies under risk for a class of utility functions, *Econometrica* **38**(5), 587–607.
- [33] Hellwig, M.F. (1982). Rational expectations equilibrium with conditioning on past prices: a mean-variance example, *Journal of Economic Theory* **26**, 279–312.
- [34] Hicks, J.R. (1939). *Value and Capital: An Inquiry into some Fundamental Principles of Economic Theory*, Clarendon Press, Oxford.
- [35] Hill, R. (1979). Profit regulation in property-liability insurance, *Bell Journal of Economics* **10**, 172–191.
- [36] Hindy, A. & Huang, M. (1995). *Asset Pricing With Linear Collateral Constraints*. unpublished manuscript, Graduate School of Business, Stanford University. March.
- [37] Jensen, M.C. (ed) (1972). *Studies in the Theory of Capital Markets*, Praeger, New York.
- [38] Jensen, M.C. (1972). The foundations and current state of capital market theory, in *Studies in the Theory of Capital Markets*, M.C. Jensen, ed., Praeger, New York, pp. 3–43.
- [39] Kahane, Y. (1979). The theory of insurance risk premiums—a re-examination in the light of recent developments in capital market theory, *ASTIN Bulletin* **10**(2), 223–239.
- [40] Korajczyk, R.A. (1999). *Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics*, Risk Books, London.
- [41] Kraus, A. & Litzenberger, R.H. (1975). Market equilibrium in a multiperiod state-preference model with logarithmic utility, *Journal of Finance* **30**(5), 1213–1227.
- [42] Kyle, A.S. (1985). Continuous auctions and insider trading, *Econometrica* **53**(3), 1315–1335.
- [43] LeRoy, S.F. (2002). *Theoretical Foundations for Conditional CAPM*. unpublished manuscript, University of California, Santa Barbara. May.
- [44] Levhari, D. & Srinivasan, T.N. (1969). Optimal savings under uncertainty, *The Review of Economic Studies* **36**(106), 153–163.
- [45] Levy, H. & Sarnat, M. (eds) (1977). *Financial Decision Making under Uncertainty*, Academic Press, New York.
- [46] Lindbeck, A. (1990). The sveriges riksbank prize in economic sciences in memory of Alfred Nobel 1990 presentation speech, *Nobel Lectures, Economics 1981–1990*, K.-G. Mäler, ed., World Scientific Publishing Co., Singapore, 1992.
- [47] Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets, *The Review of Economics and Statistics* **47**, 13–37.
- [48] Lintner, J. (1965). Securities prices, risk, and maximal gains from diversification, *Journal of Finance* **20**(4), 587–615.
- [49] Lintner, J. (1969). The aggregation of investor’s diverse judgment and preferences in purely competitive securities markets, *Journal of Financial and Quantitative Analysis* **4**, 347–400.
- [50] Long Jr, J.B. (1972). Consumption-investment decisions and equilibrium in the securities markets, in *Studies in the Theory of Capital Markets*, M.C. Jensen, ed., Praeger, New York, pp. 146–222.
- [51] Long Jr, J.B. (1974). Stock prices, inflation and the term structure of interest rates, *Journal of Financial Economics* **2**, 131–170.
- [52] Lucas Jr, R.E. (1978). Asset prices in an exchange economy, *Econometrica* **46**(6), 1429–1445.
- [53] Markowitz, H.M. (1952). Portfolio selection, *Journal of Finance* **7**(1), 77–91.

- [54] Markowitz, H.M. (1959). *Portfolio Selection: Efficient Diversification of Investments*, Cowles Foundation for Research in Economics at Yale University, Monograph #6. John Wiley & Sons, Inc., New York. (2nd Edition, 1991, Basil Blackwell, Inc., Cambridge).
- [55] Markowitz, H.M. (2000). *Mean-Variance Analysis in Portfolio Choice and Capital Markets*, Frank J. Fabozzi Associates, New Hope.
- [56] Marschak, J. (1938). Money and the theory of assets, *Econometrica* **6**, 311–325.
- [57] Mayers, D. (1972). Nonmarketable assets and capital market equilibrium under uncertainty, in *Studies in the Theory of Capital Markets*, M.C. Jensen, ed., Praeger, New York, pp. 223–248.
- [58] Mehrling, P. (2005). *Fischer Black and the Revolutionary Idea of Finance*, Wiley, Hoboken.
- [59] Merton, R.C. (1969). Lifetime portfolio selection under uncertainty: the continuous time case, *The Review of Economics and Statistics* **51**, 247–257; Reprinted as chapter 4 of Merton, R.C. (1990). *Continuous-Time Finance*, Blackwell, Cambridge, pp. 97–119.
- [60] Merton, R.C. (1971). Optimum consumption and portfolio rules in a continuous time model, *Journal of Economic Theory* **3**, 373–413; Reprinted as chapter 5 of Merton, R.C. (1990). *Continuous-Time Finance*, Blackwell, Cambridge pp. 120–165.
- [61] Merton, R.C. (1973). An intertemporal capital asset pricing model, *Econometrica* **41**, 867–887; Reprinted as chapter 15 of Merton, R.C. (1990). *Continuous-Time Finance*, Blackwell, Cambridge, pp. 475–523.
- [62] Merton, R.C. (1990). *Continuous-Time Finance*, Blackwell, Cambridge. (revised paperback edition, 1999 reprint).
- [63] Mirrlees, J.A. (1965). *Optimum Accumulation Under Uncertainty*. unpublished manuscript. December.
- [64] Modigliani, F. & Miller, M.H. (1958). The cost of capital, corporation finance, and the theory of investment, *The American Economic Review* **48**, 261–297.
- [65] Mossin, J. (1966). Equilibrium in a capital asset market, *Econometrica* **34**(4), 768–783.
- [66] Nielsen, L.T. (1990). Equilibrium in CAPM without a riskless asset, *The Review of Economic Studies* **57**, 315–324.
- [67] Nielsen, L.T. (1990). Existence of equilibrium in CAPM, *Journal of Economic Theory* **52**, 223–231.
- [68] Phelps, E.S. (1962). The accumulation of risky capital: a sequential utility analysis, *Econometrica* **30**(4), 729–743.
- [69] Poitras, G. (2000). *The Early History of Financial Economics*, Edward Elgar, Chenttenham.
- [70] Roll, R. (1977). A critique of the asset pricing theory's tests, *Journal of Financial Economics* **4**(2), 129–176.
- [71] Rosenberg, B. (1974). Extra-market component of covariance in security returns, *Journal of Financial and Quantitative Analysis* **9**(2), 263–273.
- [72] Rosenberg, B. & McKibben, W. (1973). The prediction of systematic and specific risk in security returns, *Journal of Financial and Quantitative Analysis* **8**(3), 317–333.
- [73] Ross, S.A. (1975). Uncertainty and the heterogeneous capital good model, *The Review of Economic Studies* **42**(1), 133–146.
- [74] Ross, S.A. (1976). The arbitrage theory of capital asset pricing, *Journal of Economic Theory* **13**(3), 341–360.
- [75] Ross, S.A. (1976). Risk, return and arbitrage, in *Risk and Return in Finance*, I. Friend & J. Bicksler, eds, Ballinger, Cambridge, pp. 1–34.
- [76] Ross, S.A. (1978). Mutual fund separation in financial theory—the separating distributions, *Journal of Economic Theory* **17**(2), 254–286.
- [77] Ross, S.A. (1987). Finance, in *The New Palgrave Finance*, J. Eatwell, M. Milgate & P. Newman, eds, W.W. Norton, New York, pp. 1–34.
- [78] Roy, A.D. (1952). Safety first and the holding of assets, *Econometrica* **20**(3), 431–439.
- [79] Rubinstein, M. (1973). The fundamental theorem of parameter-preference security valuation, *Journal of Financial and Quantitative Analysis* **8**, 61–69.
- [80] Rubinstein, M. (1974). *A Discrete-Time Synthesis of Financial Theory*, Working Paper 20, Haas School of Business, University of California at Berkeley; Reprinted in *Research in Finance*, JAI Press, Greenwich, Vol. 3, pp. 53–102.
- [81] Rubinstein, M. (1976). The valuation of uncertain income streams and the pricing of options, *Bell Journal of Economics* **7**, Autumn, 407–425.
- [82] Rubinstein, M. (2006). *A History of the Theory of Investments My Annotated Bibliography*, Wiley, Hoboken.
- [83] Samuelson, P.A. (1969). Lifetime portfolio selection by dynamic stochastic programming, *The Review of Economics and Statistics* **57**(3), 239–246.
- [84] Sharpe, W.F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk, *Journal of Finance* **19**(3), 425–442.
- [85] Sharpe, W.F. (1990). Autobiography, in *Les Prix Nobel 1990*, Tore Frängsmyr, ed., Nobel Foundation, Stockholm.
- [86] Solnik, B. (1974). An equilibrium model of international capital markets, *Journal of Economic Theory* **8**(4), 500–524.
- [87] Stapleton, R.C. & Subrahmanyam, M. (1978). A multiperiod equilibrium asset pricing model, *Econometrica* **46**(5), 1077–1095.
- [88] Stone, B.K. (1970). *Risk, Return, and Equilibrium, a General Single-Period Theory of Asset Selection and Capital-Market Equilibrium*, MIT Press, Cambridge.
- [89] Stulz, R.M. (1981). A model of international asset pricing, *Journal of Financial Economics* **9**(4), 383–406.
- [90] Tobin, J. (1958). Liquidity preference as behavior towards risk, *The Review of Economic Studies* (67), 65–86. Reprinted as Cowles Foundation Paper 118.
- [91] Treynor, J.L. (1961). *Market Value, Time and Risk*. unpublished manuscript dated 8/8/61.
- [92] Treynor, J.L. (1962). *Toward a Theory of Market Value of Risky Assets*, unpublished manuscript. “Rough Draft”

- dated by Mr. Treynor to the fall of 1962. A final version was published in 1999, in *Asset Pricing and Portfolio Performance*, R.A. Korajczyk, ed., Risk Books, London, pp. 15–22.
- [93] Treynor, J.L. (2007). *Treynor on Institutional Investing*, Wiley, Hoboken.
- [94] Turner, A.L. (1987). Insurance in an equilibrium asset pricing model, in *Fair Rate of Return in Property-Liability Insurance*, J.D. Cummins & S.E. Harrington, eds, Kluwer Academic Publishers, Norwell.
- [95] Williams, J.B. (1938). *The Theory of Investment Value*, Harvard University Press, Cambridge.
- [96] Williams, J.T. (1977). Capital asset prices with heterogeneous beliefs, *Journal of Financial Economics* **5**, 219–239.
- [97] Yaari, M.E. (1965). Uncertain lifetime, life insurance, and the theory of the consumer, *The Review of Economic Studies* **32**(2), 137–150.
- [98] The Royal Swedish Academy of Sciences (1990). *The Sveriges Riskbank Prize in Economic Sciences in Memory of Alfred Nobel 1990*, Press release 16 October 1990.
- of the American Economic Association, Boston, December; Subsequently extended and published as (1965). Investment decision under uncertainty: choice-theoretic approaches, *The Quarterly Journal of Economics* **79**(5), 509–536; Also, see (1966). Investment decision under uncertainty: applications of the state-preference approach, *The Quarterly Journal of Economics* **80**(2), 252–277.
- Itô, K. (1944). Stochastic integrals, *Proceedings of the Imperial Academy Tokyo* **22**, 519–524.
- Itô, K. (1951). Stochastic differentials, *Applied Mathematics and Optimization* **1**, 374–381.
- Itô, K. (1998). My sixty years in studies of probability theory, acceptance speech of the Kyoto prize in basic sciences, in *The Inamori Foundation Yearbook 1998*, Inamori Foundation, Kyoto.
- Jensen, M.C. (1968). The performance of mutual funds in the period 1945–64, *Journal of Finance* **23**(2), 389–416.
- Jensen, M.C. (1969). Risk, the pricing of capital assets, and the evaluation of investment portfolios, *Journal of Business* **42**(2), 167–247.
- Keynes, J.M. (1936). *The General Theory of Employment, Interest, and Money*, Harcourt Brace, New York.
- Leontief, W. (1947). Postulates: Keynes' general theory and the classicists, in *The New Economics: Keynes' Influence on Theory and Public Policy*, S.E. Harris, ed., Knopf, New York, Chapter 19, pp. 232–242.
- Lintner, J. (1965). Securities Prices and Risk; the Theory and a Comparative Analysis of AT&T and Leading Industrials, *Paper Presented at the Bell System Conference on the Economics of Regulated Public Utilities*, University of Chicago Business School, Chicago, June.
- Lintner, J. (1970). The market price of risk, size of market and investor's risk aversion, *The Review of Economics and Statistics* **52**, 87–99.
- Lintner, J. (1971). The effects of short selling and margin requirements in perfect capital markets, *Journal of Financial and Quantitative Analysis* **6**, 1173–1196.
- Lintner, J. (1972). *Finance and Capital Markets*, National Bureau of Economic Research, New York.
- Mandelbrot, B.B. (1987). Louis Bachelier, in *The New Palgrave Finance*, J. Eatwell, M. Milgate & P. Newman, eds, W.W. Norton, New York, pp. 86–88.
- Markowitz, H.M. (1952). The utility of wealth, *The Journal of Political Economy* **60**(2), 151–158.
- Markowitz, H.M. (1956). The optimization of a quadratic function subject to linear constraints, *Naval Research Logistics Quarterly* **3**, 111–133.
- Markowitz, H.M. (1957). The elimination form of the inverse and its application to linear programming, *Management Science* **3**, 255–269.
- Marschak, J. (1950). Rational behavior, uncertain prospects, and measurable utility, *Econometrica* **18**(2), 111–141.
- Marschak, J. (1951). Why “Should” statisticians and businessmen maximize “moral expectation”?, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, pp. 493–506. Reprinted as Cowles Foundation Paper 53.

Further Reading

- Arrow, K.J. (1953). Le Rôle des Valuers Boursières pour la Répartition la Meilleure des Risques, *Économetrie, Colloques Internationaux du Centre National de la Recherche Scientifique* **11**, 41–47.
- Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *The Journal of Political Economy* **81**(3), 637–654.
- Cootner, P.H. (ed.) (1964). *The Random Character of Stock Market Prices*, MIT Press, Cambridge. (Revised edition, First MIT Press Paperback Edition, July 1967).
- Courtault, J.M., Kabanov, Y., Bru, B., Crépel, P., Lebon, I. & Le Marchand, A. (2000). Louis Bachelier on the centenary of théorie de la spéculation, *Mathematical Finance* **10**(3), 341–353.
- Cvitanić, J., Lazrak, A., Martinelli, L. & Zapatero, F. (2002). *Revisiting Treynor and Black (1973): an Intertemporal Model of Active Portfolio Management*, unpublished manuscript. The University of Southern California and the University of British Columbia.
- Duffie, D. (1996). *Dynamic Asset Pricing Theory*, 2nd Edition, Princeton University Press, Princeton.
- Eatwell, J., Milgate, M. & Newman, P. (eds) (1987). *The New Palgrave Finance*, W.W. Norton, New York.
- Friedman, M. & Jimmie Savage, L. (1948). The utility analysis of choices involving risk, *The Journal of Political Economy* **56**(4), 279–304.
- Friend, I. & Bicksler, J.L. (1976). *Risk and Return in Finance*, Ballinger, Cambridge.
- Hakansson, N.H. (1987). Portfolio analysis, in *The New Palgrave Finance*, J. Eatwell, M. Milgate & P. Newman, eds, W.W. Norton, New York, pp. 227–236.
- Hirshleifer, J. (1963). Investment Decision Under Uncertainty, *Papers and Proceedings of the Seventy-Sixth Annual Meeting*

- Marshall, A. (1890, 1891). *Principles of Economics*, 2nd Edition, Macmillan and Co., London and New York.
- Merton, R.C. (1970). *A Dynamic General Equilibrium Model of the Asset Market and Its Application to the Pricing of the Capital Structure of the Firm*, Working Paper 497-70, Sloan School of Management, MIT, Cambridge; Reprinted as chapter 11 of Merton, R.C. (1990). *Continuous-Time Finance*, Blackwell, Cambridge, pp. 357–387.
- Merton, R.C. (1972). An analytic derivation of the efficient portfolio frontier, *Journal of Financial and Quantitative Analysis* 7, 1851–1872.
- Miller, M.H. & Modigliani, F. (1961). Dividend policy, growth and the valuation of shares, *Journal of Business* 34, 235–264.
- Modigliani, F. & Miller, M.H. (1963). Corporate income taxes and the cost of capital, *The American Economic Review* 53, 433–443.
- Mossin, J. (1968). Optimal multiperiod portfolio policies, *Journal of Business* 4(2), 215–229.
- Mossin, J. (1969a). A note on uncertainty and preferences in a temporal context, *The American Economic Review* 59(1), 172–174.
- Mossin, J. (1969b). Security pricing and investment criteria in competitive markets, *The American Economic Review* 59(5), 749–756.
- Mossin, J. (1973). *Theory of Financial Markets*, Prentice-Hall, Englewood Cliffs.
- Mossin, J. (1977). *The Economic Efficiency of Financial Markets*, Lexington, Lanham.
- von Neumann, J.L. & Morgenstern, O. (1953). *Theory of Games and Economic Behavior*, 3rd Edition, Princeton University Press, Princeton.
- Roy, A.D. (1956). Risk and rank or safety first generalised, *Economica* 23(91), 214–228.
- Rubinstein, M. (1970). Addendum (1970), in *Portfolio Selection: Efficient Diversification of Investments*, Cowles Foundation for Research in Economics at Yale University, Monograph #6, H.M. Markowitz, ed., 1959. John Wiley & Sons, Inc., New York. (2nd Edition, 1991, Basil Blackwell, Inc., Cambridge), pp. 308–315.
- Savage, L.J. (1954). *The Foundations of Statistics*, John Wiley & Sons, New York.
- Sharpe, W.F. (1961a). *Portfolio Analysis Based on a Simplified Model of the Relationships Among Securities*, unpublished doctoral dissertation. University of California at Los Angeles, Los Angeles.
- Sharpe, W.F. (1961b). *A Computer Program for Portfolio Analysis Based on a Simplified Model of the Relationships Among Securities*, unpublished mimeo. University of Washington, Seattle.
- Sharpe, W.F. (1963). A simplified model for portfolio analysis, *Management Science* 9(2), 277–293.
- Sharpe, W.F. (1966). Mutual fund performance, *Journal of Business* 39(Suppl), 119–138.
- Sharpe, W.F. (1970). *Portfolio Theory and Capital Markets*, McGraw-Hill, New York.
- Sharpe, W.F. (1977). The capital asset pricing model: a ‘multi-Beta’ interpretation, in *Financial Decision Making Under Uncertainty*, H. Levy & M. Sarnatt, eds, Harcourt Brace Jovanovich, Academic Press, New York, pp. 127–136.
- Sharpe, W.F. & Alexander, G.J. (1978). *Investments*, 4th Edition, (1990), Prentice-Hall, Englewood Cliffs.
- Taqqi, M.S. (2001). Bachelier and his times: a conversation with Bernard Bru, *Finance and Stochastics* 5(1), 3–32.
- Treynor, J.L. (1963). *Implications for the Theory of Finance*, unpublished manuscript. “Rough Draft” dated by Mr. Treynor to the spring of 1963.
- Treynor, J.L. (1965). How to rate management of investment funds, *Harvard Business Review* 43, 63–75.
- Treynor, J.L. & Black, F. (1973). How to use security analysis to improve portfolio selection, *Journal of Business* 46(1), 66–88.

Related Articles

Bernoulli, Jacob; Black–Litterman Approach; Risk–Return Analysis; Markowitz, Harry; Mutual Funds; Sharpe, William F.

CRAIG W. FRENCH

Long-Term Capital Management

Background

Long-Term Capital Management (LTCM) launched its flagship fund on February 24, 1994, with \$1.125 billion in capital, making it the largest start-up hedge fund to date. Over \$100 million came from the partners themselves, especially those who came from the proprietary trading operation that John Meriwether had headed at Salomon Brothers. At Salomon, the profit generated by this group had regularly exceeded the profit generated by the entire firm, and the idea of LTCM was to continue this record on their own. To help them, they also recruited a dream team of academic talent, most notably Myron Scholes and Robert Merton (*see Merton, Robert C.*), who would win the 1997 Nobel Prize in Economics for their pioneering work in financial economics. But they were not alone; half of the founding partners taught finance at major business schools.

The first few years of the fund continued the success of the Salomon years (Table 1).

The fund was closed to new capital in 1995 and quickly grew to \$7.5 billion of capital by the end of 1997. At this time the partners decided, given the lack of additional opportunities, to pay a dividend of \$2.7 billion, which left the capital at the beginning of 1998 at \$4.8 billion.

Investment Style

The fund invested in relative-value convergence trades. They would buy cheap assets and hedge as many of the systematic risk factors as possible by selling rich assets. The resulting “spread” trade had significantly less risk than the outright trade,

so LTCM would lever the spread trade to raise the overall risk level, as well as the expected return on invested capital.

An example of such a trade is an on-the-run *versus* off-the-run trade. In August 1998, 30-year treasuries (the on-the-run bond) had a yield to maturity of 5.50%. The 29-year bond (the off-the-run issue) was 12 basis points (bp) cheaper, with a yield to maturity of 5.62%. The outright risk of 30-year treasury bonds was a standard deviation of around 85 bp per year. The spread trade only had a risk level of around 3.5 bp per year, so the spread trade could be levered 25 to 30 to 1, bringing it in line with the market risk of 30-year treasuries.

LTCM would never do a trade that mathematically looked attractive according to its models unless they qualitatively understood why the trade worked and what were the forces that would bring the “spreads” to convergence. In the case of the on-the-run versus off-the-run trade, the main force leading to a difference in yields between the two bonds is liquidity. The 30-year bond is priced higher by 12 bp (approximately 1.2 points on a par bond) because some investors are willing to pay more to own a more liquid bond. But in six months’ time, when the treasury issues a new 30-year bond, that new bond will be the most liquid one and the old 30-year bond will lose its liquidity premium. This means that in six months’ time, it will trade at a yield similar to that of the old 29-year bond, thus bringing about a convergence of the spread.

LTCM was involved in many such relative-value trades, in many different and seemingly unrelated markets and instruments. These included trades in Government bond spreads, swap spreads, yield curve arbitrage, mortgage arbitrage, volatility spreads, risk arbitrage, and equity relative value trades. In each case, the bet was that some spread would converge over time.

Risk Management

LTCM knew that a major risk to pursuing relative-value convergence trades was the ability to hold the trades until they converged. To ensure this, LTCM insisted that investors lock in equity capital for 3 years, so there would be no premature liquidation from investor cashout. This equity lock-in also gave counterparties comfort that LTCM had long-lasting

Table 1 LTCM returns

Year	Net return (%)	Gross return (%)	Dollar profits (\$)	Ending capital (\$)
1994	20	28	0.4	1.6
1995	43	59	1.3	3.6
1996	41	57	2.1	5.2
1997	17	25	1.4	7.5

credit worthiness, and that enabled LTCM to acquire preferential financing.

As a further protection, LTCM also made extensive use of term financing. If the on-the-run/off-the-run trade might take six months to converge, LTCM would finance the securities for six months, instead of rolling the financing overnight. LTCM also had a two-way mark-to-market provisions in all of its over-the-counter contracts. Thus for its relative value trades that consisted of both securities and contractual agreements it had fully symmetric marks, so that the only time LTCM had to put additional equity capital into a trade was if the spreads widened out. The fund also had term debt and backstop credit lines in place as alternative funding.

LTCM also stress tested its portfolio relative to potential economic shocks to the system, and hedged against the consequences. As an example, in 1995, LTCM had a large swapped position in Italian government bonds. The firm got very worried that if the Republic of Italy defaulted, it would have a sizable loss. So it purchased insurance against this potential default by doing a credit default swap on the Italian government bonds.

But the primary source of risk management relied on the benefit that the portfolio obtained due to

diversification. If the relative value strategies had very low correlations with each other, then the risk of the overall portfolio would be low. LTCM assumed that in the long run these correlations were low because of the loose economic ties between the trades, although in the short run these correlations could be significantly higher. LTCM also assumed that the downside risk on some of the trades was diminished, as spreads got very wide, on the assumption that other leveraged funds would rush in to take advantage. In retrospect, these assumptions were all falsified by experience.

Before the crisis, LTCM had a historical risk level of a \$45 million daily standard deviation of return on the fund. See Figure 1 for historical daily returns.

After the fund reached global scale in 1995, the risk level was remarkably stable. In fact, the partners had actually predicted a higher risk level for the fund as they assumed that the correlations among the relative value trades would be higher than historical levels. But in 1998, all this changed.

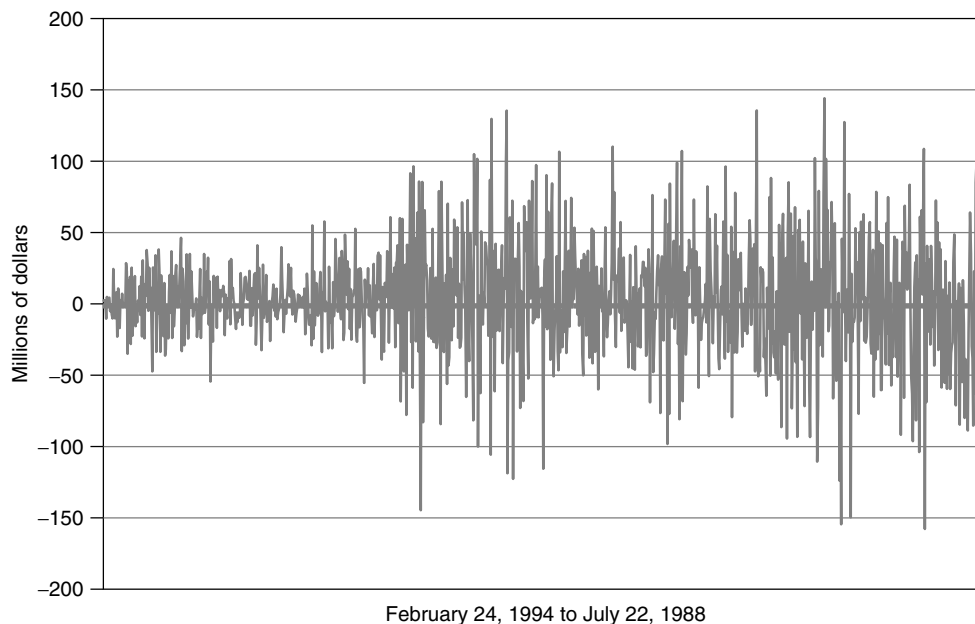


Figure 1 Historical daily returns

The 1998 Crisis

In 1998, LTCM was up slightly in the first four months of the year. Then, in May, the portfolio lost 6% and in June, it lost 10%. In early July, the portfolio rebounded by about 7% and the partners reduced the underlying risk of the portfolio accordingly by about 10%.

The crisis was triggered by the Russian default on its domestic bonds on August 17, 1998. While LTCM did not have many Russian positions so that its direct losses were small, the default did initiate the process that was to follow as unrelated markets all over the world reacted. On Friday August 21, LTCM had a one-day loss of \$550 million. (A risk arb deal that was set to close on that day, that of Ciena and Tellabs, broke, causing a \$160 million loss. Swap spreads that normally move about 1 bp a day were out 21 bp intraday.) The Russian debt crisis had triggered a flight out of all relative-value positions. In the illiquid days at the end of August, these liquidations caused a downward spiral as new losses led to more liquidations and more losses. The result was that by the end of August LTCM was down by 53% for the year, with the capital now at \$2.3 billion.

While the Russian default triggered the economic crisis in August, it was an LTCM crisis in September. Would the fund fail? Many other institutions with similar positions liquidated them in advance of the potential failure. Some market participants bet against the firm and counterparties marked contractual agreements at extremely wide levels to obtain additional cushions against bankruptcy. The partners hired Goldman Sachs to help them raise additional capital and to sell off assets; for this, they received 50% of the management company.

The leverage of the firm went to an enormous levels involuntarily (Figure 2), not because of increase in assets but because of equity falling. In the event, attempts to raise additional funds failed and on Monday, September 21, the fund lost another \$550 million, putting its capital for the first time below \$1 billion. On Wednesday, at the behest of the Federal Reserve, the 15 major counterparties met at the New York Fed to discuss the situation.

During the meeting, at 11:00 AM the partners received a telephone call from Warren Buffett, who was on a satellite phone while vacationing with Bill Gates in Alaska. He said that LTCM was about to receive a bid on its entire portfolio from him and that he hoped they would seriously consider it. At 11:30 AM LTCM received the fax message given in Figure 3.

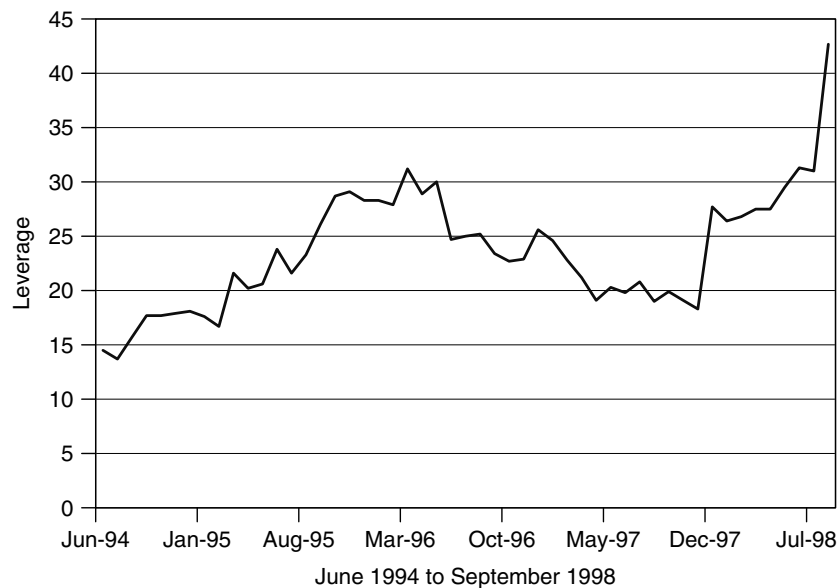


Figure 2 Leverage

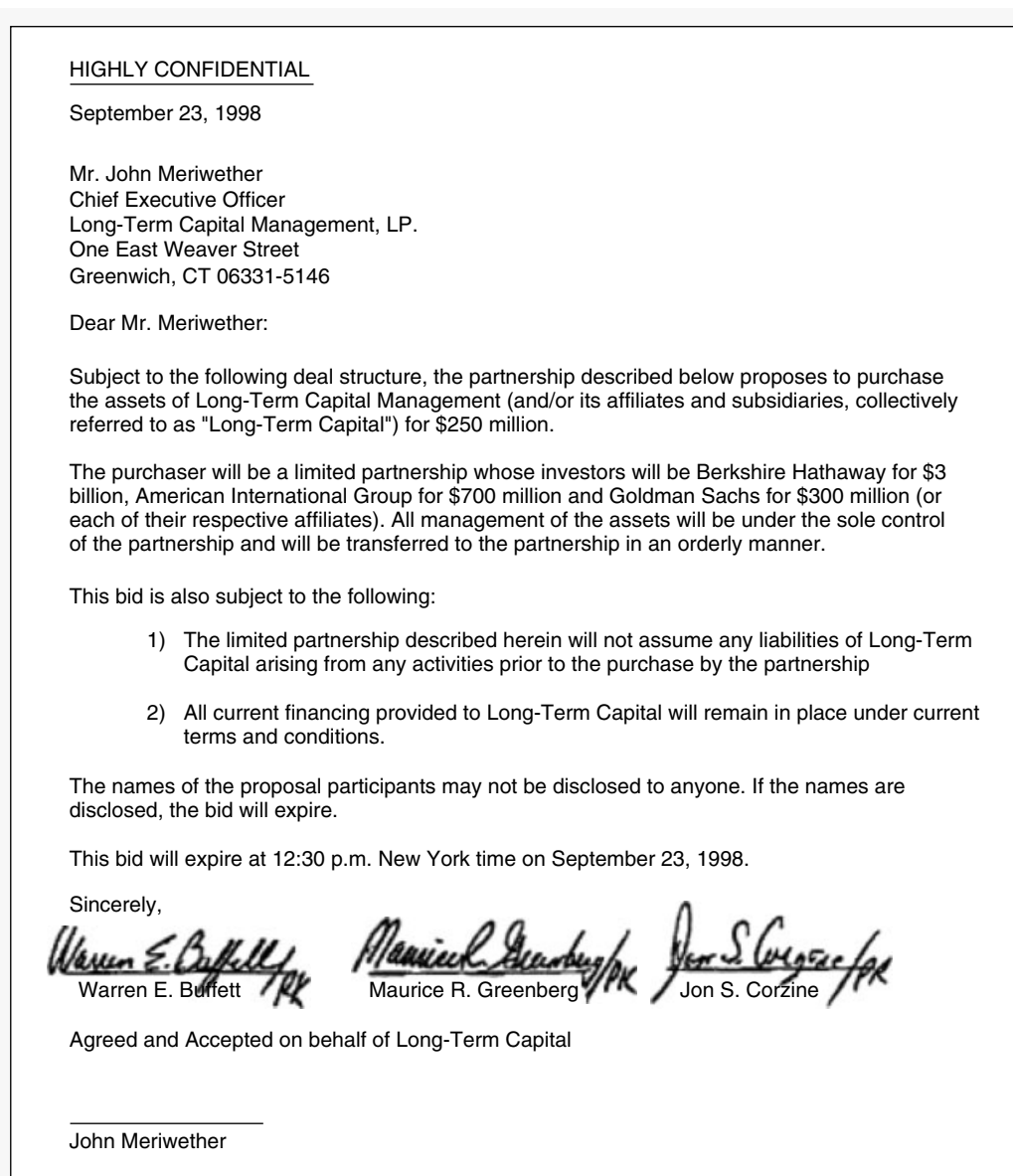


Figure 3 Copy of the \$250 million offer for Long-Term Capital Management

The partners were unable to accept the proposal as it was crafted. The fund had approximately 15 000 distinct positions. Each of these positions was a credit counterparty transaction (i.e., a repo or swap

contract). Transfer of those positions to the Buffett-led group would require the approval of all the counterparties. Clearly, all of LTCM's counterparties would prefer to have Warren Buffett as a creditor

as opposed to an about-to-be-bankrupt hedge fund. But it was going to be next to impossible to obtain complete approval in one hour.

The partners proposed, as an alternative, that the group make an emergency equity infusion into the fund in return for 90% ownership and the right to kick the partners out as managers. Under this plan, all the financing would stay in place and the third party investors could be redeemed at anytime. Unfortunately, the lawyers were not able to get Buffett back on his satellite phone and no one was prepared to consummate the deal without his approval.

At the end of the day, 14 financial institutions (everyone with the exception of Bear Stearns) agreed to make an emergency \$3.625 billion equity infusion into the fund. The plan was essentially a no-fault bankruptcy where the creditors of a company (in this case, the secured creditors) make an equity investment, cramming down the old equity holders, in order to liquidate the company in an orderly manner.

Why did the Fed orchestrate the bailout? The answer has to do with how the bankruptcy laws are applied with respect to financial firms. When LTCM did the on-the-run versus off-the-run strategy, the risk of the two sides of the trade netted within the fund. But in bankruptcy, each side of the trade liquidates its collateral separately, and sends a bill to LTCM. The risk involved in the position is thus no longer netted at 3.5 bp but is actually 85 bp per side. Although the netted risk of LTCM was \$45 million per day, the gross risk was much larger, more like \$30 million per day with each of 15 counterparties.

As conditions worsened, early in September, the partners had been going around to the counterparties and explaining this enormous potential risk factor in the event of bankruptcy and the large losses that the counterparties would potentially face. They separately asked each dealer to make an equity infusion to shore up LTCM's capital situation. But it was a classic Prisoner's Dilemma problem. No dealer would commit unless everyone else did. It was necessary to get everyone in the same room, so that they would all know the full extent of the exposures and all commit together, and that could not happen until bankruptcy was imminent.

In this event, the private bailout was a success. No counterparty had any losses on their collateral. By the end of the first quarter of 1999, the fund had rallied 25% from its value at the time of the

bailout. At that time third-party investors were paid off. The consortium of banks decided to continue the liquidation at a faster pace and, by December 1999, the liquidation was complete. The banks had no losses and had made a 10% return on their investment.

Investors who had made a \$1 investment at the beginning of 1998 would have seen their investment fall to 8 cents at the time of the bailout, and would have received 10 cents on April 1, 1999. But in its earlier years, LTCM had made high returns and paid out high dividends such that of its 100 investors only 12 actually lost money, and only 6 lost more than \$2 million. The median investor actually had a 19% internal rate of return (IRR) even including the loss. The partners did not fare as well. Their capital was about \$2 billion at the beginning of 1998 and they received no final payout.

Lessons Learned

The LTCM crisis illustrates some of the pitfalls of a VaR-based risk management system (*see Value-at-Risk*), where the risk of the portfolio is determined by the exogenous economic relationships among the trades. During the crisis, all of LTCM's trades moved together with correlations approaching one, even though the trades were economically diverse. It was hard to believe that the returns from US mortgage arbitrage trades would be highly related to LTCM's Japanese warrant and convertible book or highly related to their European government bond spread trades. Yet, during the crisis these correlations all moved toward one, resulting in a failure of diversification and creating enormous risk for the fund.

What was the common thread in all of these trades? It was not that they were economically related, but more that they had similar holders of the trades with common risk tolerances. When these hedge funds and proprietary trading groups at the banks lost money in the Russian crisis they were ordered by senior management to reduce their risk exposures. The trades that they took off were the relative-value trades. As they unwound their positions in the illiquid days of August, the spreads went out further, causing more losses and further unwinds.

This risk might be better classified as endogenous risk, risk that comes about not from the fundamental

economic relationships of the cash flows of the securities but in a crisis through the common movements of the holders of the trades. Prudent risk management practices need to manage the portfolio risk not just for normal times but for crisis times, taking into account the endogenous aspect of risk.

Related Articles

Merton, Robert C.; Risk Management: Historical Perspectives; Value-at-Risk.

ERIC ROSENFELD

Bubbles and Crashes

The two acclaimed classic books—Galbraith’s “The Great Crash 1929” [40] and Kindleberger’s “Manias, Panics and Crash” [61]—provide the most commonly accepted explanation of the 1929 boom and crash. Galbraith argues that a bubble was formed in the stock market during the rapid economic growth in the 1920s. Both he and Kindleberger, in his extensive historical compendium of financial excesses, emphasize the irrational element—the mania—that induced the public to invest in the bull “overheating” market. The rise in the stock market, according to Galbraith’s account (1954 and 1988, pp. xii-xiii), depended on “the vested interest in euphoria [that] leads men and women, individuals and institutions to believe that all will be better, that they are meant to be richer and to dismiss as intellectually deficient what is in conflict with that conviction.” This eagerness to buy stocks was then fueled by an expansion of credit in the form of brokers’ loans that encouraged investors to become dangerously leveraged. In this respect, Shiller [91] argues that the increase in stock price was driven by irrational euphoria among individual investors, fed by emphatic media, which maximized TV ratings and catered to investor demand for pseudonews.

Kindleberger [61] summarizes his compilation of many historical bubbles as follows.

- The upswing usually starts with an opportunity—new markets, new technologies, or some significant political change—and investors looking for good returns.
- It proceeds through the euphoria of rising prices, particularly of assets, while an expansion of credit inflates the bubble.
- In the manic phase, investors scramble to get out of money and into illiquid investments such as stocks, commodities, real estate, or tulip bulbs: “a larger and larger group of people seeks to become rich without a real understanding of the processes involved.”
- Ultimately, the markets stop rising and people who have borrowed heavily find themselves overstretched. This is “distress”, which generates unexpected failures, followed by “revulsion” or “discredit”.
- The final phase is a self-feeding panic, where the bubble bursts. People of wealth and credit

scramble to unload whatever they have bought at greater and greater losses and cash becomes king.

Although this makes for compelling reading, many questions remain unanswered. There is little consideration about how much fundamentals contributed to the bull market and what might have triggered the speculative mania. Galbraith [40] cited margin buying, the formation of closed-end investment trusts, the transformation of financiers into celebrities, and other qualitative signs of euphoria to support his view. Recent evidence supports the concept of the growth of a social procyclical mood that promotes the attraction for investing in the stock markets by a larger and larger fraction of the population as the bubble grows [88].

Furthermore, Galbraith’s and Kindleberger’s accounts are vague about the causes of the market crash, believing that almost any event could have triggered irrational investors to sell toward the end of bubble, not really explaining the reason for the crash. Instead, they sidestep the thorny question of the occurrence and timing of the crash by focusing on the inevitability of the bubble’s collapse and suggest several factors that could have exploded public confidence and caused prices to plummet. Furthermore, little has been done to identify the precise role of external events in provoking the collapse.

In the words of Shiller [91], a crash is a time when “the investing public en masse capriciously changes its mind.” However, as with the more rational theories, this explanation again leaves unanswered the question of why such tremendous capricious changes in sentiment occur. Alternatively, it amounts to sur-rendering the explanation to the vagaries of “capricious changes”. Other studies have argued that even though fundamentals appeared high in 1929, Fisher [35], for example, argued throughout 1929 and 1930 that the high level of prices in 1929 reflected an expectation that future corporate cash flows would be very high. Fisher believed this expectation to be warranted after a decade of steadily increasing earnings and dividends, of rapidly improving technologies, and of monetary stability. In hindsight, it has become clear that even though fundamentals appeared high in 1929, the stock market rise was clearly excessive. A recent empirical study [25] concludes that the stocks making up the S&P500 composite were priced at least

30% above fundamentals in late summer 1929. White [107] suggests that the 1929 boom cannot be readily explained by fundamentals, represented by expected dividend growth or changes in the equity premium.

While Galbraith's and Kindleberger's classical views have been most often cited by the mass media, they had received little scholarly attention. Since the 1960s, in parallel with the emergence of the efficient-market hypothesis, their position has lost ground among economists and especially among financial economists. More recent works, described at the end of this article, revive their views in the form of quantitative diagnostics.

Efficient-market Hypothesis

The efficient-markets hypothesis (*see Efficient Market Hypothesis*) states that asset prices reflect fundamental value, defined as the discounted sum of expected future cash flows where, in forming expectations, investors "correctly process" all available information. Therefore, in an efficient market, there is "no free lunch": no investment strategy can earn excess risk-adjusted average returns or average returns greater than are warranted for its risk. Proponents of the efficient-markets hypothesis, Friedman and Schwartz [39] and Fama, [34], argue that rational speculative activity would eliminate riskless arbitrage opportunities. Fama ([34], p.38) states that, if there are many sophisticated traders in the market, they may cause these bubbles to burst before they have a chance to really get under way.

However, after years of effort, it has become clear that some basic empirical facts about the stock markets cannot be understood in this framework [106]. The efficient-markets hypothesis entirely lost ground after the burst of the Internet bubble in 2000, providing one of the recent most striking episodes of anomalous price behavior and volatility in one of the most developed capital markets of the world. The movement of Internet stock prices during the late 1990s was extraordinary in many respects. The Internet sector earned over 1000% returns on its public equity in the two-year period from early 1998 through February 2000. The valuations of these stocks began to collapse shortly thereafter and by the end of the same year, they had returned to pre-1998 levels, losing nearly 70% from the peak. The extraordinary returns of 1998–February 2000 had

largely disappeared by the end of 2000. Although in February 2000 the vast majority of Internet-related companies had negative earnings, the Internet sector in the United States was equal to 6% of the market capitalization of all US public companies and 20% of the publicly traded volume of the US stock market [82, 83].

Ofek and Richardson [83] used the financial data from 400 companies in the Internet-related sectors and analyzed to what extent their stock prices differed from their fundamental values estimated by using Miller and Modigliani [79] model for stock valuation [38]. Since almost all companies in the Internet sector had negative earnings, they estimated the (implied) price-to-earnings (P/E) ratios, which are derived from the revenue streams of these firms rather than their earnings that would be read from the 1999 financial data. Their results are striking. Almost 20% of the Internet-related firms have P/E ratios in excess of 1500, while over 50% exceed 500, and the aggregate P/E ratio of the entire Internet sector is 605. Under the assumptions that the aggregate long-run P/E ratio is 20 on average (which is already on the large end member from a historical point of view), the Internet sector would have needed to generate 40.6% excess returns over a 10-year period to justify the P/E ratio of 605 implied in 2000. The vast majority of the implied P/E s are much too high relative to the P/E s usually obtained by firms. By almost any standard, this clearly represented "irrational" valuation levels. These and similar figures led many to believe that this set of stocks was in the midst of an asset price bubble.

From the theoretical point of view, some rational equilibrium asset-pricing models allow for the presence of bubbles, as pointed out for infinite-horizon models in discrete-time setups by Blanchard and Watson [9]. Loewenstein and Willard [70, 71] characterized the necessary and sufficient conditions for the absence of bubbles in complete and incomplete markets equilibria with several types of borrowing constraints and in which agents are allowed to trade continuously. For zero net supply assets, including financial derivatives with finite maturities, they show that bubbles can generally exist and have properties different from their discrete-time, infinite-horizon counterparts. However, Lux and Sornette [73] demonstrated that exogenous rational bubbles are hardly reconcilable with some of the stylized facts of financial data at a very elementary level.

Jarrow *et al.* [53] showed that if financial agents prefer more to less (no dominance assumption), then bubbles in complete markets can only exist which are uniformly integrable martingales, and these can exist with an infinite lifetime. Under these conditions, the put–call parity holds and there are no bubbles in standard call and put options. Their analysis implies that if one believes that asset price bubbles exist, then asset markets must be incomplete. Jarrow *et al.* [54] extend their discussion in [53] to characterize all possible price bubbles in an incomplete market, satisfying the “no free lunch with vanishing risk” and “no dominance” assumptions. Their [54] new theory for bubbles is formulated in terms of different local martingale measures across time, which leads to some testable predictions on derivative pricing in the presence of bubbles.

Heterogeneous Beliefs and Limits to Arbitrage

The collapsing Internet bubble has thrown new light on the old subject and raised the acute question of why rational investors have not moved earlier into the market and driven the Internet stock prices back to their fundamental valuations.

Two conditions are, in general, invoked as being necessary for prices to deviate from the fundamental value. First, there must be some degree of irrationality in the market; that is, investors’ demand for stocks must be driven by something other than fundamentals, such as overconfidence in the future. Second, even if a market has such investors, the general argument is that rational investors will drive prices back to fundamental value. To avoid this, there needs to be some limit on arbitrage. Shleifer and Vishny [92] provide a description for various limits of arbitrage. With respect to the equity market, clearly the most important impediment to arbitrage is short-sales restrictions. Roughly 70% of mutual funds explicitly state (in the Securities and Exchange Commission (SEC) form N-SAR) that they are not permitted to sell short [2]. Seventy-nine percent of equity mutual funds make no use of derivatives whatsoever (either futures or options), suggesting further that funds do not take synthetically short positions [64]. These figures indicate that the vast majority of funds never take short positions.

Recognizing that the world has limited arbitrage and significant numbers of irrational investors,

the finance literature has evolved to increasingly recognize the evidence of deviations from the fundamental value. One important class of theories shows that there can be large movements in asset prices caused by the combined effects of heterogeneous beliefs and short-sales constraints. The basic idea finds its root back to the original capital asset pricing model (CAPM) theories, in particular, to Lintner’s model of asset prices with investors having heterogeneous beliefs [69]. In his model, asset prices are a weighted average of beliefs about asset payoffs with the weights being determined by the investor’s risk aversion and beliefs about asset price covariances. Lintner [69] and many others after him show that widely inflated prices can occur.

Many other asset-pricing models in the spirit of Lintner [69] have been proposed [19, 29, 48, 52, 78, 89]. In these models that assume heterogeneous beliefs and short-sales restrictions, the asset prices are determined at equilibrium to the extent that they reflect the heterogeneous beliefs about payoffs, but short-sales restrictions force the pessimistic investors out of the market, leaving only optimistic investors and thus inflated asset price levels. However, when short-sales restrictions no longer bind investors, then prices fall. This provides a possible account of the bursting of the Internet bubble that developed in 1998–2000. As documented by Ofek and Richardson [83], and by Cochrane [20], typically as much as 80% of Internet-related shares were locked up. This is due to the fact that many Internet companies had gone through recent initial public offerings (IPOs) and regulations impose that shares held by insiders and other pre-IPO equity holders cannot be traded for at least six months after the IPO date. The float of the Internet sector dramatically increased as the lockups of many of these stocks expired. The unlocking of literally hundreds of billions of dollars of shares in the Internet sector in Spring 2000 was equivalent of removing short-sales restrictions. And the collapse of Internet stock prices coincided with a dramatic expansion in the number of publicly tradable shares of Internet companies. Among many others, Hong *et al.* [49] explicitly model the relationship between the number of publicly tradable shares of an asset and the propensity for speculative bubbles to form. So far, the theoretical models based on agents with heterogeneous beliefs facing short-sales restrictions are considered among the most

convincing models to explain the burst of the Internet bubbles.

Another test of this hypothesis on the origin of the 2000 market crash is provided by the search for possible discrepancies between option and stock prices. Indeed, even though it is difficult for rational investors to borrow Internet stocks for short selling due to the lockup period discussed above, they should have been able to construct equivalent synthetic short positions by purchasing puts and writing calls in the option market and either borrowing or lending cash, without the need for borrowing the stocks. The question is now transformed into finding some evidence for the use or the absence of such strategy and the reason for its absence in the latter case. One possible thread is that, if short selling through option positions was difficult or impractical, prices in the stock and options markets should decouple [67]. Using a sample of closing bid and ask prices for 9026 option pairs for three days in February 2000 along with closing trade prices for the underlying equities, Ofek and Richardson [83] find that 36% of the Internet stocks had put–call parity violations as compared to only 23.8% of the other stocks. One reason for put–call parity violations may be that short-sale restrictions prevent arbitrage from equilibrating option and stock prices. Hence, one interpretation of the finding that there are more put–call parity violations for Internet stocks is that short-sale constraints are more frequently binding for Internet stocks. Furthermore, Ofek *et al.* [84] provide a comprehensive comparison of the prices of stocks and options, using closing options quotes and closing trades on the underlying stock for July 1999 through November 2001. They find that there are large differences between the synthetic stock price and the actual stock price, which implies the presence of apparent arbitrage opportunities involving selling actual shares and buying synthetic shares. They interpret their findings as evidence that short-sale constraints provide meaningful limits to arbitrage that can allow prices of identical assets to diverge.

By defining a bubble as a price process that, when discounted, is a local martingale under the risk-neutral measure but not a martingale, Cox and Hobson [21] provide a complementary explanation for the failure of put–call parity. Intuitively, the local martingale model views a bubble as a stopped stochastic process for which the expectation exhibits a discontinuity when it ends. It can then be shown

that several standard results fail for local martingales: put–call parity does not hold, the price of an American call exceeds that of a European call, and call prices are no longer increasing in maturity (for a fixed strike).

Thus, it would seem that the issue of the origin of the 2000 crash is settled. However, Battalio and Schultz [6] arrive at the opposite conclusion, using proprietary intraday option trade and quote data generated in the days surrounding the collapse of the Internet bubble. They find that the general public could cheaply short synthetically using options, and this information could have been transmitted to the stock market, in line with the absence of evidence that synthetic stock prices diverged from actual stock prices. The difference between the work of Ofek and Richardson [83] and Ofek *et al.* [84], on the one hand, and Battalio and Schultz [6], on the other, is that the former used closing option quotes and last stock trade prices from the OptionMetrics Ivy database. As pointed out by Battalio and Schultz [6], OptionMetrics matches closing stock trades that occurred no later than 4:00 pm, and perhaps much earlier, with closing option quotes posted at 4:02 pm. Furthermore, option market makers that post closing quotes on day t are not required to trade at those quotes on day $t + 1$. Likewise, dealers and specialists in the underlying stocks have no obligation to execute incoming orders at the price of the most recent transaction. Hence, closing option quotes and closing stock prices obtained from the OptionMetrics database do not represent contemporaneous prices at which investors could have simultaneously traded. To address this problem, Battalio and Schultz [6] use a unique set of intraday option price data. They first ensure that the synthetic and the actual stock prices that they compare are synchronous, and then, they discard quotes that, according to exchange rules, are only indicative of the prices at which liquidity demanders could have traded. They find that almost all of the remaining apparent put–call parity violations disappear when they discard locked or crossed quotes and quotes from fast options markets. In other words, the apparent arbitrage opportunities almost always arise from quotes upon which investors could not actually trade. Battalio and Schultz [6] conclude that short-sale constraints were not responsible for the high prices of Internet stocks at the peak of the bubble and that small investors could have

sold short synthetically using options, and this information would have been transmitted to the stock market. The fact that investors did not take advantage of these opportunities to profit from overpriced Internet stocks suggests that the overpricing was not as obvious then as it is now, with the benefit of hindsight. Schultz [90] provides additional evidence that contemporaneous lockup expirations and equity offerings do not explain the collapse of Internet stocks because the stocks that were restricted to a fixed supply of shares by lockup provisions actually performed worse than stocks with an increasing supply of shares. This shows that current explanations for the collapse of Internet stocks are incomplete.

Riding Bubbles

One cannot understand crashes without knowing the origin of bubbles. In a nutshell, speculative bubbles are caused by “precipitating factors” that change public opinion about markets or that have an immediate impact on demand and by “amplification mechanisms” that take the form of price-to-price feedback, as stressed by Shiller [91]. Consider the example of a housing-market bubble. A number of fundamental factors can influence price movements in housing markets. The following characteristics have been shown to influence the demand for housing: demographics, income growth, employment growth, changes in financing mechanisms, interest rates, as well as changes in the characteristics of the geographic location such as accessibility, schools, or crime, to name a few. On the supply side, attention has been paid to construction costs, the age of the housing stock, and the industrial organization of the housing market. The elasticity of supply has been shown to be a critical factor in the cyclical behavior of home prices. The cyclical process that we observed in the 1980s in those cities experiencing boom-and-bust cycles was caused by the general economic expansion, best proxied by employment gains, which drove up the demand. In the short run, those increases in demand encountered an inelastic supply of housing and developable land, inventories of for-sale properties shrank, and vacancy declined. As a consequence, prices accelerated. This provided an amplification mechanism as it led buyers to anticipate further gains, and the

bubble was born. Once prices overshoot or supply catches up, inventories begin to rise, time on the market increases, vacancy rises, and price increases slow down, eventually encountering downward stickiness. The predominant story about home prices is always the prices themselves [91, 93]; the feedback from initial price increases to further price increases is a mechanism that amplifies the effects of the precipitating factors. If prices are going up rapidly, there is much word-of-mouth communication, a hallmark of a bubble. The word of mouth can spread optimistic stories and thus help cause an overreaction to other stories, such as ones about employment. The amplification can work on the downside as well.

Hedge funds are among the most sophisticated investors, probably closer to the ideal of “rational arbitrageurs” than any other class of investors. It is therefore particularly telling that successful hedge-fund managers have been repeatedly reported to ride rather than attack bubbles, suggesting the existence of mechanisms that entice rational investors to surf bubbles rather than attempt to arbitrage them. However, the evidence may not be that strong and could even be circular, since only successful hedge-fund managers would survive a given 2–5 year period, opening the possibility that the mentioned evidence could result in large part from a survival bias [14, 44]. Keeping this in mind, we now discuss two classes of models, which attempt to justify why sophisticated “rational” traders would be willing to ride bubbles. These models share a common theme: rational investors try to ride bubbles, and the incentive to ride the bubble stems from predictable “sentiment”—anticipation of continuing bubble growth [1] and predictable feedback trader demand [26, 27]. An important implication of these theories is that rational investors should be able to reap gains from riding a bubble at the expense of less-sophisticated investors.

Positive Feedback Trading by Noise Traders

The term *noise traders* was introduced first by Kyle [65] and Black [8] to describe irrational investors. Thereafter, many scholars exploited this concept to extend the standard models by introducing the simplest possible heterogeneity in terms

of two interacting populations of rational and irrational agents. One can say that the one-representative-agent theory is being progressively replaced by a two-representative-agents theory, analogously to the progress from the one-body to the two-body problems in astronomy.

De Long *et al.* [26, 27] introduced a model of market bubbles and crashes, which exploits this idea of the possible role of noise traders in the development of bubbles as a possible mechanism for why asset prices may deviate from the fundamentals over rather long time periods. Their inspiration came from the observation of successful investors such as George Soros, who reveal that they often exploit naive investors following positive feedback strategies or momentum investment strategies. Positive feedback investors are those who buy securities when prices rise and sell when prices fall. In the words of Jegadeesh and Titman [55], positive feedback investors are buying winners and selling losers. In a description of his own investment strategy, Soros [101] stresses that the key to his success was not to counter the irrational wave of enthusiasm that appears in financial markets, but rather to *ride this wave* for a while and sell out much later. The model of De Long *et al.* [26, 27] assumes that when rational speculators receive good news and trade on this news, they recognize that the initial price increase will stimulate buying by noise traders who will follow positive feedback trading strategies with a delay. In anticipation of these purchases, rational speculators buy more today, and so drive prices up today higher than fundamental news warrants. Tomorrow, noise traders buy in response to increase in today's price and so keep prices above the fundamentals. The key point is that trading between rational arbitrageurs and positive feedback traders gives rise to bubble-like price patterns. In their model, rational speculators destabilize prices because their trading triggers positive feedback trading by other investors. Positive feedback trading reinforced by arbitrageurs' jumping on the bandwagon leads to a positive autocorrelation of returns at short horizons. Eventually, selling out or going short by rational speculators will pull the prices back to the fundamentals, entailing a negative autocorrelation of returns at longer horizons. In summary, De Long *et al.* [26, 27] model suggests the coexistence of intermediate-horizon momentum and long-horizon reversals in stock returns.

Their work was followed by a number of behavioral models based on the idea that trend chasing by one class of agents produces momentum in stock prices [5, 22, 50]. The most influential empirical evidence on momentum strategies came from the work of Jegadeesh and Titman [55, 56], who established that stock returns exhibit momentum behavior at intermediate horizons. Strategies that buy stocks that have performed well in the past and sell stocks that have performed poorly in the past generate significant positive returns over 3- to 12-month holding periods. De Bondt and Thaler [24] documented long-term reversals in stock returns. Stocks that perform poorly in the past perform better over the next 3–5 years than stocks that perform well in the past. These findings present a serious challenge to the view that markets are semistrong-form efficient.

In practice, do investors engage in momentum trading? A growing number of empirical studies address momentum trading by investors, with somewhat conflicting results. Lakonishok *et al.* [66] analyzed the quarterly holdings of a sample of pension funds and found little evidence of momentum trading. Grinblatt *et al.* [45] examined the quarterly holdings of 274 mutual funds and found that 77% of the funds in their sample engaged in momentum trading [105]. Nofsinger and Sias [81] examined total institutional holdings of individual stocks and found evidence of intraperiod momentum trading. Using a different sample, Gompers and Metrick [41] investigated the relationship between institutional holdings and lagged returns and concluded that once they controlled for the firm size, there was no evidence of momentum trading. Griffin *et al.* [43] reported that, on a daily and intraday basis, institutional investors engaged in trend chasing in NASDAQ 100 stocks. Finally, Badrinath and Wahal [4] documented the equity trading practices of approximately 1200 institutions from the third quarter of 1987 through the third quarter of 1995. They decomposed trading by institutions into (i) the initiation of new positions (entry), (ii) the termination of previous positions (exit), and (iii) the adjustments to ongoing holdings. Institutions were found to act as momentum traders when they enter stocks but as contrarian traders when they exit or make adjustments to ongoing holdings. Badrinath and Wahal [4] found significant differences in trading practices among different types of institutions. These studies are limited in their ability to capture the full range of trading

practices, in part because they focus almost exclusively on the behavior of institutional investors. In summary, many experimental studies and surveys suggest that positive feedback trading exists in greater or lesser degrees.

Synchronization Failures among Rational Traders

Abreu and Brunnermeier [1] propose a completely different mechanism justifying why rational traders ride rather than arbitrage bubbles. They consider a market where arbitrageurs face synchronization risk and, as a consequence, delay usage of arbitrage opportunities. Rational arbitrageurs are supposed to know that the market will eventually collapse. They know that the bubble will burst as soon as a sufficient number of (rational) traders will sell out. However, the dispersion of rational arbitrageurs' opinions on market timing and the consequent uncertainty on the synchronization of their sell-off are delaying this collapse, allowing the bubble to grow. In this framework, bubbles persist in the short and intermediate term because short sellers face synchronization risk, that is, uncertainty regarding the timing of the correction. As a result, arbitrageurs who conclude that other arbitrageurs are yet unlikely to trade against the bubble find it optimal to ride the still growing bubble for a while.

Like other institutional investors, hedge funds with large holdings in US equities have to report their quarterly equity positions to the SEC on Form 13F. Brunnermeier and Nagel [15] extracted hedge-fund holdings from these data, including those of well-known managers such as Soros, Tiger, Tudor, and others in the period from 1998 to 2000. They found that, over the sample period 1998–2000, hedge-fund portfolios were heavily tilted toward highly priced technology stocks. The proportion of their overall stock holdings devoted to this segment was higher than the corresponding weight of technology stocks in the market portfolio. In addition, the hedge funds in their sample skillfully anticipated price peaks of individual technology stocks. On a stock-by-stock basis, hedge funds started cutting back their holdings before prices collapsed, switching to technology stocks that still experienced rising prices. As a result, hedge-fund managers captured the upturn, but avoided much of the downturn. This

is reflected in the fact that hedge funds earned substantial excess returns in the technology segment of the NASDAQ.

Complex Systems Approach to Bubbles and Crashes

Bhattacharya and Yu [7] provide a summary of recent efforts to expand on the above concepts, in particular, to address the two main questions of (i) the cause(s) of bubbles and crashes and (ii) the possibility to diagnose them *ex ante*. Many financial economists recognize that positive feedbacks and, in particular, herding are the key factors for the growth of bubbles. Herding can result from a variety of mechanisms, such as anticipation by rational investors of noise traders' strategies [26, 27], agency costs and monetary incentives given to competing fund managers [23] sometimes leading to the extreme Ponzi schemes [28], rational imitation in the presence of uncertainty [88], and social imitation.

The Madoff Ponzi scheme is a significant recent illustration, revealed by the unfolding of the financial crisis that started in 2007 [97]. It is the world's biggest fraud allegedly perpetrated by long-time investment adviser Bernard Madoff, arrested on December 11, 2008 and sentenced on June 29, 2009 to 150 years in prison, the maximum allowed. His fraud led to 65 billion US dollars losses that caused reverberations around the world as the list of victims included many wealthy private investors, charities, hedge funds, and major banks in the United States, Europe, and Asia. The Madoff Ponzi scheme surfed on the general psychology, characterizing the first decade of the twenty-first century, of exorbitant unsustainable expected financial gains. It is a remarkable illustration of the problem of implementing sound risk management, due diligence processes, and of the capabilities of the SEC, the US markets watchdog, when markets are booming and there is a general sentiment of a new economy and new financial era, in which old rules are believed not to apply anymore [75]. Actually, the Madoff Ponzi scheme is only the largest of a surprising number of other Ponzi schemes revealed by the financial crisis in many different countries (see accounts from village.albourne.com).

Discussing social imitation is often considered off-stream among financial economists but warrants

some scrutiny, given its pervasive presence in human affairs. On the question of the *ex ante* detection of bubbles, Gurkaynak [46] summarizes the dismal state of the econometric approach, stating that the “econometric detection of asset price bubbles cannot be achieved with a satisfactory degree of certainty. For each paper that finds evidence of bubbles, there is another one that fits the data equally well without allowing for a bubble. We are still unable to distinguish bubbles from time-varying or regime-switching fundamentals, while many small sample econometrics problems of bubble tests remain unresolved.” The following discusses an arguably off-stream approach that, by using concepts and tools from the theory of complex systems and statistical physics, suggests that *ex ante* diagnostic and partial predictability might be possible [93].

Social Mimetism, Collective Phenomena, Bifurcations, and Phase Transitions

Market behavior is the aggregation of the individual behavior of the many investors participating in it. In an economy of traders with completely rational expectations and the same information sets, no bubbles are possible [104]. Rational bubbles can, however, occur in infinite-horizon models [9], with dynamics of growth and collapse driven by noise traders [57, 59]. However, the key issue is to understand by what detailed mechanism the aggregation of many individual behaviors can give rise to bubbles and crashes. Modeling social imitation and social interactions requires using approaches, little known to financial economists, that address the fundamental question of how global behaviors can emerge at the macroscopic level. This extends the representative agent approach, but it also goes well beyond the introduction of heterogeneous agents. A key insight from statistical physics and complex systems theory is that systems with a large number of interacting agents, open to their environment, self-organize their internal structure and their dynamics with novel and sometimes surprising “emergent” out-of-equilibrium properties. A central property of a complex system is the possible occurrence and coexistence of many large-scale collective behaviors with a very rich structure, resulting from the repeated nonlinear interactions among its constituents.

How can this help address the question of what is/are the cause(s) of bubbles and crashes? The crucial insight is that a system, made of competing investors subjected to the myriad of influences, both exogenous news and endogenous interactions and reflexivity, can develop into endogenously self-organized self-reinforcing regimes, which would qualify as bubbles, and that crashes occur as a global self-organized transition. Mathematicians refer to this behavior as a *bifurcation* or more specifically as a *catastrophe* [103]. Physicists call these phenomena *phase transitions* [102]. The implication of modeling a market crash as a bifurcation is to solve the question of what makes a crash: in the framework of bifurcation theory (or phase transitions), sudden shifts in behavior arise from small changes in circumstances, with qualitative changes in the nature of the solutions that can occur abruptly when the parameters change smoothly. A minor change of circumstances, of interaction strength, or heterogeneity may lead to a sudden and dramatic change, such as during an earthquake and a financial crash.

Most approaches for explaining crashes search for possible mechanisms or effects that operate at very short timescales (hours, days, or weeks at most). According to the “bifurcation” approach, the underlying cause of the crash should be found in the preceding months and years, in the progressively increasing buildup of market cooperativity, or effective interactions between investors, often translated into accelerating ascent of the market price (the bubble). According to this “critical” point of view, the specific manner in which prices collapsed is not the most important problem: a crash occurs because the market has entered an unstable phase and any small disturbance or process may reveal the existence of the instability.

Ising Models of Social Imitation and Phase Transitions

Perhaps the simplest and historically most important model describing how the aggregation of many individual behaviors can give rise to macroscopic out-of-equilibrium dynamics such as bubbles, with bifurcations in the organization of social systems due to slight changes in the interactions, is the Ising model [16, 80]. In particular, Orléan [85, 86] captured the paradox of combining rational and imitative behavior under the name *mimetic rationality*, by developing

models of mimetic contagion of investors in the stock markets, which are based on irreversible processes of opinion forming. Roehner and Sornette [88], among others, showed that the dynamical updating rules of the Ising model are obtained in a natural way as the optimal strategy of rational traders with limited information who have the possibility to make up for their lack of information *via* information exchange with other agents within their social network. The Ising model is one of the simplest models describing the competition between the ordering force of imitation or contagion and the disordering impact of private information or idiosyncratic noise (see [77] for a technical review).

Starting with a framework suggested by Blume [10, 11], Brock [12], Durlauf [30–33], and Phan *et al.* [87] summarize the formalism starting with different implementation of the agents' decision processes whose aggregation is inspired from statistical mechanics to account for social influence in individual decisions. Lux and Marchesi [72], Brock and Hommes [13], Kaizoji [60], and Kirman and Teyssiere [63] also developed related models in which agents' successful forecasts reinforce the forecasts. Such models have been found to generate swings in opinions, regime changes, and long memory. An essential feature of these models is that agents are wrong for some of the time, but whenever they are in the majority they are essentially right. Thus, they are not systematically irrational [62]. Sornette and Zhou [99] show how Bayesian learning added to the Ising model framework reproduces the stylized facts of financial markets. Harras and Sornette [47] show how overlearning from lucky runs of random news in the presence of social imitation may lead to endogenous bubbles and crashes.

These models allow one to combine the questions on the cause of both bubbles and crashes, as resulting from the collective emergence of herding *via* self-reinforcing imitation and social interactions, which are then susceptible to phase transitions or bifurcations occurring under minor changes in the control parameters. Hence, the difficulty in answering the question of “what causes a bubble and a crash” may, in this context, be attributed to this distinctive attribute of a dynamical out-of-equilibrium system to exhibit bifurcation behavior in its dynamics. This line of thought has been pursued by Sornette and his coauthors, to propose a novel operational diagnostic of bubbles.

V-3 Bubble as Superexponential Price Growth, Diagnostic, and Prediction

Bubbles are often defined as exponentially explosive prices, which are followed by a sudden collapse. As summarized, for instance, by Gurkaynak [46], the problem with this definition is that any exponentially growing price regime—that one would call a bubble—can be also rationalized by a fundamental valuation model. This is related to the problem that the fundamental price is not directly observable, giving no strong anchor to understand observed prices. This was exemplified during the last Internet bubble by fundamental pricing models, which incorporated real options in the fundamental valuation, justifying basically any price. Mauboussin and Hiler [76] were among the most vocal proponents of the proposition, offered close to the peak of the Internet bubble that culminated in 2000, that better business models, the network effect, first-to-scale advantages, and real options effect could account rationally for the high prices of dot-com and other New Economy companies. These interesting views expounded in early 1999 were in synchrony with the bull market of 1999 and preceding years. They participated in the general optimistic view and added to the strength of the herd. Later, after the collapse of the bubble, these explanations seemed less attractive. This did not escape the US Federal Reserve chairman Greenspan [42], who said: “Is it possible that there is something fundamentally new about this current period that would warrant such complacency? Yes, it is possible. Markets may have become more efficient, competition is more global, and information technology has doubtless enhanced the stability of business operations. But, regrettably, history is strewn with visions of such new eras that, in the end, have proven to be a mirage. In short, history counsels caution.” In this vein, the buzzword “new economy” so much used in the late 1990s was also in use in the 1960s during the “tronic boom” also followed by a market crash and during the bubble of the late 1920s before the October 1929 crash. In the latter case, the “new” economy was referring to firms in the utility sector. It is remarkable how traders do not learn the lessons of their predecessors.

A better model derives from the mechanism of positive feedbacks discussed above, which generically gives rise to faster-than-exponential growth of

price (termed as *superexponential*) [95, 96]. An exponential growing price is characterized by a constant expected growth rate. The geometric random walk is the standard stochastic price model embodying this class of behaviors. A superexponential growing price is such that the growth rate grows itself as a result of positive feedbacks of price, momentum, and other characteristics on the growth rate [95]. As a consequence of the acceleration, the mathematical models generalizing the geometric random walk exhibit so-called finite-time singularities. In other words, the resulting processes are not defined for all times: the dynamics has to end after a finite life and to transform into something else. This captures well the transient nature of bubbles, and the fact that the crashes ending the bubbles are often the antechambers to different market regimes.

Such an approach may be thought of, at first sight, to be inadequate or too naive to capture the intrinsic stochastic nature of financial prices, whose null hypothesis is the geometric random walk model [74]. However, it is possible to generalize this simple deterministic model to incorporate nonlinear positive feedback on the stochastic Black–Scholes model, leading to the concept of stochastic finite-time singularities [3, 36, 37, 51, 95]. Much work still needs to be done on this theoretical aspect.

In a series of empirical papers, Sornette and his collaborators have used this concept to empirically test for bubbles and prognosticate their demise often in the form of crashes. Johansen and Sornette [58] provide perhaps the most inclusive series of tests of this approach. First, they identify the most extreme cumulative losses (drawdowns) in a variety of asset classes, markets, and epochs, and show that they belong to a probability density distribution, which is distinct from the distribution of 99% of the smaller drawdowns (the more “normal” market regime). These drawdowns can thus be called *outliers* or *kings* [94]. Second, they show that, for two-thirds of these extreme drawdowns, the market prices followed a superexponential behavior before their occurrences, as characterized by the calibration of the power law with a finite-time singularity.

This provides a systematic approach to diagnose for bubbles *ex ante*, as shown in a series of real-life tests [98, 100, 108–111]. Although this approach has enjoyed a large visibility in the professional financial community around the world (banks, mutual funds, hedge funds, investment houses, etc.), it has not yet

received the attention from the academic financial community that it perhaps deserves given the stakes. This is probably due to several factors, which include the following: (i) the origin of the hypothesis coming from analogies with complex critical systems in physics and the theory of complex systems, which constitutes a well-known obstacle to climb the ivory towers of standard financial economics; (ii) the non-standard (from an econometric viewpoint) formulation of the statistical tests performed until present (in this respect, see the attempts in terms of a Bayesian analysis of log-periodic power law (LPPL) precursors [17] to focus on the time series of returns instead of prices, and of regime-switching model of LPPL [18]), (iii) the nonstandard expression of some of the mathematical models underpinning the hypothesis; and (iv) perhaps an implicit general belief in academia that forecasting financial instabilities is inherently impossible. Lin *et al.* [68] have recently addressed problem (ii) by combining a mean-reverting volatility process and a stochastic conditional return, which reflects nonlinear positive feedbacks and continuous updates of the investors’ beliefs and sentiments. When tested on the S&P500 US index from January 3, 1950 to November 21, 2008, the model correctly identifies the bubbles that ended in October 1987, in October 1997, in August 1998, and the information and communication technologies (ICT) bubble that ended in the first quarter of 2000. Using Bayesian inference, Lin *et al.* [68] find a very strong statistical preference for their model compared with a standard benchmark, in contradiction with Chang and Feigenbaum [17], who used a unit-root model for residuals.

V-4 Bubbles and the Great Financial Crisis of 2007

It is appropriate to end this article with some comments on the relationship between the momentous financial crisis and bubbles. The financial crisis, which started with an initially well-defined epicenter focused on mortgage-backed securities (MBS), has been cascading into a global economic recession, whose increasing severity and uncertain duration are continuing to lead to massive losses and damage for billions of people. At the time of writing (July 2009), the world still suffers from a major financial crisis that has transformed into the worst economic recession since the Great Depression, perhaps on its way

to surpass it. Heavy central bank interventions and government spending programs have been launched worldwide and especially in the United States and Europe, with the hope to unfreeze credit and bolster consumption.

The current financial crisis is a perfect illustration of the major role played by financial bubbles. We refer to the analysis, figures, and references in [97], which articulate a general framework, suggesting that the fundamental cause of the unfolding financial and economic crisis is the accumulation of five bubbles:

1. the “new economy” ICT bubble that started in the mid-1990s and ended with the crash of 2000;
2. the real-estate bubble launched in large part by easy access to a large amount of liquidity as a result of the active monetary policy of the US Federal Reserve lowering the fed rate from 6.5% in 2000 to 1% in 2003 and 2004 in a successful attempt to alleviate the consequence of the 2000 crash;
3. the innovations in financial engineering with the collateralized debt obligations (CDOs) and other derivatives of debts and loan instruments issued by banks and eagerly bought by the market, accompanying and fueling the real-estate bubble;
4. the commodity bubble(s) on food, metals, and energy; and
5. the stock market bubble that peaked in October 2007.

These bubbles, by their interplay and mutual reinforcement, have led to the illusion of a “perpetual money machine”, allowing financial institutions to extract wealth from an unsustainable artificial process. This realization calls to question the soundness of many of the interventions to address the recent liquidity crisis that tend to encourage more consumption.

References

- [1] Abreu, D. & Brunnermeier, M.K. (2003). Bubbles and crashes, *Econometrica* **71**, 173–204.
- [2] Almazan, A., Brown, K.C., Carlson, M. & Chapman, D.A. (2004). Why constrain your mutual fund manager? *Journal of Financial Economics* **73**, 289–321.
- [3] Andersen, J.V. & Sornette, D. (2004). Fearless versus fearful speculative financial bubbles, *Physica A* **337**(3–4), 565–585.
- [4] Badrinath, S.G. & Wahal, S. (2002). Momentum trading by institutions, *Journal of Finance* **57**(6), 2449–2478.
- [5] Barberis, N., Shleifer, A. & Vishny, R. (1998). A model of investor sentiment, *Journal of Financial Economics* **49**, 307–343.
- [6] Battalio, R. & Schultz, P. (2006). Option and the bubble, *Journal of Finance* **61**(5), 2071–2102.
- [7] Bhattacharya, U. & Yu, X. (2008). The causes and consequences of recent financial market bubbles: an introduction, *Review of Financial Studies* **21**(1), 3–10.
- [8] Black, F. (1986). Noise, *The Journal of Finance* **41**(3), 529–543. Papers and Proceedings of the Forty-Fourth Annual Meeting of the American Finance Association, New York, NY, December 28–30, 1985.
- [9] Blanchard, O.J. and Watson, M.W. (1982). Bubbles, rational expectations and speculative markets, in *Crisis in Economic and Financial Structure: Bubbles, Bursts, and Shocks*, P. Wachtel, ed., Lexington Books, Lexington.
- [10] Blume, L.E. (1993). The statistical mechanics of strategic interaction, *Game and Economic Behavior* **5**, 387–424.
- [11] Blume, L.E. (1995). The statistical mechanics of best-response strategy revisions, *Game and Economic Behavior* **11**, 111–145.
- [12] Brock, W.A. (1993). Pathways to randomness in the economy: emergent nonlinearity and chaos in economics and finance, *Estudios Económicos* **8**, 3–55.
- [13] Brock, W.A. & Hommes, C.H. (1999). Rational animal spirits, in *The Theory of Markets*, P.J.J. Herings, G. van derLaan & A.J.J. Talman, eds, North-Holland, Amsterdam, pp. 109–137.
- [14] Brown, S.J., Goetzmann, W., Ibbotson, R.G. & Ross, S.A. (1992). Survivorship bias in performance studies, *Review of Financial Studies* **5**(4), 553–580.
- [15] Brunnermeier, M.K. & Nagel, S. (2004). Hedge funds and the technology bubble, *Journal of Finance* **59**(5), 2013–2040.
- [16] Callen, E. & Shapero, D. (1974). A theory of social imitation, *Physics Today* July, 23–28.
- [17] Chang, G. & Feigenbaum, J. (2006). A Bayesian analysis of log-periodic precursors to financial crashes, *Quantitative Finance* **6**(1), 15–36.
- [18] Chang, G. & Feigenbaum, J. (2007). Detecting log-periodicity in a regime-switching model of stock returns, *Quantitative Finance* **8**, 723–738.
- [19] Chen, J., Hong, H. & Stein, J. (2002). Breadth of ownership and stock returns, *Journal of Financial Economics* **66**, 171–205.
- [20] Cochrane, J.H., 2003. Stocks as money: convenience yield and the tech-stock bubble, in *Asset Price Bubbles*, W.C. Hunter, G.G. Kaufman & M. Pomerleano, eds, MIT Press, Cambridge.
- [21] Cox, A.M.G. & Hobson, D.G. (2005). Local martingales, bubbles and option prices, *Finance and Stochastics* **9**(4), 477–492.

- [22] Daniel, K., Hirshleifer, D. & Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions, *The Journal of Finance* **53**(6), 1839–1885.
- [23] Dass, N., Massa, M. & Patgiri, R. (2008). Mutual funds and bubbles: the surprising role of contracted incentives, *Review of Financial Studies* **21**(1), 51–99.
- [24] De Bondt, W.F.M. & Thaler, R.I.-I. (1985). Does the stock market overreact? *Journal of Finance* **40**, 793–805.
- [25] De Long, B.J. & Shleifer, A. (1991). The stock market bubble of 1929: evidence from closed-end mutual funds, *The Journal of Economic History* **51**(3), 675–700.
- [26] De Long, J.B., Shleifer, A., Summers, L.H. & Waldmann, R.J. (1990a). Positive feedback investment strategies and destabilizing rational speculation, *The Journal of Finance* **45**(2), 379–395.
- [27] De Long, J.B., Shleifer, A., Summers, L.H. & Waldmann, R.J. (1990b). Noise trader risk in financial markets, *The Journal of Political Economy* **98**(4), 703–738.
- [28] Dimitriadi, G.G. (2004). What are “Financial Bubbles”: approaches and definitions, *Electronic journal “INVESTIGATED in RUSSIA”* <http://zhurnal.ape.relarn.ru/articles/2004/245e.pdf>
- [29] Duffie, D., Garleanu, N. & Pedersen, L.H. (2002). Security lending, shorting and pricing, *Journal of Financial Economics* **66**, 307–339.
- [30] Durlauf, S.N. (1991). Multiple equilibria and persistence in aggregate fluctuations, *American Economic Review* **81**, 70–74.
- [31] Durlauf, S.N. (1993). Nonergodic economic growth, *Review of Economic Studies* **60**(203), 349–366.
- [32] Durlauf, S.N., (1997). Statistical mechanics approaches to socioeconomic behavior, in *The Economy as an Evolving Complex System II*, Santa Fe Institute Studies in the Sciences of Complexity, B. Arthur, S. Durlauf & D. Lane, eds, Addison-Wesley, Reading, MA, Vol. XXVII.
- [33] Durlauf, S.N. (1999). How can statistical mechanics contribute to social science? *Proceedings of the National Academy of Sciences of the USA* **96**, 10582–10584.
- [34] Fama, E.F. (1965). The Behavior of Stock-Market Prices, *Journal of Business*, **38**(1), 34–105.
- [35] Fisher, I. (1930). *The Stock Market Crash-and After*, Macmillan, New York.
- [36] Fogedby, H.C. (2003). Damped finite-time-singularity driven by noise, *Physical Review E* **68**, 051105.
- [37] Fogedby, H.C. & Poukaradze, V. (2002). Power laws and stretched exponentials in a noisy finite-time-singularity model, *Physical Review E* **66**, 021103.
- [38] French, K.R. & Poterba, J.M. (1991). Were Japanese stock prices too high? *Journal of Financial Economics* **29**(2), 337–363.
- [39] Friedman, M. & Schwartz, A.J. (1963). *A Monetary History of the United States, 1867-1960*, Princeton University Press, Princeton.
- [40] Galbraith, J.K. (1954/1988). *The Great Crash 1929*, Houghton Mifflin Company, Boston.
- [41] Gompers, P.A. & Metrick, A. (2001). Institutional investors and equity prices, *Quarterly Journal of Economics* **116**, 229–259.
- [42] Greenspan, A. (1997). *Federal Reserve’s Semiannual Monetary Policy Report, before the Committee on Banking, Housing, and Urban Affairs, U.S. Senate*, February 26.
- [43] Griffin, J.M., Harris, J. & Topaloglu, S. (2003). The dynamics of institutional and individual trading, *Journal of Finance* **58**, 2285–2320.
- [44] Grinblatt, M. & Titman, S. (1992). The persistence of mutual fund performance, *Journal of Finance* **47**, 1977–1984.
- [45] Grinblatt, M., Titman, S. & Wermers, R. (1995). Momentum investment strategies, portfolio performance and herding: a study of mutual fund behavior, *The American Economic Review* **85**(5), 1088–1105.
- [46] Gurkaynak, R.S. (2008). Econometric tests of asset price bubbles: taking stock, *Journal of Economic Surveys* **22**(1), 166–186.
- [47] Harras, G. & Sornette, D. (2008). *Endogenous versus Exogenous Origins of Financial Rallies and Crashes in an Agent-based Model with Bayesian Learning and Imitation*, ETH Zurich preprint (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1156348)
- [48] Harrison, M. & Kreps, D. (1978). Speculative investor behavior in a stock market with heterogeneous expectations, *Quarterly Journal of Economics* **92**, 323–336.
- [49] Hong, H., Scheinkman, J. & Xiong, W. (2006). Asset float and speculative bubbles, *Journal of Finance* **59**(3), 1073–1117.
- [50] Hong, H. & Stein, J.C. (2003). Differences of Opinion, short-sales constraints, and market crashes, *The Review of Financial Studies* **16**(2), 487–525.
- [51] Ide, K. & Sornette, D. (2002). Oscillatory finite-time singularities in finance, population and rupture, *Physica A* **307**(1–2), 63–106.
- [52] Jarrow, R. (1980). Heterogeneous expectations, restrictions on short sales, and equilibrium asset prices, *Journal of Finance* **35**, 1105–1113.
- [53] Jarrow, R., Protter, P. & Shimbo, K. (2007). Asset price bubbles in a complete market, in *Advances in Mathematical Finance*, (Festschrift in honor of Dilip Madan’s 60th birthday), M.C. Fu, R.A. Jarrow, J.-Y. Yen & R.J. Elliott, eds, Birkhäuser, pp. 97–122.
- [54] Jarrow, R., Protter, P. & Shimbo, K. (2008). Asset price bubbles in incomplete markets, *Mathematical Finance* to appear.
- [55] Jegadeesh, N. & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* **48**, 65–91.
- [56] Jegadeesh, N. & Titman, S. (2001). Profitability of momentum strategies: An evaluation of alternative explanations, *Journal of Finance* **54**, 699–720.

-
- [57] Johansen, A., Ledoit, O. & Sornette, D. (2000). Crashes as critical points, *International Journal of Theoretical and Applied Finance* **3**(2), 219–255.
- [58] Johansen, A. & Sornette, D. (2004). *Endogenous versus Exogenous Crashes in Financial Markets*, preprint at http://papers.ssrn.com/paper.taf?abstract_id=344980, published as “Shocks, Crashes and Bubbles in Financial Markets,” *Brussels Economic Review* (Cahiers économiques de Bruxelles), 49 (3/4), Special Issue on Nonlinear Analysis (2006) (<http://ideas.repec.org/s/bxrbxrce.html>)
- [59] Johansen, A., Sornette, D. & Ledoit, O. (1999). Predicting financial crashes using discrete scale invariance, *Journal of Risk* **1**(4), 5–32.
- [60] Kaizoji, T. (2000). Speculative bubbles and crashes in stock markets: an interacting agent model of speculative activity, *Physica A* **287**(3–4), 493–506.
- [61] Kindleberger, C.P. (1978). *Manias, Panics and Crashes: A History of Financial Crises*, Basic Books, New York.
- [62] Kirman, A.P. (1997). *Interaction and Markets, G.R.E.Q.A.M. 97a02*, Université Aix-Marseille III.
- [63] Kirman, A.P. & Teyssiere, G. (2002). Micro-economic models for long memory in the volatility of financial time series, in *The Theory of Markets*, P.J.J. Herings, G. VanderLaan & A.J.J. Talman, eds, North-Holland, Amsterdam, pp. 109–137.
- [64] Koski, J.L. & Pontiff, J. (1999). How Are derivatives used? Evidence from the mutual fund industry, *Journal of Finance* **54**(2), 791–816.
- [65] Kyle, A.S. (1985). Continuous auctions and insider trading, *Econometrica* **53**, 1315–1335.
- [66] Lakonishok, J., Shleifer, A. & Vishny, R.W. (1992). The impact of institutional trading on stock prices, *Journal of Financial Economics* **32**, 23–43.
- [67] Lamont, O.A. & Thaler, R.H. (2003). Can the market add and subtract? Mispricing in tech stock carve-outs, *Journal of Political Economy* **111**(2), 227–268. University of Chicago Press.
- [68] Lin, L., Ren, R.E. & Sornette, D. (2009). *A Consistent Model of ‘Explosive’ Financial Bubbles With Mean-Reversing Residuals*, preprint at <http://papers.ssrn.com/abstract=1407574>
- [69] Lintner, J. (1969). The aggregation of investors’ diverse judgments and preferences in purely competitive security markets, *Journal of Financial and Quantitative Analysis* **4**, 347–400.
- [70] Loewenstein, M. & Willard, G.A. (2000a). Rational equilibrium asset-pricing bubbles in continuous trading models, *Journal of Economic Theory* **91**(1), 17–58.
- [71] Loewenstein, M. & Willard, G.A. (2000b). Local martingales, arbitrage and viability: free snacks and cheap thrills, *Economic Theory* **16**, 135–161.
- [72] Lux, T. & Marchesi, M. (1999). Scaling and criticality in a stochastic multi-agent model of a financial market, *Nature* **397**, 498–500.
- [73] Lux, T. & Sornette, D. (2002). On rational bubbles and fat tails, *Journal of Money, Credit and Banking Part I* **34**(3), 589–610.
- [74] Malkiel, B.G. (2007). *A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing*, W.W. Norton & Co.. Revised and Updated edition (December 17, 2007).
- [75] Markopolos, H. (2009). *Testimony of Harry Markopolos, CFA, CFE Chartered Financial Analyst, Certified fraud examiner, before the U.S. House of Representatives, Committee on Financial Services*. Wednesday, February 4, 2009, 9:30am, McCarter & English LLP, Boston.
- [76] Mauboussin, M.J. & Hiler, B. (1999). *Rational Exuberance? Equity Research, Credit Suisse First Boston*, pp. 1–6. January 26, 1999.
- [77] McCoy, B.M. & Wu, T.T. (1973). *The Two-Dimensional Ising Model*, Harvard University, Cambridge, MA.
- [78] Miller, E. (1977). Risk, uncertainty and divergence of opinion, *Journal of Finance* **32**, 1151–1168.
- [79] Miller, M.H. & Modigliani, F. (1961). Dividend policy, growth, and the valuation of shares, *Journal of Business*, **34**(4), 411–433.
- [80] Montroll, E.W. & Badger, W.W. (1974). *Introduction to Quantitative Aspects of Social Phenomena*, Gordon and Breach, New York.
- [81] Nofsinger, J.R. & Sias, R.W. (1999). Herding and feedback trading by institutional and individual investors, *Journal of Finance* **54**, 2263–2295.
- [82] Ofek, E. & Richardson, M. (2002). The valuation and market rationality of internet stock prices, *Oxford Review of Economic Policy* **18**(3), 265–287.
- [83] Ofek, E. & Richardson, M. (2003). DotCom mania: the rise and fall of internet stock prices, *The Journal of Finance* **58**(3), 1113–1137.
- [84] Ofek, E., Richardson, M. & Whitelaw, R.F. (2004). Limited arbitrage and short sale constraints: evidence from the options market, *Journal of Financial Economics* **74**(2), 305–342.
- [85] Orléan, A. (1989). Mimetic contagion and speculative bubbles, *Theory and Decision* **27**, 63–92.
- [86] Orléan, A. (1995). Bayesian interactions and collective dynamics of opinion – herd behavior and mimetic contagion, *Journal of Economic Behavior and Organization* **28**, 257–274.
- [87] Phan, D., Gordon, M.B. & Nadal, J.-P. (2004). Social interactions in economic theory: an insight from statistical mechanics, in *Cognitive Economics – An Interdisciplinary Approach*, P. Bourguin & J.-P. Nadal, eds, Springer, Berlin.
- [88] Roehner, B.M. & Sornette, D. (2000). Thermometers of speculative frenzy, *European Physical Journal B* **16**, 729–739.
- [89] Scheinkman, J. & Xiong, W. (2003). Overconfidence and speculative bubbles, *Journal of Political Economy* **111**, 1183–1219.

- [90] Schultz, P. (2008). Downward-sloping demand curves, the supply of shares, and the collapse of internet stock prices, *Journal of Finance* **63**, 351–378.
- [91] Shiller, R. (2000). *Irrational Exuberance*, Princeton University Press, Princeton, NJ.
- [92] Shleifer, A. & Vishny, R. (1997). Limits of arbitrage, *Journal of Finance* **52**, 35–55.
- [93] Sornette, D. (2003). *Why Stock Markets Crash (Critical Events in Complex Financial Systems)*, Princeton University Press, Princeton NJ.
- [94] Sornette, D. (2009). Dragon-Kings, Black Swans and the Prediction of Crises, in press in the *International Journal of Terraspace Science and Engineering* (<http://ssrn.com/abstract=1470006>).
- [95] Sornette, D. & Andersen, J.V. (2002). A nonlinear super-exponential rational model of speculative financial bubbles, *International Journal of Modern Physics C* **13**(2), 171–188.
- [96] Sornette, D., Takayasu, H. & Zhou, W.-X. (2003). Finite-time singularity signature of hyperinflation, *Physica A: Statistical Mechanics and Its Applications* **325**, 492–506.
- [97] Sornette, D. & Woodard, R. (2009). Financial bubbles, real estate bubbles, derivative bubbles, and the financial and economic crisis, to appear in the Proceedings of APFA7 (Applications of Physics in Financial Analysis), in *New Approaches to the Analysis of Large-Scale Business and Economic Data*, M. Takayasu, T. Watanabe & H. Takayasu, eds., Springer (2010) (e-print at <http://arxiv.org/abs/0905.0220>)
- [98] Sornette, D., Woodard, R. & Zhou, W.-X. (2008). *The 2006–2008 Oil Bubble and Beyond*, ETH Zurich preprint (<http://arXiv.org/abs/0806.1170>)
- [99] Sornette, D. & Zhou, W.-X. (2006a). Importance of positive feedbacks and over-confidence in a self-fulfilling ising model of financial markets, *Physica A: Statistical Mechanics and its Applications* **370**(2), 704–726.
- [100] Sornette, D. & Zhou, W.-X. (2006b). Predictability of large future changes in major financial indices, *International Journal of Forecasting* **22**, 153–168.
- [101] Soros, G. (1987). *The Alchemy of Finance: Reading the Mind of the Market*, Wiley, Chichester.
- [102] Stanley, H.E. (1987). *Introduction to Phase Transitions and Critical Phenomena*, Oxford University Press, USA.
- [103] Thom, R. (1989). *Structural Stability and Morphogenesis: An Outline of a General Theory of Models*, Addison-Wesley, Reading, MA.
- [104] Tirole, J. (1982). On the possibility of speculation under rational expectations, *Econometrica* **50**, 1163–1182.
- [105] Wermers, R. (1999). Mutual fund herding and the impact on stock prices, *Journal of Finance* **54**(2), 581–622.
- [106] West, K.D. (1988). Bubbles, fads and stock price volatility tests: a partial evaluation, *Journal of Finance* **43**(3), 639–656.
- [107] White, E.N. (2006). *Bubbles and Busts: The 1990s in the Mirror of the 1920s* NBER Working Paper No. 12138.
- [108] Zhou, W.-X. & Sornette, D. (2003). 2000–2003 real estate bubble in the UK but not in the USA, *Physica A* **329**, 249–263.
- [109] Zhou, W.-X. & Sornette, D. (2006). Is there a real-estate bubble in the US? *Physica A* **361**, 297–308.
- [110] Zhou, W.-X. & Sornette, D. (2007). *A Case Study of Speculative Financial Bubbles in the South African Stock Market 2003–2006*, ETH Zurich preprint (<http://arxiv.org/abs/physics/0701171>)
- [111] Zhou, W.-X. & Sornette, D. (2008). Analysis of the real estate market in Las Vegas: bubble, seasonal patterns, and prediction of the CSW indexes, *Physica A* **387**, 243–260.

Further Reading

- Abreu, D & Brunnermeier, M.K. (2002). Synchronization risk and delayed arbitrage, *Journal of Financial Economics* **66**, 341–360.
- Farmer, J.D. (2002). Market force, ecology and evolution, *Industrial and Corporate Change* **11**(5), 895–953.
- Narasimhan, J. & Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency, *The Journal of Finance* **48**(1), 65–91.
- Narasimhan, J. & Titman, S. (2001). Profitability of momentum strategies: an evaluation of alternative explanations, *The Journal of Finance* **56**(2), 699–720.
- Shleifer, A & Summers, L.H. (1990). The noise trader approach to finance, *The Journal of Economic Perspectives* **4**(2), 19–33.

TAISEI KAIZOJI & DIDIER SORNETTE

Ross, Stephen

The central focus of the work of Ross (1944–) has been to tease out the consequences of the assumption that all riskless arbitrage opportunities have already been exploited and none remain. The empirical relevance of the no arbitrage assumption is especially high in the area of financial markets for two simple reasons: there are many actors actively searching for arbitrage opportunities, and the exploitation of such opportunities is relatively costless. For finance, therefore, the principle of no arbitrage is not merely a convenient assumption that makes it possible to derive clean theoretical results but even more an idealization of observable empirical reality, and a characterization of the deep and simple structure underlying multifarious surface phenomena. For one whose habits of mind were initially shaped by the methods of natural science, specifically physics as taught by Richard Feynman (B.S. California Institute of Technology, 1965), finance seemed to be an area of economics where a truly scientific approach was possible.

It was exposure to the Black–Scholes option pricing theory, when Ross was starting his career as an assistant professor at the University of Pennsylvania, that first sparked his interest in the line of research that would occupy him for the rest of his life. If the apparently simple and eminently plausible assumption of no arbitrage could crack the problem of option pricing, perhaps it could crack other problems in finance as well. In short order, Ross produced what he later called the *fundamental theorem of asset pricing* [7, p. 101], which linked the absence of arbitrage with the existence of a positive linear pricing rule [12, 15] (see **Fundamental Theorem of Asset Pricing**).

Perhaps the most important practical implication of this theorem is that it is possible to price assets that are not yet traded simply by reference to the price of assets that are already traded, and to do so without the need to invoke any particular theory of asset pricing. This opened the possibility of creating new assets, such as options, that would in practical terms “complete” markets, and so help move the economy closer to the ideal efficient frontier characterized by Kenneth Arrow (see **Arrow, Kenneth**) as a complete set of markets for state-contingent securities [11]. Here, in the abstract, is

arguably the vision that underlies the entire field of financial engineering.

The general existence of a linear pricing rule has further implications that Ross would later group together in what he called the *pricing rule representation theorem* [7, p. 104]. Most important for practical purposes is the existence of positive risk-neutral probabilities and an associated riskless rate of interest, a feature first noted in [4, 5]. It is this general feature that makes it possible to model option prices by treating the underlying stock price as a binomial random variable in discrete time, as first introduced by Cox *et al.* [6] in an approach that is now ubiquitous in industry practice. It is this same general feature that makes it possible to characterize asset prices generally as following a martingale under the equivalent martingale measure [9], a characterization that is also now routine in financial engineering practice.

What is most remarkable about these consequences of the no arbitrage point of view is how little economics has to do with it. Ross, a trained economist (Harvard, PhD, 1969), might well have built a rather different career, perhaps in the area of agency theory where he made one of the early seminal contributions [10], but once he found finance he never looked back. (His subsequent involvement in agency theory largely focused on financial intermediation in a world with no arbitrage, as in [14, 18].)

When Ross was starting his career, economists had already begun making inroads into finance, and one of the consequences was the Sharpe–Lintner capital asset pricing model (CAPM) (see **Modern Portfolio Theory**). Ross [16] reinterpreted the CAPM as a possible consequence of no arbitrage and then proposed his own arbitrage pricing theory [13] as a more general consequence that would be true whenever asset prices were generated by a linear factor model such as

$$R_i = E_i + \sum \beta_{ij} f_j + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

where E_i is the expected return on asset i , f_i is an exogenous systematic factor, and ε_i is the random noise.

In such a world, it follows from no arbitrage that the expected return on asset i , in excess of the risk-free rate of return r , is equal to a linear combination

of the factor loadings β_{ij} :

$$E_i - r = \Sigma \lambda_j \beta_{ij} \quad (2)$$

This is the APT generalization of the CAPM security market line that connects the mean–variance of the market (r_M , σ_M) to that of the risk-free asset (r , 0).

It also follows that the optimal portfolio choice for any agent can be characterized as a weighted sum of n mutual funds, one for each factor. This is the APT generalization of the CAPM two-fund separation theorem, and unlike CAPM it does not depend on any special assumptions about either utility functions or the stochastic processes driving asset returns. In a certain sense, it does not depend on economics.

We can understand the work of Cox *et al.* [1–3] as an attempt to connect the insights of no arbitrage back to economic “fundamentals”. “In work on contingent claims analysis, such as option pricing, it is common, and to a first approximation reasonable, to insist only on a partial equilibrium between the prices of the primary and derivative assets. For something as fundamental as the rate of interest, however, a general equilibrium model is to be preferred” [1, p. 773]. They produce a general equilibrium model driven by a k -dimensional vector of state variables, but are forced to specialize the model considerably in order to achieve definite results for the dynamics of interest rates and the term structure. Here, more than anywhere else in Ross’s wide-ranging work, we see the tension between the methodologies of economics and finance. It is this experience, one supposes, that lies behind his subsequent defense of the “isolated and eccentric tradition” that is unique to finance [17, p. 34]. The tradition to which he refers is the practice of approaching financial questions from the perspective of no arbitrage, without the apparatus of utility and production functions and without demand and supply.

Not content with having established the core principles and fundamental results of the no arbitrage approach to finance, Ross devoted his subsequent career to making sure that the significance and wide applicability of these results was appreciated by both academicians and practitioners. Toward that end, his own voluminous writings have been multiplied by the work of the many students whom he trained at the University of Pennsylvania, then Yale, and then MIT [8].

References

- [1] Cox, J.C., Ingersoll Jr, J.E. & Ross, S. (1981). A re-examination of traditional hypotheses about the term structure of interest rates, *Journal of Finance* **36**(4), 769–799.
- [2] Cox, J.C., Ingersoll Jr, J.E. & Ross, S. (1985a). An intertemporal general equilibrium model of asset prices, *Econometrica* **53**(2), 363–384.
- [3] Cox, J.C., Ingersoll Jr, J.E. & Ross, S. (1985b). A theory of the term structure of interest rates, *Econometrica* **53**(2), 385–407.
- [4] Cox, J.C. & Ross, S.A. (1976a). The valuation of options for alternative stochastic processes, *Journal of Financial Economics* **3**, 145–166.
- [5] Cox, J.C. & Ross, S.A. (1976b). A survey of some new results in financial option pricing theory, *Journal of Finance* **31**(2), 383–402.
- [6] Cox, J.C., Ross, S.A. & Rubinstein, M. (1979). Option pricing: a simplified approach, *Journal of Financial Economics* **7**, 229–263.
- [7] Dybvig, P.H. & Ross, S.A. (1987). Arbitrage, in *New Palgrave, A Dictionary of Economics*, J. Eatwell, M. Milgate & P. Newman, eds, Macmillan, London, pp. 100–106.
- [8] Grinblatt, M. (ed) (2008). *Stephen A. Ross, Mentor: Influence Through Generations*, McGraw Hill, New York.
- [9] Harrison, J.M. & Kreps, D. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**(3), 381–408.
- [10] Ross, S.A. (1973). The economic theory of agency: the principal’s problem, *American Economic Review* **63**(2), 134–139.
- [11] Ross, S.A. (1976a). Options and efficiency, *Quarterly Journal of Economics* **90**(1), 75–89.
- [12] Ross, S.A. (1976b). Return, risk, and arbitrage, in *Risk and Return in Finance*, I. Friend & J. Bicksler, eds, Ballinger, Cambridge, pp. 189–217.
- [13] Ross, S.A. (1976c). The Arbitrage theory of capital asset pricing, *Journal of Economic Theory* **13**, 341–360.
- [14] Ross, S.A. (1977). The determination of financial structure: the incentive-signalling approach, *Bell Journal of Economics* **8**(1), 23–40.
- [15] Ross, S.A. (1978b). A simple approach to the valuation of risky streams, *Journal of Business* **51**(3), 453–475.
- [16] Ross, S.A. (1982). On the general validity of the man-variance approach in large markets, in *Financial Economics: Essays in Honor of Paul Cootner*, W. Sharpe & P. Cootner, eds, Prentice-Hall.
- [17] Ross, S.A. (1987). The interrelations of finance and economics: theoretical perspectives, *American Economic Review* **77**(2), 29–34.

- [18] Ross, S.A. (2004). Markets for agents: fund management, in *The Legacy of Fischer Black*, B.N. Lehman, ed, Oxford University Press.

Further Reading

- Ross, S.A. (1974). Portfolio Turnpike theorems for constant policies, *Journal of Financial Economics* **1**, 171–198.
 Ross, S.A. (1978a). Mutual fund separation in financial theory: the separating distributions, *Journal of Economic Theory* **17**(2), 254–286.

Related Articles

Arbitrage: Historical Perspectives; Arbitrage Pricing Theory; Black, Fischer; Equivalent Martingale Measures; Martingale Representation Theorem; Option Pricing Theory: Historical Perspectives; Risk-neutral Pricing.

PERRY MEHRLING

Fisher, Irving

The American economist Irving Fisher (born 1867, died 1947) advanced the use of formal mathematical and statistical techniques in economics and finance, both in his own pioneering research in monetary and capital theory and in his roles as a mentor to a handful of talented doctoral students and as founding president of the Econometric Society. As an undergraduate and a graduate student at Yale University, Fisher studied with the physicist J. Willard Gibbs and the economist and sociologist William Graham Sumner. Fisher's 1891 doctoral dissertation in economics and mathematics, *Mathematical Investigations in the Theory of Value and Prices* (reprinted in [12], Vol. 1), was the first North American use of general equilibrium analysis—indeed, an independent rediscovery of general equilibrium, because Fisher did not read the works of Léon Walras and F.Y. Edgeworth until his thesis was nearly completed. To accompany this thesis, Fisher constructed a hydraulic mechanism to simulate the determination of equilibrium prices and quantities, a remarkable achievement in the days before electronic computers (see Brainard and Scarf in [5] and Schwalbe in [14]). Initially appointed to teach mathematics at Yale, Fisher soon switched to political economy, teaching at Yale until he retired in 1935. Stricken with tuberculosis in 1898, Fisher was on leave for three years, and did not resume a full teaching load until 1903. This ordeal turned Fisher into a relentless crusader for healthier living and economic reforms, dedicated to improving the world and confident of overcoming adversity and daunting obstacles [1, 5, 14]. As a scientific economist and as a reformer, Fisher was a brilliant and multifaceted innovator, but he never managed to pull his ideas together in a grand synthesis.

In *The Nature of Capital and Income*, Fisher [7] popularized the concept of net present value, viewing capital as the present discounted value of an expected income stream. Controversially, Fisher excluded saving from his definition of income, and advocated a spending tax instead of a tax on income as usually defined. Since saving is the acquisition of assets whose market value is the net present value of the expected taxable income from owning the assets, a tax on income (as usually defined) would involve double taxation and would introduce a distortion

favoring consumption at the expense of saving, a view now increasingly held by economists. Fisher [7] also discussed the pricing and allocation of risk in financial markets, using a “coefficient of caution” to represent subjective attitudes to risk tolerance [2, 3, 18]. In *The Rate of Interest*, Fisher [8] drew on the earlier work of John Rae and Eugen von Böhm-Bawerk to examine how intertemporal allocation and the real interest rate depend on impatience (time preference) and opportunity to invest (expected rate of return over cost). He illustrated this analysis with the celebrated “Fisher diagram” showing optimal smoothing of consumption over two periods. According to the “Fisher separation theorem,” the time pattern of consumption is independent of the time pattern of income (assuming perfect credit markets), because the net present value of expected lifetime income is the relevant budget constraint for consumption and saving decisions, rather than income in a particular period. Fisher's analysis of consumption smoothing across time periods provided the basis for later permanent-income and life-cycle models of consumption, and was extended by others to consumption smoothing across possible states of the world. John Maynard Keynes later identified his concept of the marginal efficiency of capital with Fisher's rate of return over costs.

Fisher's *Appreciation and Interest* [6] presented the “Fisher equation,” decomposing nominal interest into real interest and expected inflation, formalizing and expounding an idea that had been briefly noted by, among others, John Stuart Mill and Alfred Marshall. With i as the nominal interest rate, j as the real interest rate, and a as the expected rate of appreciation of the purchasing power of money ([6] appeared at the end of two decades of falling prices),

$$(1 + j) = (1 + a)(1 + i) \quad (1)$$

in Fisher's notation. This analysis of the relationship between interest rates expressed in two different standards (money and goods, gold and silver, dollars and pounds sterling) led Fisher [6] to uncovered interest parity (the difference between nominal interest rates in two currencies is the expected rate of change of the exchange rate) and to a theory of the term structure of interest rates as reflecting expectations about future changes in the purchasing power of money. In later work (see [12], Vol. 9), Fisher correlated nominal interest with a distributed lag of

past price level changes, deriving expected inflation adaptively from past inflation. Distributed lags were introduced into economics by Fisher, who was also among the first economists to use correlation analysis. Long after Fisher's death, his pioneering 1926 article [10], correlating unemployment with a distributed lag of inflation, was reprinted in 1973, under the title "I Discovered the Phillips Curve."

In *The Purchasing Power of Money*, Fisher [13] upheld the quantity theory of money, arguing that changes in the quantity of money affect real output and real interest during adjustment periods of up to 10 years, but affect only nominal variables in the long run. He extended the quantity theory's equation of exchange to include bank deposits:

$$MV + M'V' = PT \quad (2)$$

where M is currency, M' bank deposits, V and V' the velocities of circulation of currency and bank deposits, respectively, P the price level, and T an index of the volume of transactions. Fisher attributed economic fluctuations to the slow adjustment of nominal interest to monetary shocks, resulting from what he termed "*the money illusion*" in the title of a 1928 book (in [12], Vol. 8). The economy would be stable if, instead of pegging the dollar price of gold, monetary policy followed Fisher's "compensated dollar" plan of regularly varying the price of gold to target an index number of prices. Inflation targeting is a modern version of Fisher's proposed price level target (without attempting a variable peg of the price of gold, which would have made Fisher's plan vulnerable to speculative attacks). Failing to persuade governments to stabilize the purchasing power of money, Fisher attempted to neutralize the effects of price level changes by advocating the creation of indexed financial instruments, persuading Rand Kardex (later Remington Rand) to issue the first indexed bond (see [12], Vol. 8). Fisher tried to educate the public against money illusion, publishing a weekly index of wholesale prices calculated by an index number institute operating out of his house in New Haven, Connecticut. Indexed bonds, the compensated dollar, statistical verification of the quantity theory, and eradication of money illusion all called for a measure of the price level. In *The Making of Index Numbers*, Fisher [9] argued that a simple formula, the geometric mean of the Laspeyres (base-year weighted) index and the Paasche (current-year weighted) index, was the best index number for that and all other purposes, as

it came closer than any other formula to satisfying seven tests for such desirable properties as determinateness, proportionality, and independence of the units of measurement. Later research demonstrated that no formula can satisfy more than six of the seven tests, although, which one should be dropped remains an open question. Three quarters of a century later, the "Fisher ideal index" began to be adopted by governments.

Beyond his work, Fisher encouraged quantitative research by others, notably Yale dissertations by J. Pease Norton [16] and Chester A. Phillips [17], and through his role as founding president of the Econometric Society. Norton's *Statistical Studies of the New York Money Market* is now recognized as a landmark in time-series analysis, while Phillips's *Bank Credit* (together with later work by Fisher's former student James Harvey Rogers) analyzed the creation and absorption of bank deposits by the banking system [4]. Arguing that fluctuations in the purchasing power of money make money and bonds risky assets, contrary to the widespread "money illusion," Fisher and his students advocated common stocks as a long-term investment, with the return on stocks more than compensating for their risk, once risk is calculated in real rather than in nominal terms.

Fisher was swept up in the "New Economy" rhetoric of the 1920s stock boom. He promoted several ventures, of which by far the most successful was his "Index Visible," a precursor of the Rolodex. Fisher sold Index Visible to Rand Kardex for shares and stock options, which he exercised with borrowed money. In mid-1929, Fisher's net worth was 10 million dollars. Had he died then, he would have been remembered like Keynes as a financial success as well as a brilliant theorist; however, a few years later, Fisher's debts exceeded his assets by a million dollars—a loss of 11 million dollars, which, as John Kenneth Galbraith remarked, was "a substantial sum of money, even for a professor of economics" [1, 3]. Worst of all for his public and professional reputation, Fisher memorably asserted in October 1929, on the eve of the Wall Street crash, that stock prices appeared to have reached a permanently high plateau. McGrattan and Prescott [15] hold that Fisher was right to deny that stocks were overvalued in 1929 given the prices/earnings multiples of the time. Whether or not Fisher could reasonably be faulted for not

predicting the subsequent errors of public policy that converted the downturn into the Great Depression, and even though many others were just as mistaken about the future course of stock prices, Fisher's mistaken prediction was particularly pithy, quotable, and memorable, and his reputation suffered as severely as his personal finances. Fisher's 1933 article on "The Debt-Deflation Theory of Great Depressions" [11], linking the fragility of the financial system to the nonneutrality of inside nominal debt whose real value grew as the price level fell, was much later taken up by such economists as Hyman Minsky, James Tobin, Ben Bernanke, and Mervyn King [5, 14], but in the 1930s Fisher had lost his audience. Fisher's 1929 debacle (together with his enthusiastic embrace of causes ranging from a new world map projection, the unhealthiness of smoking, and the usefulness of mathematics in economics, through the League of Nations, universal health insurance, and a low-protein diet to, more regrettably, prohibition and eugenics) long tarnished his public and professional reputation, but he has increasingly come to be recognized as a great figure in the development of theoretical and quantitative economics, including financial economics.

References

- [1] Allen, R.L. (1993). *Irving Fisher: A Biography*, Blackwell, Cambridge, MA.
- [2] Crockett, J.H. Jr. (1980). Irving Fisher on the financial economics of uncertainty, *History of Political Economy* **12**, 65–82.
- [3] Dimand, R. (2007). Irving Fisher and financial economics: the equity premium puzzle, the predictability of stock prices, and intertemporal allocation under risk, *Journal of the History of Economic Thought* **29**, 153–166.
- [4] Dimand, R. (2007). Irving Fisher and his students as financial economists, in *Pioneers of Financial Economics*, G. Poitras, ed., Edward Elgar, Cheltenham, UK, Vol. 2, pp. 45–59.
- [5] Dimand, R. & Geanakoplos, J. (eds) (2005). *Celebrating Irving Fisher*, Blackwell, Malden, MA.
- [6] Fisher, I. (1896). *Appreciation and Interest*, Macmillan for American Economic Association, New York. (reprinted in Fisher [12], **1**).
- [7] Fisher, I. (1906). *The Nature of Capital and Income*, Macmillan, New York. (reprinted in Fisher [12], **2**).
- [8] Fisher, I. (1907). *The Rate of Interest*, Macmillan, New York. (reprinted in Fisher [12], **3**).
- [9] Fisher, I. (1922). *The Making of Index Numbers*, Houghton Mifflin, Boston. (reprinted in Fisher [12], **7**).
- [10] Fisher, I. (1926). A statistical relation between unemployment price changes, *International Labour Review* **13**, 785–792. reprinted as Lost and found: (1973) I discovered the Phillips curve – Irving Fisher, *Journal of Political Economy* **81**, 496–502.
- [11] Fisher, I. (1933). The debt-deflation theory of great depressions, *Econometrica* **1**, 337–357. (reprinted in Fisher [12], **10**).
- [12] Fisher, I. (1997). *The Works of Irving Fisher*, W.J. Barber, ed, Pickering & Chatto, London.
- [13] Fisher, I. & Brown, H.G. (1911). *The Purchasing Power of Money*, Macmillan, New York. (reprinted in Fisher [12], **4**).
- [14] Loef, H. & Monissen, H. (eds) *The Economics of Irving Fisher*, Edward Elgar, Cheltenham, UK.
- [15] McGrattan, E. & Prescott, E. (2004). The 1929 stock market: Irving Fisher was right, *International Economic Review* **45**, 91–1009.
- [16] Norton, J.P. (1902). *Statistical Studies in the New York Money Market*, Macmillan, New York.
- [17] Phillips, C. (1920). *Bank Credit*, Macmillan, New York.
- [18] Stabile, D. & Putnam, B. (2002). Irving Fisher and statistical approaches to risk, *Review of Financial Economics* **11**, 191–203.

ROBERT W. DIMAND

Modigliani, Franco

An Italian-born economist who fled the fascist regime of Benito Mussolini at the outbreak of WWII, Modigliani pursued the study of economics at the New School of Social Research (renamed New School University) in New York where he received his doctorate in 1944. He taught at several universities but, from 1962 on, he stayed at the Massachusetts Institute of Technology. His famous dissertation on the Keynesian system served as a springboard for many of his lifetime contributions, which include stabilization policies, the FRB–MIT–Penn–SSRC Model (MPS), the Modigliani–Miller (M&M) theorem (**Modigliani–Miller Theorem**) and the life cycle hypothesis (LCH). Modigliani was awarded the Nobel Memorial Prize in economics in 1985 for research in the latter two areas.

Modigliani contributed to making the disciplines of financial economics and macroeconomics operational, and thus more quantitative from a neoclassical perspective. The influence of his teachers, particularly J. Marschak and A. Wald, is seen in his quantitative MPS model based on Keynesian economic thought and his M&M hypothesis in financial economics. The macroeconomic framework that Modigliani built emphasized the savings, consumption, investment, and liquidity components of the Keynesian model. He explained the anomalous fluctuations of the savings (S) to income (Y) ratio during the 1940s and 1950s. He explained the S/Y ratio by the relative position in the income distribution of individuals, and by secular and cyclical changes in income ([3], Vol. 2). The secular changes represent differences in real income per capita above the highest level reached in any preceding year, signifying his contribution to the relative income hypothesis in consumption theory. The cyclical changes represent variation in money income measured by an index, $(Y_t - Y_t^0)/Y_t$, where Y_t is real income per capita in current time, and Y_t^0 is the past peak level of such income. He estimated that the secular and the cyclical affects on income were approximately 0.1% and 0.125%, respectively. These coefficients translate to an S/Y ratio of about 11.7%. Klein and Ozmucur [1] revisited Modigliani's S/Y specification with a much larger sample size and were able to reaffirm the robustness of the model.

In 1954, Modigliani laid the groundwork for the now-famous life cycle hypothesis (LCH) ([5], Vol. 6, pp. 3–45). The LCH bracketed broader macroeconomic problems such as why S/Y is larger in rich countries than in poor countries; why S is greater for farm families than urban families; why lower status urban families save less than other urban families; why when a higher future income is expected, more of current income will be consumed now; why in countries with rising income that is expected to continue to increase, S/Y will be smaller; and why property income that mostly accrues to the rich is largely saved, whereas wages that are mostly earned by the poor are largely spent. To answer these questions, the LCH model maintains the relative income concept of the early S/Y model. The income concept is, however, more encompassing in being high or low relative to the individual's lifetime or permanent income, marking Modigliani's contribution to the permanent income hypothesis in consumption theory. The LCH captures how individuals save when they are young, spend when they are old, and make bequests to their children. In that scenario, consumption, C is uniform over time, T , or $C(T) = (N/L)\bar{Y}$, where L is the number of years the representative individual lives; $N < L$ is the number of years the individual earns labor income, and $\bar{Y}$ is average income. Average income is represented by a flat line, $Y(T)$ up to N , which falls to zero after N , when the individual retires. Since income is earned for N periods, lifetime income is NY , and savings is defined as the excess of $Y(T)$ over $C(T)$.

The empirical estimate of the LCH included a wealth-effect variable on consumption. Saving during an individual's early working life is one way in which wealth accumulates. Such an accumulation of wealth reaches a peak during the person's working age when income is highest. Individuals also inherit wealth. If the initial stock of wealth is A_0 , then, at a certain age, τ , a person's consumption can be expressed as $(L - \tau)C = A + (N - \tau)Y$. Thus, we have a model of consumption explained by income and wealth or assets that can be confronted with data. An early estimate of the coefficient of this LCH model yielded $C = 0.76Y + 0.073A$ (Modigliani, *ibid.*, 70). The result reconciled an early controversy that the short-run propensity to consume from income was between 70% and 80%, and the long-run propensity was approximately 100%. The

reconciliation occurs because the short-run marginal propensity to consume (MPC) is 0.766, and assuming assets, A , is approximately five times income, while labor income is approximately 80% of income, then a long-run MPC is approximately $0.98 = 0.8(.76Y) + 5(.073Y)$.

Modigliani's largest quantitative effort was the MPS model. Working with the board of governors of the Federal Reserve Banks (FRB) and the Social Science Research Council (SSRC), Modigliani built the MIT–Penn–SSRC (MPS) econometric model in the 1960s. The 1968 version, which had 171 endogenous and 119 exogenous variables, predicted poorly in the 1970s and 1980s. In 1996, the FRB/US model replaced the MPS by incorporating rational and vector autoregression types of expectations with a view to improve forecasts. The financial sector was the dominant module in the MPS model. The net worth of consumers took the form of the real value of money and debt. The demand for money depended on the nominal interest rate and the current value of output. Unborrowed reserves influenced the short-term money rate of interest and the nominal money supply, and through the term structure effect, the short-term rate affected the long-term rate and hence savings, which is essential for the expansion of output and employment. Out of this process came the following two fitted demand and supply equations that characterized the financial sector:

$$M_d = -0.0021iY - 0.0043r_sY + 0.542Y + 0.0046NP + 0.833Md_{t-1} \quad (1)$$

$$FR = (0.001 - 0.00204S_2 - 0.00237S_3 - 0.00223S_4)\overline{D}_{t-1} + 0.00122i\overline{D}_{t-1} + 0.00144d\overline{D}_{t-1} + 0.646(1 - \delta)\Delta RU - 0.502\delta\Delta CL + 0.394RD + 0.705FR_{t-1} \quad (2)$$

where M_d is demand for deposits held by the public, Y is gross national product (GNP), r_s is the savings deposit rate, i is the available return on short-term assets, P is expected profits, FR is free reserves, S_i are seasonal adjustments, $\overline{D}$ is the expected value of the stock of member banks deposits, RU is unborrowed reserves, CL is commercial loans, RL

is a reserve release term, and δ is a constant. The equations indicate that the cause and effect between unborrowed reserves to GNP works through lags, causing delay responses to policy measures.

Another of Modigliani's noteworthy contributions to quantitative analysis is the Modigliani and Miller (M&M) theorem [6], which has created a revolution in corporate finance equivalent to the revolution in portfolio theory by H. Markowitz and W. Sharpe. The M&M hypothesis stands on two major propositions, namely that "... market value of any firm is independent of its capital structure and is given by capitalizing its expected return at the rate ρ_k appropriate to its class," and that "*the average cost of capital to any firm is completely independent of the capital structure and is equal to the capitalization rate of a pure equity stream of its class*" (Italics original) ([4], Vol. 3, 10–11). The M&M model can be demonstrated for a firm with no growth, no new net investment, and no taxes. The firm belongs to a risk group in which its shares can be substituted for one another.

The value of the firm can be written as $V_j \equiv S_j + D_j = \overline{X}_j / \rho_j$, where $\overline{X}_j$ measures expected return on assets, ρ_j measures interest rate for a given risk class, D_j is market value of bonds, and S_j is the market value of stocks. For instance, if the earnings before interest and taxes (EBIT) are \$5000 and if the low-risk interest is 10%, then the net operating income is \$50 000.

The proposition of the M&M hypothesis is often expressed as an invariance principle based on the idea that the value of a firm is independent of how it is financed. The proof of this invariance is based on arbitrage. As stated by Modigliani, "... an investor can buy and sell stocks and bonds in such a way as to exchange one income stream for another ... the value of the overpriced shares will fall and that of the under priced shares will rise, thereby tending to eliminate the discrepancy between the market values of the firms" (ibid., p. 11). For example, an investor can get a 6% return either by holding the stocks of an unlevered firm ($0.06X_1$), or holding the stocks and debts of a levered firm, that is, $[0.06(X_2 - rD_2)$ of stocks + $0.06rD_2$ of debts], where the subscripts refer to firms, X is stock, D is debt, and r is return.

The M&M hypothesis was a springboard for many new works in finance. A first extension of the model by the authors reflected the effect of corporate tax effects. Further analysis incorporating the effects of

personal and corporate income taxes does not change the value of the firm because both personal and corporate tax rates tend to cancel out. Researchers dealt with questions that arise when the concept of risk class used in the computation of a firm's value is replaced with perfect market assumptions, and when mean–variance models are used instead of arbitrage. The value of the firm was also found to be independent of dividend policy. By changing the discount rate for the purpose of calculating a firm's present value, it was found that bankruptcy can have an effect on the value of a firm. Macroeconomic variables such as the inflation rate can result in the underestimation of the value of a firm's equity. The M&M theorem has been extended into many areas of modern research. It supports the popular Black–Scholes capital structure model. It has been used to validate the effect of the Tax Reform Act of 1986 on values of the firm. Modern capital asset pricing model (CAPM) scholars such as Sharpe (**Sharpe, William F.**), J. Lintner, and J. Treynor [2] were influenced by the M&M result in the construction of their financial models and ratios.

On a personal level, Modigliani was an outstandingly enthusiastic, passionate, relentless, and focus-driven teacher and exceptional researcher whose arena was both economic theory and the real empirical world.

References

- [1] Klein, L.R. & Ozmur, S. (2005). *The Wealth Effect: A Contemporary Update*, paper presented at the New School University.
- [2] Mehrling, P. (2005). *Fisher Black and the Revolutionary Idea of Finance*, John Wiley & Sons, Inc, Hoboken.
- [3] Modigliani, F. (1980). Fluctuations in the saving-income ratio: a problem in economic forecasting, in *The Collected Papers of Franco Modigliani, The Life Cycle Hypothesis of Savings*, A. Abel, & S. Johnson, eds, The MIT Press, Cambridge, MA, Vol. 2.
- [4] Modigliani, F. (1980). The cost of capital, corporate finance and the theory of investment, in *The Collected Papers of Franco Modigliani, The Theory of Finance and Other Essays*, A. Abel, ed., The MIT Press, Cambridge, MA, Vol.3.
- [5] Modigliani, F. (2005). *Collected Papers of Franco Modigliani*, F. Modigliani, ed., The MIT Press, Cambridge, MA, Vol. 6.
- [6] Modigliani, F. & Miller, M. (1958). The cost of capital, corporation finance and the theory of investment, *American Economic Review* 48(3), 261–297.

Further Reading

- Modigliani, F. (2003). The Keynesian Gospel according to Modigliani, *The American Economist* 47(1), 3–24.
- Ramrattan, L. & Szenberg, M. (2004). Franco Modigliani 1918–2003, in memoriam, *The American Economist* 43(1), 3–8.
- Szenberg, M. & Ramrattan, L. (2008). *Franco Modigliani, A Mind That Never Rests with a Foreword by Robert M. Solow*, Palgrave Macmillan, Houndmills, Basingstoke and New York.

Related Articles

Modigliani–Miller Theorem.

MICHAEL SZENBERG & LALL RAMRATTAN

Arrow, Kenneth

Most financial decisions are made under conditions of uncertainty. Yet a formal analysis of markets under uncertainty emerged only recently, in the 1950s. The matter is complex as it involves explaining how individuals make decisions when facing uncertain situations, the behavior of market instruments such as insurance, securities, and their prices, the welfare properties of the distribution of goods and services under uncertainty, and how risks are shared among the traders. It is not even obvious how to formulate market clearing under conditions of uncertainty. A popular view in the middle of the last century was that markets would only clear on the average and asymptotically in large economies.^a This approach was a reflection of how insurance markets work, and followed a notion of actuarially fair trading.

A different formulation was proposed in the early 1950s by Arrow and Debreu [10, 12, 30]. They introduced an economic theory of markets in which the treatment of uncertainty follows basic principles of physics. The contribution of Arrow and Debreu is as fundamental as it is surprising. For Arrow and Debreu, markets under uncertainty are formally identical to markets without uncertainty. In their approach, uncertainty all but disappears.^b

It may seem curious to explain trade with uncertainty as though uncertainty did not matter. The disappearing act of the issue at stake is an unusual way to think about financial risk, and how we trade when facing such risks. But the insight is valuable. Arrow and Debreu produced a rigorous, consistent, general theory of markets under uncertainty that inherits the most important properties of markets without uncertainty. In doing so, they forced us to clarify what is intrinsically different about uncertainty.

This article summarizes the theory of markets under uncertainty that Arrow and Debreu created, including critical issues that arise from it, and also its legacy. It focuses on the way Arrow introduced securities: how he defined them and the limits of his theory. It mentions the theory of insurance that Arrow pioneered together with Malinvaud and others [6], as well as the theory of risk bearing that Arrow developed on the basis of expected utility [7], following the axioms of Von Neumann

and Morgenstern [41], Harsanyi and Milnor [33], De Groot [31], and Villegas [40]. The legacy of Arrow's work is very extensive and some of it surprising. This article describes his legacy along three lines: (i) individual and idiosyncratic risks, (ii) rare risks and catastrophic events, and (iii) endogenous uncertainty.

Biographical Background

Kenneth Joseph Arrow is American economist and joint winner of the Nobel Memorial Prize in Economics with John Hicks in 1972. Arrow taught at Stanford University and Harvard University. He is one of the founders of modern (post World War II) economic theory, and one of the most important economists of the twentieth century. For a full biographical note, the reader is referred to [18]. Born in 1921 in New York City to Harry and Lilian Arrow, Kenneth was raised in the city. He graduated from Townsend Harris High School and earned a bachelor's degree from the City College of New York studying under Alfred Tarski. After graduating in 1940, he went to Columbia University and after a hiatus caused by World War II, when he served with the Weather Division of the Army Air Forces, he returned to Columbia University to study under the great statistician Harold Hotelling at Columbia University. He received a master's degree in 1941 studying under A. Wald, who was the supervisor of his master's thesis on stochastic processes. From 1946 to 1949 he spent his time partly as a graduate student at Columbia and partly as a research associate at the Cowles Commission for Research in Economics at the University of Chicago; it was in Chicago that he met his wife Selma Schweitzer. During that time, he also held the position of Assistant Professor of Economics at the University of Chicago. Initially interested in following a career as an actuary, in 1951 he earned his doctorate in economics from Columbia University working under the supervision of Harold Hotelling and Albert Hart. His published work on risk started in 1951 [3]. In developing his own approach to risk, Arrow grapples with the ideas of Shackle [39], Knight [35], and Keynes [34] among others, seeking and not always finding a rigorous mathematical foundation. His best-known works on financial markets date back to 1953 [3]. These works provide a solid foundation based on the

role of securities in the allocation of risks [4, 5, 7, 9, 10]. His approach can be described as a state contingent security approach to the allocations of risks in an economy, and is largely an extension of the same approach he followed in his work on general equilibrium theory with Gerard Debreu, for which he was awarded the Nobel Prize in 1972 [8]. Nevertheless, his work connects also with social issues of risk allocation and with the French literature of the time, especially [1, 2].

Markets under Uncertainty

The Arrow–Debreu theory conceptualizes uncertainty with a number of possible states of the world $s = 1, 2, \dots$ that may occur. Commodities can be in one of several states, and are traded separately in each of the states of nature. In this theory, one does not trade a good, but a “contingent good”, namely, a good in each state of the world: apples when it rains and apples when it shines [10, 12, 30]. This way the theory of markets with N goods and S states of nature is formally identical to the theory of markets without uncertainty but with $N \times S$ commodities. Traders trade “state contingent commodities”. This simple formulation allows one to apply the results of the theory of markets without uncertainty, to markets with uncertainty. One recovers most of the important results such as (i) the existence of a market equilibrium and (ii) the “invisible hand theorem” that establishes that market solutions are always Pareto efficient. The approach is elegant, simple, and general.

Along with its elegance and simplicity, the formulation of this theory can be unexpectedly demanding. It requires that we all agree on all the possible states of the world that describe “collective uncertainty”, and that we trade accordingly. This turns out to be more demanding than it seems: for example, one may need to have a separate market for apples when it rains than when it does not, and separate market prices for each case. The assumption requires $N \times S$ markets to guarantee market efficiency, a requirement that in some cases militates against the applicability of the theory. In a later article, Arrow simplified the demands of the theory and reduced the number of markets needed for efficiency by defining “securities”, which are different payments of money exchanged among the traders in different states of

nature [4, 5]. This new approach no longer requires trading “contingent” commodities but rather trading a combination of commodities and securities. Arrow proves that by trading commodities and securities, one can achieve the same results as trading state contingent commodities [4, 5]. Rather than needing $N \times S$ markets, one needs a fewer number of markets, namely, N markets for commodities and $S - 1$ markets for securities. This approach was a great improvement and led to the study of securities in a rigorous and productive manner, an area in which his work has left a large legacy. The mathematical requirement to reach Pareto efficiency was simplified gradually to require that the securities traded should provide for each trader a set of choices with the same dimension as the original state contingent commodity approach. When this condition is not satisfied, the markets are called “incomplete”. This led to a large literature on incomplete markets, for example, [26, 32], in which Pareto efficiency is not assured, and government intervention may be required, an area that exceeds the scope of this article.

Individual Risk and Insurance

The Arrow–Debreu theory is not equally well suited for all types of risks. In some cases, it could require an unrealistically large number of markets to reach efficient allocations. A clear example of this phenomenon arises for those risks that pertain to one individual at a time, called *individual risks*, which are not readily interpreted as states of the world on which we all agree and are willing to trade. Individuals’ accidents, illnesses, deaths, and defaults, are frequent and important risks that fall under this category. Arrow [6] and Malinvaud [37] showed how individual uncertainty can be reformulated or reinterpreted as collective uncertainty. Malinvaud formalized the creation of states of collective risks from individual risks, by lists that describe all individuals in the economy, each in one state of individual risk. The theory of markets can be reinterpreted accordingly [14, 37, 38], yet remains somewhat awkward. The process of trading under individual risk using the Arrow–Debreu theory requires an unrealistically large number of markets. For example with N individuals, each in one of two individual states G (good) and B (bad), the number of (collective) states that are required to apply the Arrow–Debreu

theory is $S = 2^N$. The number of markets required is as above, either $S \times N$ or $N + S - 1$. But with $N = 300$ million people, as in the US economy, applying the Arrow–Debreu approach would require $N \times S = N \times 2^{300 \text{ million}}$ markets to achieve Pareto efficiency, more markets than the total amount of particles in the known universe [25]. For this reason, individual uncertainty is best treated with another formulation of uncertainty involving individual states of uncertainty and insurance rather than securities, in which market clearing is defined on the average and may never actually occur. In this new approach, instead of requiring $N + S - 1$ markets, one requires only N commodity markets and, with two states of individual risk, just one security: an insurance contract suffices to obtain asymptotic efficiency [37, 38]. This is a satisfactory theory of individual risk and insurance, but it leads only to asymptotic market clearing and Pareto efficiency. More recently, the theory was improved and it was shown that one can obtain exact market-clearing solutions and Pareto-efficient allocations based on N commodity markets with the introduction of a limited number of financial instruments called *mutual insurance* [14]. It is shown in [14] that if there are N households (consisting of H types), each facing the possibility of being in S individual states together with T collective states, then ensuring Pareto optimality requires only $H(S - 1)T$ independent mutual insurance policies plus T pure Arrow securities.

Choice and Risk Bearing

Choice under uncertainty explains how individuals rank risky outcomes. In describing how we rank choices under uncertainty, one follows principles that were established to describe the way nature ranks what is most likely to occur, a topic that was widely explored and is at the foundation of statistics [31, 40]. To explain how individuals choose under conditions of uncertainty, Arrow used behavioral axioms that were introduced by Von Neumann and Morgenstern [41] for the theory of games^c and axioms defined by De Groot [31] and Villegas [40] for the foundation of statistics. The main result obtained in the middle of the twentieth century was that under rather simple behavioral assumptions, individuals behave as though they were optimizing

an “expected utility function”. This means that they behave as though they have (i) a utility u for commodities, which is independent of the state of nature, and (ii) subjective probabilities about how likely are the various states of nature. Using the classic axioms one constructs a ranking of choice under uncertainty obtaining a well-known expected utility approach. Specifically, traders choose over “lotteries” that achieve different outcomes in different states of nature. When states of nature and outcomes are represented by real numbers in R , a lottery is a function $f : R \rightarrow R^N$, a utility is a function $u : R^N \rightarrow R$, and a subjective probability is $p : R \rightarrow [0, 1]$ with $\int_R p(s) = 1$. Von Neumann, Arrow, and Hershstein and Milnor, all obtained the same classic “representation theorem” that identifies choice under uncertainty by the ranking of lotteries according to a real-valued function W , where W has the now familiar “expected utility” form:

$$W(f) = \int_{s \in R} p(s) \cdot u(f(s)) \, ds \quad (1)$$

The utility function u is typically bounded to avoid paradoxical behavior. The expected utility approach just described has been generally used since the mid-twentieth century. Despite its elegance and appeal, from the very beginning, expected utility has been unable to explain a host of experimental evidence that was reported in the work of Allais [2] and others. There has been a persistent conflict between theory and observed behavior, but no axiomatic foundation to replace Von Neumann’s foundational approach. The reason for this discrepancy has been identified more recently, and it is attributed to the fact that expected utility is dominated by frequent events and neglects rare events—even those that are potentially catastrophic, such as widespread default in today’s economies. That expected utility neglects rare events was shown in [17, 19, 23]. In [23], the problem was traced back to Arrow’s axiom of monotone continuity [7], which Arrow attributed to Villegas [40], and to the corresponding continuity axioms of Hershstein and Milnor, and De Groot [31], who defined a related continuity condition denoted “ SP_4 ”. Because of this property, on which Arrow’s work is based, the expected utility approach has been characterized as the “dictatorship” of frequent events, since it is dominated by the consideration of “normal” and frequent events [19]. To correct this bias, and to represent more accurately how we choose

under uncertainty, and to arrive at a more realistic meaning of rationality, a new axiom was added in [17, 19, 21], requiring equal treatment for frequent and for rare events. The new axiom was subsequently proven to be the logic negation of Arrow's monotone continuity that was shown to neglect small probability events [23].

The new axioms led to a "representation theorem" according to which the ranking of lotteries is a modified expected utility formula

$$W(f) = \int_{s \in R} p(s) \cdot u(f(s)) ds + \phi(f) \quad (2)$$

where ϕ is a continuous linear function on lotteries defined by a finite additive measure, rather than a countably additive measure [17, 19]. This measure assigns most weight to rare events. The new formulation has both types of measures, so the new characterization of choice under uncertainty incorporates both (i) frequent and (ii) rare events in a balanced manner, conforming more closely to the experimental evidence on how humans choose under uncertainty [15]. The new specification gives well-deserved importance to catastrophic risks, and a special role to fear in decision making [23], leading to a more realistic theory of choice under uncertainty and foundations of statistics, [15, 23, 24]. The legacy of Kenneth Arrow's work is surprising but strong: the new theory of choice under uncertainty coincides with the old when there are no catastrophic risks so that, in reality, the latter is an extension of the former to incorporate rare events. Some of the most interesting applications are to environmental risks such as global warming [25]. Here Kenneth Arrow's work was prescient: Arrow was a contributor to the early literature on environmental risks and irreversibilities [11], along with option values.

Endogenous Uncertainty and Widespread Default

Some of the risks we face are not created by nature. They are our own creation, such as global warming or the financial crisis of 2008 and 2009 anticipated in [27]. In physics, the realization that the observer matters, that the observer is a participant and creates uncertainty, is called *Heisenberger's uncertainty principle*. The equivalent in economics is an uncertainty principle that describes how we create risks

through our economic behavior. This realization led to the new concept of "markets with endogenous uncertainty", created in 1991, and embodied in early articles [16, 27, 28] that established some of the basic principles and welfare theorems in markets with endogenous uncertainty. This, and other later articles ([20, 25, 27, 36]), established basic principles of existence and the properties of the general equilibrium of markets with endogenous uncertainty. It is possible to extend the Arrow–Debreu theory of markets to encompass markets with endogenous uncertainty and also to prove the existence of market equilibrium under these conditions [20]. But in the new formulation, Heisenberg's uncertainty principle rears its quizzical face. It is shown that it is no longer possible to fully hedge the risks that we create ourselves [16], no matter how many financial instruments we create. The equivalent of Russel's paradox in mathematical logic appears also in this context due to the self-referential aspects of endogenous uncertainty [16, 20]. Pareto efficiency of equilibrium can no longer be ensured. Some of the worst economic risks we face are endogenously determined—for example, those that led to the 2008–2009 global financial crisis [27]. In [27] it was shown that the creation of financial instruments to hedge individual risks—such as credit default insurance that is often a subject of discussion in today's financial turmoil—by themselves induce collective risks of widespread default. The widespread default that we experience today was anticipated in [27], in 1991, and in 2006, when it was attributed to endogenous uncertainty created by financial innovation as well as to our choices of regulation or deregulation of financial instruments. Examples are the extent of reserves that are required for investment banking operations, and the creation of mortgage-backed securities that are behind many of the default risks faced today [29]. Financial innovation of this nature, and the attendant regulation of new financial instruments, causes welfare gains for individuals—but at the same time creates new risks for society that bears the collective risks that ensue, as observed in 2008 and 2009. In this context, an extension of the Arrow–Debreu theory of markets can no longer treat markets with endogenous uncertainty as equivalent to markets with standard commodities. The symmetry of markets with and without uncertainty is now broken. We face a brave new world of financial innovation and the

endogenous uncertainty that we create ourselves. Creation and hedging of risks are closely linked, and endogenous uncertainty has acquired a critical role in market performance and economic welfare, an issue that Kenneth Arrow has more recently tackled himself through joint work with Frank Hahn [13].

Acknowledgments

Many thanks are due to Professors Rama Cont and Perry Mehrling of Columbia University and Barnard College, respectively, for their comments and excellent suggestions.

End Notes

^aSee [37, 38]; later on Werner Hildenbrand followed this approach.

^bThey achieved the same for their treatment of economic dynamics. Trading over time and under conditions of uncertainty characterizes financial markets.

^cAnd similar axioms used by Hershman and Milnor [33].

^dSpecifically to avoid the so-called St. Petersburg paradox, see [7].

References

- [1] Allais, M. (ed) (1953). *Fondements et Applications de la Theorie du Risque en Econometrie*, CNRS, Paris.
- [2] Allais, M. (1987). The general theory of random choices in relation to the invariant cardinality and the specific probability function, in *Risk Decision and Rationality*, B.R. Munier, ed., Reidel, Dordrech The Netherlands, pp. 233–289.
- [3] Arrow, K. (1951). Alternative approaches to the theory of choice in risk – taking situations, *Econometrica* **19**(4), 404–438.
- [4] Arrow, K. (1953). Le Role des Valeurs Boursiers pour la Repartition la Meilleure des Risques, *Econometrie* **11**, 41–47. Paris CNRS, translated in English in *RES* 1964 (below).
- [5] Arrow, K. (1953). The role of securities in the optimal allocation of risk bearing, *Proceedings of the Colloque sur les Fondements et Applications de la Theorie du Risque en Econometrie*. CNRS, Paris. English Transation published in *The Review of Economic Studies* Vol. 31, No. 2, April 1964, p. 91–96.
- [6] Arrow, K. (1953). Uncertainty and the welfare economics of medical care, *American Economic Review* **53**, 941–973.
- [7] Arrow, K. (1970). *Essays on the Theory of Risk Bearing*, North Holland, Amsterdam.
- [8] Arrow, K. (1972). *General Economic Equilibrium: Purpose Analytical Techniques Collective Choice*, Les Prix Nobel en 1972, Stockholm Nobel Foundation pp. 253–272.
- [9] Arrow, K. (1983). *Collected Papers of Kenneth Arrow*, Belknap Press of Harvard University Press.
- [10] Arrow, K.J. & Debreu, G. (1954). Existence of an equilibrium for a competitive economy, *Econometrica* **22**, 265–290.
- [11] Arrow, K.J. & Fischer, A. (1974). Environmental preservation, uncertainty and irreversibilities, *Quarterly Journal of Economics* **88**(2), 312–319.
- [12] Arrow, K. & Hahn, F. (1971). *General Competitive Analysis*, Holden Day, San Francisco.
- [13] Arrow, K. & Hahn, F. (1999). Notes on sequence economies, transaction costs and uncertainty, *Journal of Economic Theory* **86**, 203–218.
- [14] Cass, D., Chichilnisky, G. & Wu, H.M. (1996). Individual risk and mutual insurance, *Econometrica* **64**, 333–341.
- [15] Chanel, O. & Chichilnisky, G. (2009). The influence of fear in decisions: experimental evidence, *Journal of Risk and Uncertainty* **39**(3).
- [16] Chichilnisky, G. (1991, 1996). Markets with endogenous uncertainty: theory and policy, *Columbia University Working paper 1991 and Theory and Decision* **41**(2), 99–131.
- [17] Chichilnisky, G. (1996). Updating Von Neumann Morgenstern axioms for choice under uncertainty with catastrophic risks. *Proceedings of Conference on Catastrophic Risks*, Fields Institute for Mathematical Sciences, Toronto, Canada.
- [18] Chichilnisky, G. (ed) (1999). *Markets Information and Uncertainty: Essays in Honor of Kenneth Arrow*, Cambridge University Press.
- [19] Chichilnisky, G. (2000). An axiomatic treatment of choice under uncertainty with catastrophic risks, *Resource and Energy Economics* **22**, 221–231.
- [20] Chichilnisky, G. (1999/2008). Existence and optimality of general equilibrium with endogenous uncertainty, in *Markets Information and Uncertainty: Essays in Honor of Kenneth Arrow*, 2nd Edition, G. Chichilnisky, ed., Cambridge University Press, Chapter 5.
- [21] Chichilnisky, G. (2009). The foundations of statistics with Black Swans, *Mathematical Social Sciences*, DOI:10.1016/j.mathsocsci.2009.09.007.
- [22] Chichilnisky, G. (2009). The limits of econometrics: non parametric estimation in Hilbert spaces, *Econometric Theory* **25**, 1–17.
- [23] Chichilnisky, G. (2009). “The Topology of Fear” invited presentation at NBER conference in honor of Gerard Debreu, UC Berkeley, December 2006, *Journal of Mathematical Economics* **45**(11–12), December 2009. Available online 30 June 2009, ISSN 0304–4068, DOI: 10.1016/j.jmateco.2009.06.006.
- [24] Chichilnisky, G. (2009a). Subjective Probability with Black Swans, *Journal of Probability and Statistics* (in press, 2010).

-
- [25] Chichilnisky, G. & Heal, G. (1993). Global environmental risks, *Journal of Economic Perspectives, Special Issue on the Environment* Fall, 65–86.
- [26] Chichilnisky, G. & Heal, G. (1996). On the existence and the structure pseudo-equilibrium manifold, *Journal of Mathematical Economics* **26**, 171–186.
- [27] Chichilnisky, G. & Wu, H.M. (1991, 2006). General equilibrium with endogenous uncertainty and default, Working Paper Stanford University, 1991, *Journal of Mathematical Economics* **42**, 499–524.
- [28] Chichilnisky, G., Heal, G. & Dutta, J. (1991). *Endogenous Uncertainty and Derivative Securities in a General Equilibrium Model*, Working Paper Columbia University.
- [29] Chichilnisky, G., Heal, G. & Tsomocos, D. (1995). Option values and endogenous uncertainty with asset backed securities, *Economic Letters* **48**(3–4), 379–388.
- [30] Debreu, G. (1959). *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*, John Wiley & Sons, New York.
- [31] De Groot, M.H. (1970, 2004). *Optimal Statistical Decisions*, John Wiley & Sons, Hoboken New Jersey.
- [32] Geanakoplos, J. (1990). An introduction to general equilibrium with incomplete asset markets, *Journal of Mathematical Economics* **19**, 1–38.
- [33] Hershstein, N. & Milnor, J. (1953). An axiomatic approach to measurable utility, *Econometrica* **21**, 219–297.
- [34] Keynes, J.M. (1921). *A Treatise in Probability*, MacMillan and Co., London.
- [35] Knight, F. (1921). *Risk Uncertainty and Profit*, Houghton Mifflin and Co., New York.
- [36] Kurz, M. & Wu, H.M. (1996). Endogenous uncertainty in a general equilibrium model with price - contingent contracts, *Economic Theory* **6**, 461–488.
- [37] Malinvaud, E. (1972). The allocation of individual risks in large markets, *Journal of Economic Theory* **4**, 312–328.
- [38] Malinvaud, E. (1973). Markets for an exchange economy with individual; Risks, *Econometrica* **41**, 383–410.
- [39] Shackle, G.L. (1949). *Expectations in Economics*, Cambridge University Press, Cambridge, UK.
- [40] Villegas, C. (1964). On quantitative probability σ -algebras, *Annals of Mathematical Statistics* **35**, 1789–1800.
- [41] Von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ.

Related Articles

Arrow–Debreu Prices; Risk Aversion; Risk Premia; Utility Theory: Historical Perspectives.

GRACIELA CHICHILNISKY

Efficient Markets Theory: Historical Perspectives

Without any doubt, it can be said that efficient market hypothesis (EMH) was crucial in the emergence of financial economics as a proper subfield of economics. But this was not its original goal: EMH was initially created to give a theoretical explanation of the random character of stock market prices.

The historical roots of EMH can be traced back to the nineteenth century and the early twentieth century in the work of Regnault and Bachelier, but their work was isolated and not embedded in a scientific community interested in finance. More immediate roots of the EMH lie in the empirical work of Cowles, Working, and Kendall from 1933 to 1959, which laid the foundation for the key works published in the period from 1959 (Roberts) to 1976 (Fama's reply to LeRoy). More than any other single contributor, it was Fama [7] in his 1965 dissertation, building on the work of Roberts, Cowles, and Cootner, who formulated the EMH, suggesting that stock prices reflect all available information, and that, consequently, the actual value of a security is equal to its price. In addition, because new information arrives randomly, stock prices fluctuate randomly.

The idea that stock prices fluctuate randomly was not new: in 1863, a French broker, Jules Regnault [20], had already suggested it. Regnault was the first author to put forward this hypothesis, to validate it empirically, and to give it a theoretical interpretation. In 1900, Louis Bachelier [1], a French mathematician, used Regnault's hypothesis and framework to develop the first mathematical model of Brownian motion, and tested the model by using it to price futures and options. In retrospect, we can recognize that Bachelier's doctoral dissertation constitutes the first work in mathematical finance. Unfortunately for him, however, financial economics did not then exist as a scientific field, and there was no organized scientific community interested in his research. Consequently, both Regnault and Bachelier were ignored by economists until the 1960s.

Although these early authors did suggest modeling stock prices as a stochastic process, they did not formulate the EMH as it is known today. EMH was genuinely born in linking three elements that originally existed independently of each other: (i) the

mathematical model of a stochastic process (random walk, Brownian motion, or martingale); (ii) the concept of economic equilibrium; and (iii) the statistical results about the unpredictability of stock market prices. EMH's creation took place only between 1959 and 1976, when a large number of economists became familiar with these three features. Between the time of Bachelier and the development of EMH, there were no theoretical preoccupations *per se* about the random character of stock prices, and research was only empirical.

Empirical Research between 1933 and 1959

Between 1933 and the end of the 1950s, only three authors dealt with the random character of stock market prices: Cowles [3, 4], Working [24, 25], and Kendall [13]. They compared stock price fluctuations with random simulations and found similarities. One point must be underlined: these works were strictly statistical, and no theory explained these empirical results.

The situation changed at the end of the 1950s and during the 1960s because of three particular events. First, the Koopmans–Vining controversy at the end of 1940s led to a decline of descriptive approaches and to the increased use of modeling based on theoretical foundations. Second, modern probability theory, and consequently also the theory of stochastic processes, became usable for nonmathematicians. Significantly, economists were attracted to the new formalisms by some features that were already familiar consequences of economic equilibrium. Most important, the zero expected profit when prices follow a Brownian motion reminded economists of the zero marginal profit in the equilibrium of a perfectly competitive market. Third, research on the stock market became more and more popular among scholars: groups of researchers and seminars in financial economics became organized; scientific journals such as the *Journal of Financial and Quantitative Analysis* were created and a community of scholars was born. This context raised awareness about the need for theoretical investigations, and these investigations, in turn, allowed the creation of the EMH.

Theoretical Investigations during the 1960s

Financial economists did not speak immediately of EMH; they talked about “random walk theory”. Following his empirical results, Working [26] was the first author to suggest a theoretical explanation; he established an explicit link between the unpredictable arrival of information and the random character of stock market price changes. However, this paper made no link with economic equilibrium and, probably for this reason, it was not widely diffused. Instead, it was Roberts [21], a professor at the University of Chicago, who first suggested a link between economic concepts and the random walk model by using the “arbitrage proof” argument that had been popularized by Modigliani and Miller [19]. Then, Cowles [5] made an important step by identifying a link between financial econometric results and economic equilibrium. Finally, two years later, Cootner [2] linked the random walk model, information, and economic equilibrium, and exposed the idea of EMH, although he did not use that expression.

Cootner [2] had the essential idea of EMH, but he did not make the crucial empirical link because he considered that real-world stock price variations were not purely random. This point of view was defended by economists from MIT (such as Samuelson) and Stanford University (such as Working). By contrast, economists from the University of Chicago claimed that real stock markets were perfect, and so were more inclined to characterize them as efficient. Thus, it was a scholar from the University of Chicago, Eugene Fama, who formulated the EMH.

In his 1965 PhD thesis, Fama gave the first theoretical account of EMH. In that account, the key assumption is the existence of “sophisticated traders” who, due to their skills, make a better estimate of intrinsic valuation than do other agents by using all available information. Provided that such traders have predominant access to financial resources, their activity of buying underpriced assets and selling overpriced assets will tend to make prices equal the intrinsic values about which they have a shared assessment and also to eliminate any expectation of profit from trading. Linking these consequences with the random walk model, Fama added that because information arrives randomly, stock prices have to fluctuate randomly. Fama thus offered the first clear

link between empirical results about stock price variations, the random walk model, and economic equilibrium. EMH was born.

Evolution of Fama’s Definition during the 1970s

Five years after his PhD dissertation, Fama [8] offered a mathematical demonstration of the EMH. He simplified his first definition by making the implicit assumption of a representative agent. He also used another stochastic process: the martingale model, which had been introduced to model the random character of stock market prices by Samuelson [22] and Mandelbrot [17]. The martingale model is less restrictive than the random walk model: the martingale model requires only independence of the conditional expectation of price changes, whereas the random walk model requires also independence involving the higher conditional moments (i.e., variance, skewness, and kurtosis) of the probability distribution of price changes. For Fama’s [8] purposes, the most important attraction of the martingale formalism was its explicit reference to a set of information, Φ_t ,

$$E(P_{t+1}|\Phi_t) - P_t = 0 \quad (1)$$

As such, the martingale model could be used to test the implication of EMH that, if all available information is used, the expected profit is null. This idea led to the definition of an efficient market that is generally used nowadays: “a market in which prices always ‘fully reflect’ available information is called ‘efficient’” [8].

However, in 1976, LeRoy [15] showed that Fama’s demonstration is tautological and that his theory is not testable. Fama answered by modifying his definition and he also admitted that any test of the EMH is a test of both market efficiency and the model of equilibrium used by investors. In addition, it is striking to note that the test suggested by Fama [9] (i.e., markets are efficient if stock prices are equal to the prediction provided by the model of equilibrium used) does not imply any clear causality between the random character of stock market prices and the EMH; it is mostly a plausible correlation valid only for some cases.

The Proliferation of Definitions since the 1970s

Fama's modification of his definition proved to be a fateful admission. In retrospect, it is clear that the theoretical content of EMH comprised its suggestion of a link between some mathematical model, some empirical results, and some concept of economic equilibrium. The precise linkage proposed by Fama was, however, only one of many possible linkages, as subsequent literature would demonstrate. Just so, LeRoy [14] and Lucas [16] provided theoretical proofs that efficient markets and the martingale hypothesis are two distinct ideas: martingale is neither necessary nor sufficient for an efficient market. In a similar way, Samuelson [23], who gave a mathematical proof that prices may be permanently equal to the intrinsic value and fluctuate randomly, explained that it cannot be excluded that some agents make profits, contrary to the original definition of EMH. De Meyer and Saley [6] show that stock market prices follow a martingale even if all available information is not contained in stock market prices.

This proliferation at the level of theory has been matched by proliferation at the level of empirical testing, as the definition of EMH has changed depending on the emphasis placed by each author on one particular feature. For instance, Fama *et al.* [10] defined an efficient market as "a market that adjusts rapidly to new information"; Jensen [12] considered that "a market is efficient with respect to information set θ_t if it is impossible to make economic profit by trading on the basis of information set θ_t "; according to Malkiel [18] "the market is said to be efficient with respect to some information set [...] if security prices would be unaffected by revealing that information to all participants. Moreover, efficiency with respect to an information set [...] implies that it is impossible to make economic profits by trading on the basis of [that information set]".

The situation is similar regarding the tests: the type of test used depends on the definition used by the authors and on the data used (for instance, most of the tests are done with low frequency or daily data, while statistical arbitrage opportunities are discernible and exploitable at high frequency using algorithmic trading). Moreover, some authors have used the weakness of the definitions to criticize the very relevance of efficient markets. For instance, Grossman and Stiglitz [11] argued that because information

is costly, prices cannot perfectly reflect all available information. Consequently, they considered that perfectly information-efficient markets are impossible.

The history of EMH shows that the definition of this theory is plural, and the initial project of EMH (the creation of a link between a mathematical model, the concept of economic equilibrium, and statistical results about the unpredictability of stock market prices) has not been fully achieved. Moreover, this theory is not empirically refutable (since a test of the random character of stock prices does not imply a test on efficiency). Nevertheless, financial economists have considered EMH as one of the pillars of financial economics because it played a key role in the creation and history of financial economics by linking financial results with standard economics. This link is the main contribution of EMH.

References

- [1] Bachelier, L. (1900). Théorie de la spéculation reproduced in *Annales de l'Ecole Normale Supérieure*, 3ème série 17, in *Random Character of Stock Market Prices* (English Translation: P.H. Cootner, ed, (1964)), M.I.T. Press, Cambridge, MA, pp. 21–86.
- [2] Cootner, P.H. (1962). Stock prices: random vs. systematic changes, *Industrial Management Review* 3(2), 24–45.
- [3] Cowles, A. (1933). Can stock market forecasters forecast? *Econometrica* 1(3), 309–324.
- [4] Cowles, A. (1944). Stock market forecasting, *Econometrica* 12(3/4), 206–214.
- [5] Cowles, A. (1960). A revision of previous conclusions regarding stock price behavior, *Econometrica* 28(4), 909–915.
- [6] De Meyer, B. & Saley, H.M. (2003). On the strategic origin of Brownian motion in finance, *International Journal of Game Theory* 31, 285–319.
- [7] Fama, E.F. (1965). The behavior of stock-market prices, *Journal of Business* 38(1), 34–105.
- [8] Fama, E.F. (1970). Efficient capital markets: a review of theory and empirical work, *Journal of Finance* 25(2), 383–417.
- [9] Fama, E.F. (1976). Efficient capital markets: reply, *Journal of Finance* 31(1), 143–145.
- [10] Fama, E.F., Fisher, L., Jensen, M.C. & Roll, R. (1969). The adjustment of stock prices to new information, *International Economic Review* 10(1), 1–21.
- [11] Grossman, S.J. & Stiglitz, J.E. (1980). The impossibility of informationally efficient markets, *American Economic Review* 70(3), 393–407.
- [12] Jensen, M.C. (1978). Some anomalous evidence regarding market efficiency, *Journal of Financial Economics* 6, 95–101.

- [13] Kendall, M.G. (1953). The analysis of economic time-series. Part I: prices, *Journal of the Royal Statistical Society* **116**, 11–25.
- [14] LeRoy, S.F. (1973). Risk-aversion and the martingale property of stock prices, *International Economic Review* **14**(2), 436–446.
- [15] LeRoy, S.F. (1976). Efficient capital markets: comment, *Journal of Finance* **31**(1), 139–141.
- [16] Lucas, R.E. (1978). Asset prices in an exchange economy, *Econometrica* **46**(6), 1429–1445.
- [17] Mandelbrot, B. (1966). Forecasts of future prices, unbiased markets, and “Martingale” models, *Journal of Business* **39**(1), 242–255.
- [18] Malkiel, B.G. (1992). Efficient Market Hypothesis, in *The New Palgrave Dictionary of Money and Finance*, P. Newman, M. Milgate & J. Eatwell, eds, Macmillan, London.
- [19] Modigliani, F. & Miller, M.H. (1958). The cost of capital, corporation finance and the theory of investment, *The American Economic Review* **48**(3), 261–297.
- [20] Regnault, J. (1863). *Calcul des Chances et Philosophie de la Bourse*, Mallet-Bachelier and Castel, Paris.
- [21] Roberts, H.V. (1959). Stock-market “Patterns” and financial analysis: methodological suggestions, *Journal of Finance* **14**(1), 1–10.
- [22] Samuelson, P.A. (1965). Proof that properly anticipated prices fluctuate randomly, *Industrial Management Review* **6**(2), 41–49.
- [23] Samuelson, P.A. (1973). Proof that properly discounted present value of assets vibrate randomly, *Bell Journal of Economics* **4**(2), 369–374.
- [24] Working, H. (1934). A random-difference series for use in the analysis of time series, *Journal of the American Statistical Association* **29**, 11–24.
- [25] Working, H. (1949). The investigation of economic expectations, *The American Economic Review* **39**(3), 150–166.
- [26] Working, H. (1956). New ideas and methods for price research, *Journal of Farm Economics* **38**, 1427–1436.

Further Reading

- Jovanovic, F. (2008). The construction of the canonical history of financial economics, *History of Political Economy* **40**(3), 213–242.
- Jovanovic, F. & Le Gall, P. (2001). Does God practice a random walk? The “financial physics” of a 19th century forerunner, Jules Regnault, *European Journal of the History of Economic Thought* **8**(3), 323–362.
- Jovanovic, F. & Poitras, G. (eds) (2007). *Pioneers of Financial Economics: Twentieth Century Contributions*, Edward Elgar, Cheltenham, Vol. 2.
- Poitras, G. (ed) (2006). *Pioneers of Financial Economics: Contributions prior to Irving Fisher*, Edward Elgar, Cheltenham, Vol. 1.
- Rubinstein, M. (1975). Securities market efficiency in an Arrow-Debreu economy, *The American Economic Review* **65**(5), 812–824.

Related Articles

Bachelier, Louis (1870–1946); Efficient Market Hypothesis.

FRANCK JOVANOVIĆ

Econophysics

The Prehistoric Times of Econophysics

The term *econophysics* was introduced in the 1990s, endorsed in 1999 by the publication of Mantegna & Stanley's "An Introduction to Econophysics" [33]. The word "econophysics", paralleling the quests of biophysics or geophysics, suggests that there is a physics-based approach to economics.

From classical to neoclassical economics and until now, economists have been inspired by the conceptual and mathematical developments of the physical sciences and by their remarkable successes in describing and predicting natural phenomena. Reciprocally, physics has been enriched several times by developments first observed in economics. Well before the christening of econophysics as the incarnation of the multidisciplinary study of complex large-scale financial and economic systems, a multitude of small and large collisions have punctuated the development of these two fields. We now mention a few that illustrate the remarkable commonalities and interfertilization.

In his "Inquiry into the Nature and Causes of the Wealth of Nations" (1776), Adam Smith found inspiration in the *Philosophiae Naturalis Principia Mathematica* (1687) of Isaac Newton, specifically based on the (novel at the time) notion of causative forces.

The recognition of the importance of feedbacks to fathom the sheer complexity of economic systems has been at the root of economic thinking for a long time. Toward the end of the nineteenth century, the microeconomists Francis Edgeworth and Alfred Marshall drew on some of the ideas of physicists to develop the notion that the economy achieves an equilibrium state like that described for gases by Clerk Maxwell and Ludwig Boltzmann. The general equilibrium theory now at the core of much of economic thinking is nothing but a formalization of the idea that "everything in the economy affects everything else" [18], reminiscent of mean-field theory or self-consistent effective medium methods in physics, but emphasizing and transcending these ideas much beyond their initial sense in physics.

While developing the field of microeconomics in his "Cours d'Economie Politique" (1897), the economist and philosopher Vilfredo Pareto was

the first to describe, for the distribution of incomes, the eponym power laws that would later become the center of attention of physicists and other scientists observing this remarkable and universal statistical signature in the distribution of event sizes (earthquakes, avalanches, landslides, storms, forest fires, solar flares, commercial sales, war sizes, and so on) punctuating so many natural and social systems [3, 29, 35, 41].

While attempting to model the erratic motion of bonds and stock options in the Paris *Bourse* in 1900, mathematician Louis Bachelier developed the mathematical theory of diffusion (and the first elements of financial option pricing) and solved the parabolic diffusion equation five years before Albert Einstein [10] established the theory of Brownian motion based on the same diffusion equation (also underpinning the theory of random walks) in 1905. The ensuing modern theory of random walks now constitutes one of the fundamental pillars of theoretical physics and economics and finance models.

In the early 1960s, mathematician Benoit Mandelbrot [28] pioneered the use in financial economics of heavy-tailed distributions (Lévy stable laws) as opposed to the traditional Gaussian (normal) law. A cohort of economists, notably at the University of Chicago (Merton Miller, Eugene Fama, and Richard Roll), at MIT (Paul Samuelson), and at Carnegie Mellon University (Thomas Sargent), initially followed his steps. In his PhD thesis, Eugene Fama confirmed that the frequency distribution of the changes in the logarithms of prices was "leptokurtic", that is, with a high peak and fat tails. However, other notable economists (Paul Cootner and Clive Granger) opposed Mandelbrot's proposal, on the basis of the argument that "the statistical theory that exists for the normal case is nonexistent for the other members of the class of Lévy laws." The *coup de grace* was the mounting empirical evidence that the distributions of returns were becoming closer to the Gaussian law at timescales larger than one month, at odds with the self-similarity hypothesis associated with the Lévy laws [7, 23]. Much of the efforts in the econophysics literature of the late 1990s and early 2000s revisited and refined this hypothesis, confirming on one hand the existence of the variance (which rules out the class of Lévy distributions proposed by Mandelbrot), but also suggesting a power-law tail with an exponent close to 3 [16, 32]—several other groups have discussed alternatives, such as exponential [39]

or stretched exponential distributions [19, 24, 26]. Financial engineers actually care about these apparent technicalities because the tail structure controls the Value at Risk and other measures of large losses, and physicists care because the tail may constrain the underlying mechanism(s). For instance, Gabaix *et al.* [14] attribute the large movements in stock market activity to the interplay between the power-law distribution of the sizes of large financial institutions and the optimal trading of such large institutions. In this domain, econophysics focuses on models that can reproduce and explain the main stylized facts of financial time series: non-Gaussian fat tail distribution of returns, long-range autocorrelation of volatility and the absence of correlation of returns, multifractal property of the absolute value of returns, and so on.

In the late 1960s, Benoit Mandelbrot left financial economics but, inspired by this first episode, went on to explore other uncharted territories to show how nondifferentiable geometries (that he coined *fractal*), previously developed by mathematicians from the 1870s to the 1940s, could provide new ways to deal with the real complexity of the world [29]. He later returned to finance in the late 1990s in the midst of the econophysics' enthusiasm to model the multifractal properties associated with the long-memory properties observed in financial asset returns [2, 30, 31, 34, 43].

Notable Contributions

The modern econophysicists are implicitly and sometimes explicitly driven by the hope that the concept of "universality" holds in economics and finance. The value of this strategy remains to be validated [42], as most econophysicists have not yet digested the subtleties of economic thinking and failed to marry their ideas and techniques with mainstream economics. The following is a partial list of a few notable exceptions: precursory physics approach to social systems [15], agent-based models, induction, evolutionary models [1, 9, 11, 21], option theory for incomplete markets [4, 6], interest rate curves [5, 38], minority games [8], theory of Zipf law and its economic consequences [12, 13, 27], theory of large price fluctuations [14], theory of bubbles and crashes [17, 22, 40], random matrix theory applied

to covariance of returns [20, 36, 37], and methods and models of dependence between financial assets [25, 43].

At present, the most exciting progresses seem to be unraveling at the boundary between economics and the biological, cognitive, and behavioral sciences. While it is difficult to argue for a physics-based foundation of economics and finance, physics has still a role to play as a unifying framework full of concepts and tools to deal with the complex. The modeling skills of physicists explain their impressive number in investment and financial institutions, where their data-driven approach coupled with a pragmatic sense of theorizing has made them a most valuable commodity on Wall Street.

Acknowledgments

We would like to thank Y. Malevergne for many discussions and a long-term enjoyable and fruitful collaboration.

References

- [1] Arthur, W.B. (2005). Out-of-equilibrium economics and agent-based modeling, in *Handbook of Computational Economics, Vol. 2: Agent-Based Computational Economics*, K. Judd & L. Tesfatsion, eds, Elsevier, North Holland.
- [2] Bacry, E., Delour, J. & Muzy, J.-F. (2001). Multifractal random walk, *Physical Review E* **64**, 026103.
- [3] Bak, P. (1996). *How Nature Works: The Science of Self-Organized Criticality*, Copernicus, New York.
- [4] Bouchaud, J.-P. & Potters, M. (2003). Theory of financial risk and derivative pricing, *From Statistical Physics to Risk Management*, 2nd Edition, Cambridge University Press.
- [5] Bouchaud, J.-P., Sagna, N., Cont, R., El-Karoui, N. & Potters, M. (1999). Phenomenology of the interest rate curve, *Applied Mathematical Finance* **6**, 209.
- [6] Bouchaud, J.-P. & Sornette, D. (1994). The Black-Scholes option pricing problem in mathematical finance: generalization and extensions for a large class of stochastic processes, *Journal de Physique I France* **4**, 863–881.
- [7] Campbell, J.Y., Lo, A.W. & MacKinlay, A.C. (1997). *The Econometrics of Financial Markets*, Princeton University Press, Princeton.
- [8] Challet, D., Marsili, M. & Zhang, Y.-C. (2005). *Minority Games*, Oxford University Press, Oxford.
- [9] Cont, R. & Bouchaud, J.-P. (2000). Herd behavior and aggregate fluctuations in financial markets, *Journal of Macroeconomic Dynamics* **4**(2), 170–195.

- [10] Einstein, A. (1905). On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat, *Annalen der Physik* **17**, 549–560.
- [11] Farmer, J.D. (2002). Market forces, ecology and evolution, *Industrial and Corporate Change* **11**(5), 895–953.
- [12] Gabaix, X. (1999). Zipf's law for cities: an explanation, *Quarterly Journal of Economics* **114**(3), 739–767.
- [13] Gabaix, X. (2005). *The Granular Origins of Aggregate Fluctuations*, working paper, Stern School of Business, New York.
- [14] Gabaix, X., Gopikrishnan, P., Plerou, V. & Stanley, H.E. (2003). A theory of power law distributions in financial market fluctuations, *Nature* **423**, 267–270.
- [15] Galam, S. & Moscovici, S. (1991). Towards a theory of collective phenomena: consensus and attitude changes in groups, *European Journal of Social Psychology* **21**, 49–74.
- [16] Gopikrishnan, P., Plerou, V., Amaral, L.A.N., Meyer, M. & Stanley, H.E. (1999). Scaling of the distributions of fluctuations of financial market indices, *Physical Review E* **60**, 5305–5316.
- [17] Johansen, A., Sornette, D. & Ledoit, O. (1999). Predicting financial crashes using discrete scale invariance, *Journal of Risk* **1**(4), 5–32.
- [18] Krugman, P. (1996). *The Self-Organizing Economy*, Blackwell, Malden.
- [19] Laherrere, J. & Sornette, D. (1999). Stretched exponential distributions in nature and economy: fat tails with characteristic scales, *European Physical Journal B* **2**, 525–539.
- [20] Laloux, L., Cizeau, P., Bouchaud, J.-P. & Potters, M. (1999). Noise dressing of financial correlation matrices, *Physical Review Letters* **83**, 1467–1470.
- [21] Lux, T. & Marchesi, M. (1999). Scaling and criticality in a stochastic multi-agent model of financial market, *Nature* **397**, 498–500.
- [22] Lux, T. & Sornette, D. (2002). On rational bubbles and fat tails, *Journal of Money, Credit and Banking, Part 1* **34**(3), 589–610.
- [23] MacKenzie, D. (2006). *An Engine, Not a Camera: How Financial Models Shape Markets*, The MIT Press, Cambridge, London.
- [24] Malevergne, Y., Pisarenko, V.F. & Sornette, D. (2005). Empirical distributions of log-returns: between the stretched exponential and the power law? *Quantitative Finance* **5**(4), 379–401.
- [25] Malevergne, Y. & Sornette, D. (2003). Testing the Gaussian copula hypothesis for financial assets dependences, *Quantitative Finance* **3**, 231–250.
- [26] Malevergne, Y. & Sornette, D. (2006). *Extreme Financial Risks: From Dependence to Risk Management*, Springer, Heidelberg.
- [27] Malevergne, Y. & Sornette, D. (2007). *A two-factor Asset Pricing Model Based on the Fat Tail Distribution of Firm Sizes*, ETH Zurich working paper. <http://arxiv.org/abs/physics/0702027>
- [28] Mandelbrot, B.B. (1963). The variation of certain speculative prices, *Journal of Business* **36**, 394–419.
- [29] Mandelbrot, B.B. (1982). *The Fractal Geometry of Nature*, W.H. Freeman, San Francisco.
- [30] Mandelbrot, B.B. (1997). *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*, Springer, New York.
- [31] Mandelbrot, B.B., Fisher, A. & Calvet, L. (1997). *A Multifractal Model of Asset Returns*, Cowles Foundation Discussion Papers 1164, Cowles Foundation, Yale University.
- [32] Mantegna, R.N. & Stanley, H.E. (1995). Scaling behavior in the dynamics of an economic index, *Nature* **376**, 46–49.
- [33] Mantegna, R. & Stanley, H.E. (1999). *An Introduction to Econophysics: Correlations and Complexity in Finance*, Cambridge University Press, Cambridge and New York.
- [34] Muzy, J.-F., Sornette, D., Delour, J. & Arneodo, A. (2001). Multifractal returns and hierarchical portfolio theory, *Quantitative Finance* **1**, 131–148.
- [35] Newman, M.E.J. (2005). Power laws, Pareto distributions and Zipf's law, *Contemporary Physics* **46**, 323–351.
- [36] Pafka, S. & Kondor, I. (2002). Noisy covariance matrices and portfolio optimization, *European Physical Journal B* **27**, 277–280.
- [37] Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N. & Stanley, H.E. (1999). Universal and non-universal properties of cross correlations in financial time series, *Physical Review Letters* **83**(7), 1471–1474.
- [38] Santa-Clara, P. & Sornette, D. (2001). The dynamics of the forward interest rate curve with stochastic string shocks, *The Review of Financial Studies* **14**(1), 149–185.
- [39] Silva, A.C., Prange, R.E. & Yakovenko, V.M. (2004). Exponential distribution of financial returns at mesoscopic time lags: a new stylized fact, *Physica A* **344**, 227–235.
- [40] Sornette, D. (2003). *Why Stock Markets Crash, Critical Events in Complex Financial Systems*, Princeton University Press.
- [41] Sornette, D. (2006). *Critical Phenomena in Natural Sciences, Chaos, Fractals, Self-organization and Disorder: Concepts and Tools, Series in Synergetics*, 2nd Edition, Springer, Heidelberg.
- [42] Sornette, D., Davis, A.B., Ide, K., Vixie, K.R., Pisarenko, V. & Kamm, J.R. (2007). Algorithm for model validation: theory and applications, *Proceedings of the National Academy of Sciences of the United States of America* **104**(16), 6562–6567.
- [43] Sornette, D., Malevergne, Y. & Muzy, J.F. (2003). What causes crashes? *Risk* **16**, 67–71. <http://arXiv.org/abs/cond-mat/0204626>

Further Reading

Bachelier, L. (1900). *Théorie de la speculation*, *Annales de l'Ecole Normale Supérieure* (translated in the book *Random Character of Stock Market Prices*), Théorie des probabilités continues, 1906, *Journal des Mathématiques Pures*

- et Appliquées*; Les Probabilités cinématiques et dynamiques, 1913, *Annales de l'Ecole Normale Supérieure*.
- Cardy, J.L. (1996). *Scaling and Renormalization in Statistical Physics*, Cambridge University Press, Cambridge.
- Pareto, V. (1897). *Cours d'Économie Politique*, Macmillan, Paris, Vol. 2.
- Stanley, H.E. (1999). Scaling, universality, and renormalization: three pillars of modern critical phenomena, *Reviews of Modern Physics* **71**(2), S358–S366.

GILLES DANIEL & DIDIER SORNETTE

Kolmogorov, Andrei Nikolaevich

Andrei Nikolaevich Kolmogorov was born on April 25, 1903 and died on October 20, 1987 in the Soviet Union.

Springer Verlag published (in German) Kolmogorov's monograph "Foundations of the Theory of Probability" more than seventy-five years ago [3]. In this small, 80-page book, he not only provided the logical foundation of the mathematical theory of probability (axiomatics) but also defined new concepts: conditional probability as a random variable, conditional expectations, notion of independency, the use of Borel fields of probability, and so on. The "Main theorem" in Chapter III "Probability in Infinite Spaces" indicated how to construct stochastic processes starting from their finite-dimensional distributions. His approach has made the development of modern mathematical finance possible.

Before writing "Foundations of the Theory of Probability", Kolmogorov wrote his great paper "Analytical Methods in Probability Theory" [2], which gave birth to the theory of Markov processes in continuous time. In this paper, Kolmogorov presented his famous *forward and backward* differential equations, which are the often-used tools in probability theory and its applications. He also gave credit to L. Bachelier for the latter's pioneering investigations of probabilistic schemes evolving continuously in time.

The two works mentioned earlier laid the groundwork for all subsequent developments of the theory of probability and stochastic processes. Today, it is impossible to imagine the state of these sciences without Kolmogorov's contributions.

Kolmogorov developed many fundamentally important concepts that have determined the progress in different branches of mathematics and other branches of science and arts. Being an outstanding mathematician and scientist, he obtained, besides fundamental results in the theory of probability

[5], the theory of trigonometric series, measure and set theory, the theory of integration, approximation theory, constructive logic, topology, the theory of superposition of functions and Hilbert's thirteenth problem, classical mechanics, ergodic theory, the theory of turbulence, diffusion and models of population dynamics, mathematical statistics, the theory of algorithms, information theory, the theory of automata and applications of mathematical methods in humanitarian sciences (including work in the theory of poetry, the statistics of text, and history), and the history and methodology of mathematics for school children and teachers of school mathematics [4–6]. For more descriptions of Kolmogorov's works, see [1, 7].

References

- [1] Bogolyubov, N.N., Gnedenko, B.V. & Sobolev, S.L. (1983). Andrei Nikolaevich Kolmogorov (on his eighteenth birthday), *Russian Mathematical Surveys* **38**(4), 9–27.
- [2] Kolmogoroff, A. (1931). Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung, *Mathematische Annalen*, **104**, 415–458.
- [3] Kolmogoroff, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin.
- [4] Kolmogorov, A.N. (1991). Mathematics and mechanics, in *Mathematics and its Applications (Soviet Series 25)*, V.M. Tikhomirov, ed., Kluwer, Dordrecht, Vol. I, pp. xx+551.
- [5] Kolmogorov, A.N. (1992). Probability theory and mathematical statistics, in *Mathematics and its Applications (Soviet Series 26)*, A.N. Shiryaev, ed., Kluwer, Dordrecht, Vol. II, pp. xvi+597.
- [6] Kolmogorov, A.N. (1993). Information theory and the theory of algorithms, in *Mathematics and its Applications (Soviet Series 27)*, A.N. Shiryaev, ed., Kluwer, Dordrecht, Vol. III, pp. xxvi+275.
- [7] Shiryaev, A.N. (2000). Andrei Nikolaevich Kolmogorov (April 25, 1903 to October 20, 1987). A biographical sketch of his life and creative paths, in *Kolmogorov in Perspective*, American Mathematical Society, London Mathematical Society, pp. 1–87.

ALBERT N. SHIRYAEV

Bernoulli, Jacob

Jacob Bernoulli (1654–1705), the son and grandson of spice merchants in the city of Basel, Switzerland, was trained to be a protestant clergyman, but, following his own interests and talents, instead became the professor of mathematics at the University of Basel from 1687 until his death. He taught mathematics to his nephew Nicolaus Bernoulli (1687–1759) and to his younger brother Johann (John, Jean) Bernoulli (1667–1748), who was trained in medicine, but took over as professor of mathematics at Basel after Jacob's death in 1705. As a professor of mathematics, Johann Bernoulli, in turn, taught mathematics to his sons, including Daniel Bernoulli (1700–1782), known for the St. Petersburg paradox in probability, as well as for work in hydrodynamics. Jacob and Johann Bernoulli were among the first to read and understand Gottfried Wilhelm Leibniz's articles in the *Acta Eruditorum* of 1684 and 1686, in which Leibniz put forth the new algorithm of calculus. They helped to develop and spread Leibniz's calculus throughout Europe, Johann teaching calculus to the Marquis de Hôpital, who published the first calculus textbook. Nicolas Bernoulli wrote his master's thesis [1] on the basis of the manuscripts of Jacob's still unpublished *Art of Conjecturing*, and helped to spread its contents in the years between Jacob's death and the posthumous publication of Jacob's work in 1713 [2]. In the remainder of this article, the name "Bernoulli" without any first name refers to Jacob Bernoulli. (Readers should be aware that many Bernoulli mathematicians are not infrequently confused with each other. For instance, it was Jacob's *son* Nicolaus, also born in 1687, but a painter and not a mathematician, who had the Latin manuscript of [2] printed, and not his *nephew* Nicolaus, although the latter wrote a brief preface.)

As far as the application of the art of conjecturing to economics (or finance) is concerned, much of the mathematics that Jacob Bernoulli inherited relied more on law and other institutional factors than it relied on statistics or mathematical probability, a discipline that did not then exist. Muslim traders had played a significant role in Mediterranean commerce in the medieval period and in the development of mathematics, particularly algebra, as well. Muslim mathematical methods were famously transmitted to

Europe by Leonardo of Pisa, also known as *Fibonacci* [6]. Rather than relying on investments with guaranteed rates of return, which were frowned upon as involving usury, Muslim trade was often carried out by partnerships or companies, many involving members of extended families. Such partnerships would be based on a written contract between those involved, spelling out the agreed-upon division of the profits once voyagers had returned and the goods had been sold, the shares of each partner depending upon their investment of cash, supply of capital goods such as ships or warehouses, and labor. According to the Islamic law, if one of the partners in such an enterprise died before the end of the anticipated period of the venture, his heirs were entitled to demand the dissolution of the firm, so that they might receive their legal inheritances. Not infrequently, applied mathematicians were called upon to calculate the value of the partnership on a given intermediate date, so that the partnership could be dissolved fairly.

In Arabic and then Latin books of commercial arithmetic or business mathematics in general (geometry, for instance, volumes of barrels, might also be included), there were frequently problems of "societies" or partnerships, which later evolved into the so-called "problem of points" concerning the division of the stakes of a gambling game if it were terminated before its intended end. Typically, the values of the various partners' shares were calculated using (i) the amounts invested; (ii) the length of time it was invested in the company if all the partners were not equal in this regard; and (iii) the original contract, which generally specified the division of the capital and profits among partners traveling to carry out the business and those remaining at home. The actual mathematics involved in making these calculations was similar to the mathematics of calculating the price of a mixture [2, 7, 8]. (If, as was often the case, "story problems" were described only in long paragraphs, what was intended might seem much more complex than if everything could have been set out in the subsequently developed notation of algebraic equations.)

In Part IV of [2], Bernoulli had intended to apply the mathematics of games of chance, expounded in Parts I–III of the book on the basis of Huygens' work, by analogy, to civil, moral, and economic problems. The fundamental principle of Huygens' and Bernoulli's mathematics of games of chance was that the game should be fair and that players should

pay to play a game in proportion to their expected winnings. Most games, like business partnerships, were assumed to involve only the players, so that the total paid in would equal the total paid out at the end. Here, a key concept was the number of “cases” or possible alternative outcomes. If a player might win a set amount if a die came up a 1, then there were said to be six cases, corresponding to the six faces of the die, of which one, the 1, would be favorable to that player. For this game to be fair, the player should pay in one-sixth of the amount he or she would win if the 1 were thrown.

Bernoulli applied this kind of mathematics in an effort to quantify the evidence that an accused person had committed a crime by systematically combining all the various types of circumstantial evidence of the crime. He supposed that something similar might be done to judge life expectancies, except that no one knew all the “cases” that might affect life expectancy, such as the person’s inherited vigor and healthiness, the diseases to which a person might succumb, the accidents that might happen, and so forth. With the law that later came to be known as the *weak law of large numbers*, Bernoulli proposed to discover *a posteriori* from the results many times observed in similar situations what the ratios of unobserved underlying “cases” might be. Most people realize, Bernoulli said, that if you want to judge what may happen in the future by what has happened in the past, you are less liable to be mistaken if you have made more observations or have a longer time series of outcomes. What people do not know, he said, is whether, if you make more and more observations, you can be more and more sure, without limit, that your prediction is reliable. By his proof he claimed to show that there was no limit to the degree of confidence or probability one might have that the ratio of results would fall within some interval around an expected ratio. In addition, he made a rough calculation of the number of trials (later called *Bernoulli trials*) that would be needed for a proposed degree of certainty. The mathematics he used in his proof basically involved binomial expansions and the possible combinations and permutations of outcomes (“successes” or “failures”) over a long series of trials. After a long series of trials, the distribution of ratios of outcomes would take the shape of a bell curve, with increasing percentages of outcomes clustering around the central value. For a comparison of Jacob

Bernoulli’s proof with Nicolaus Bernoulli’s proof of the same theorem, see [5].

In correspondence with Leibniz, Bernoulli unsuccessfully tried to obtain from Leibniz a copy of Jan De Witt’s rare pamphlet, in Dutch, on the mathematics of annuities—this was the sort of problem to which he hoped to apply his new mathematical theory [4]. Leibniz, in reply, without having been told the mathematical basis of Bernoulli’s proof of his law for finding, *a posteriori*, ratios of cases, for instance, of surviving past a given age, objected that no such approach would work because the causes of death might be changeable over time. What if a new disease should make an appearance, leading to an increase of early deaths? Bernoulli’s reply was that, if there were such changed circumstances, then it would be necessary to make new observations to calculate new ratios for life expectancies or values of annuities [2].

But what if not only were there no fixed ratios of cases over time, but no such regularities (underlying ratios of cases) at all? For Bernoulli this was not a serious issue because he was a determinist, believing that from the point of view of the Creator everything is determined and known eternally. It is only because we humans do not have such godlike knowledge that we cannot know the future in detail. Nevertheless, we can increase the security and prudence of our actions through the application of the mathematical art of conjecturing that he proposed to develop. Even before the publication of *The Art of Conjecturing*, Abraham De Moivre had begun to carry out with great success the program that Bernoulli had begun [3]. Although, for Bernoulli, probability was an epistemic concept, and expectation was more fundamental than relative chances, De Moivre established mathematical probability on the basis of relative frequencies.

References

- [1] Bernoulli, N. (1709). De Usu Artis Conjectandi in Jure, in *Die Werke von Jacob Bernoulli III*, B.L. vander Waerden, ed., Birkhäuser, Basel, pp. 287–326. An English translation of Chapter VII can be found at http://www.york.ac.uk/depts/mathes/histstat/bernoulli_n.htm [last access December 13, 2008].
- [2] Bernoulli, J. (2006). [Ars Conjectandi (1713)], English translation in Jacob Bernoulli, *The Art of Conjecturing together with Letter to a Friend on Sets in Court Tennis*, E.D. Sylla ed., The Johns Hopkins University Press, Baltimore.

-
- [3] De Moivre, A. (1712). De Mensura Sortis, seu, de Probabilitate Eventuum in Ludis a Casu Fortuito Pendentibus *Philosophical Transactions of the Royal Society* **27**, 213–264 ; translated by Bruce McClintock in Hald, A. (1984a). A. De Moivre: ‘De Mensura Sortis’ or ‘On the Measurement of Chance’ . . . Commentary on ‘De Mensura Sortis’, *International Statistical Review* **52**, 229–262. After Bernoulli’s *The Art of Conjecturing*, De Moivre published *The Doctrine of Chances*, London 1718, 1738, 1756.
 - [4] De Witt, J. (1671). Waerdye van Lyf-renten, in *Die Werke von Jacob Bernoulli III*, B.L. vander Waerden, ed., Birkhäuser, Basel, pp. 328–350.
 - [5] Hald, A. (1984b). Nicholas Bernoulli’s theorem, *International Statistical Review* **52**, 93–99 ; Cf. Hald, A. (1990). *A History of Probability and Statistics and Their Applications before 1750*, Wiley, New York.
 - [6] Leonardo of Pisa (Fibonacci) (2002). [*Liber Abaci* (1202)], English translation in *Fibonacci’s Liber Abaci: A Translation into Modern English of Leonardo Pisano’s Book of Calculation*, Springer Verlag, New York.
 - [7] Sylla, E. (2003). Business ethics, commercial mathematics, and the origins of mathematical probability, in *Oeconomies in the Age of Newton*, M. Schabas & N.D. Marchi, eds, Annual Supplement to History of Political Economy, Duke University Press, Durham, Vol. 35, pp. 309–327.
 - [8] Sylla, E. (2006). Revised and expanded version of [7]: “Commercial Arithmetic, theology and the intellectual foundations of Jacob Bernoulli’s *Art of Conjecturing*”, in G. Poitras, ed., *Pioneers of Financial Economics*, Contributions Prior to Irving Fisher, Edward Elgar Publishing, Cheltenham UK and Northampton MA, Vol. 1.

EDITH DUDLEY SYLLA

Treynor, Lawrence Jack

Jack Lawrence Treynor was born in Council Bluffs, Iowa, on February 21, 1930 to Jack Vernon Treynor and Alice Cavin Treynor. In 1951, he graduated from Haverford College on Philadelphia's Main Line with a Bachelors of Arts degree in mathematics. He served two years in the US Army before moving to Cambridge, MA to attend Harvard Business School. After a year writing cases for Professor Robert Anthony, Treynor went to work for the Operations Research department at Arthur D. Little in 1956.

Treynor was particularly inspired by the 1958 paper coauthored by Franco Modigliani and Merton H. Miller, titled "The Cost of Capital, Corporation Finance, and the Theory of Investment." At the invitation of Modigliani, Treynor spent a sabbatical year at MIT between 1962 and 1963. While at MIT, Treynor made two presentations to the finance faculty, the first of which, "Toward a Theory of the Market Value of Risky Assets," introduced the capital asset pricing model (CAPM). The CAPM says that the return on an asset should equal the rate on a risk-free rate plus a premium proportional to its contribution to the risk in the market portfolio. The model is often referred to as the *Treynor–Sharpe–Lintner–Mossin CAPM* to reflect the fact that it was simultaneously and independently developed by multiple individuals, albeit with slight differences. Although Treynor's paper was not published until Robert Korajczyk included the unrevised version in his 1999 book, *Asset Pricing and Portfolio Performance*, it is also included in the "Risk" section of Treynor's own 2007 book, *Treynor on Institutional Investing* (Wiley, 2008). William F. Sharpe's 1964 version, which was built on the earlier work of Harry M. Markowitz, won the Nobel Prize for Economics in 1990.

The CAPM makes no assumptions about the factor structure of the market. In particular, it does not assume the single-factor structure of the so-called market model. However, in his *Harvard Business Review* papers on performance measurement, Treynor assumed a single factor. He used a regression of returns on managed funds against returns on the "market" to estimate the sensitivity of the fund to the market factor and then used the slope of that regression line to estimate the contribution of market fluctuations to a fund's rate of return, which

permitted him to isolate the portion of fund return that was actually due to the selection skills of the fund manager. In 1981, Fischer Black wrote an open letter in the *Financial Analysts Journal*, stating that Treynor had "developed the capital asset pricing model before anyone else."

In his second *Harvard Business Review* paper, Treynor and Kay Mazuy used a curvilinear regression line to test whether funds were more sensitive to the market in the years when the market went up *versus* the years when the market went down.

When Fischer Black arrived at Arthur D. Little in 1965, Black took an interest in Treynor's work and later inherited Treynor's caseload (after Treynor went to work for Merrill Lynch.) In their paper, "How to Use Security Analysis to Improve Portfolio Selection," Treynor and Black proposed viewing portfolios as having three distinct parts: a riskless part, a highly diversified part (devoid of specific risk), and an active part (which would have both specific risk and market risk). The paper spells out the optimal balance, not only between the three parts but also between the individual securities in the active part.

In 1966, Treynor was hired by Merrill Lynch where he headed Wall Street's first quantitative research group. Treynor left Merrill Lynch in 1969 to serve as the editor of the *Financial Analysts Journal*, with which he stayed until 1981. Treynor then joined Harold Arbit in starting Treynor–Arbit Associates, an investment firm based in Chicago. Treynor continues to serve on the advisory boards of the *Financial Analysts Journal* and the *Journal of Investment Management*, where he is also case editor.

In addition to his 1976 book published with William Priest and Patrick Regan titled *The Financial Reality of Pension Funding under ERISA*, Treynor coauthored *Machine Tool Leasing* in 1956 with Richard Vancil of Harvard Business School. Treynor has authored and co-authored more than 90 papers on such topics as risk, performance measurement, economics, trading (market microstructure), accounting, investment value, active management, and pensions. He has also written 20 cases, many published in the *Journal of Investment Management*.

Treynor's work has appeared in the *Financial Analysts Journal*, the *Journal of Business*, the *Harvard Business Review*, the *Journal of Finance*, and the *Journal of Investment Management*, among others. Some of Treynor's works were published under the pen-name "Walter Bagehot," a cover that offered him

anonymity while allowing him to share his often unorthodox theories. He promoted notions such as random walks, efficient markets, risk/return trade-off, and betas that others in the field actively avoided. Treynor has since become renowned not only for pushing the envelope with new ideas but also for encouraging others to do the same as well. Eighteen of his papers have appeared in anthologies.

Two papers that have not been anthologized are “Treynor’s Theory of Inflation” and “Will the Phillips Curve Cause World War III?” In these papers, he points out that, because in industry labor and capital are complements (rather than substitutes, as depicted in economics textbooks), over the business cycle they will become more or less scarce together. However, when capital gets more or less scarce, the identity of the marginal machine will change. If the real wage is determined by the marginal productivity of labor then (as Treynor argues) it is determined by the labor productivity of the marginal machine. As demand rises and the marginal machines get older and less efficient, the real wage falls, but labor negotiations fix the money wage. In order to satisfy the identity

$$\text{money prices} \equiv \frac{\text{money wage}}{\text{real wage}} \quad (1)$$

when the real wage falls, money prices must rise. According to Nobel Laureate Merton Miller, Treynor’s main competitor on the topic, the Phillips curve is “just an empirical regularity” (i.e., just data snooping).

Treynor has won the *Financial Analysts Journal*’s Graham and Dodd Scroll award in 1968, 1982, twice in 1987, for “The Economics of the Dealer Function” and “Market Efficiency and the Bean Jar Experiment,” in 1998 for “Bulls Bears and Market Bubbles”, and in 1999 for “The Investment Value of Brand Franchise.” In 1981 Treynor was again recognized for his research, winning the Graham and Dodd award for “Best Paper” titled “What Does It Take to Win the Trading Game?” In 1987, he was presented with the James R. Vertin Award of the Research Foundation of the Institute of Chartered Financial Analysts, “in recognition of his research, notable for its relevance and enduring value to investment professionals.” In addition, the Financial Analysts Association presented him with the Nicholas Molodovsky Award in 1985, “in recognition of his outstanding contributions to the profession of financial analysis of such significance

as to change the direction of the profession and raise it to higher standards of accomplishment.” He received the Roger F. Murray prize in 1994 from the Institute of Quantitative Research in Finance for “Active Management as an Adversary Game.” That same year he was also named a Distinguished Fellow of the Institute for Quantitative Research in Finance along with William Sharpe, Merton Miller, and Harry Markowitz. In 1997, he received the EBRI Lillywhite Award, which is “awarded to persons who have had distinguished careers in the investment management and employee benefits fields and whose outstanding service enhances Americans’ economic security.” In 2007, he was presented with The Award for Professional Excellence, presented periodically by the CFA Institute Board to “a member of the investment profession whose exemplary achievement, excellence of practice, and true leadership have inspired and reflected honor upon our profession to the highest degree” (Previous winners were Jack Bogle and Warren Buffett.). In 2008, he was recognized as the 2007 IAFE/SunGard Financial Engineer of the Year for his contributions to financial theory and practice.

Treynor taught investments at Columbia University while working at the *Financial Analysts Journal*. Between 1985 and 1988, Treynor taught investments at the University of Southern California.

He is currently President of Treynor Capital Management in Palos Verdes, California.

Further Reading

- Bernstein, P.L. (1992). *Capital Ideas: The Improbable Origins of Modern Wall Street*, The Free Press, New York.
- Black, F.S. (1981). An open letter to Jack Treynor, *Financial Analysts Journal* July/August, 14.
- Black, F.S. & Treynor, J.L. (1973). How to use security analysis to improve portfolio selection, *The Journal of Business* **46**(1), 66–88.
- Black, F.S. & Treynor, J.L. (1986). Corporate investment decision, in *Modern Developments in Financial Management*, S.C. Myers, ed., Praeger Publishers.
- French, C. (2003). The Treynor capital asset pricing model, *Journal of Investment Management* **1**(2), 60–72.
- Keynes, J.M. (1936). *The General Theory of Employment, Interest, and Money*, Harcourt Brace, New York.
- Korajczyk, R. (1999). *Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics*, Risk Books, London.
- Lintner, J. (1965a). The valuation of risk assets and the selection of risky investment in stock portfolios and capital budgets, *The Review of Economics and Statistics* **47**, 13–37.

- Lintner, J. (1965b). Securities prices, risk, and maximal gains from diversification, *The Journal of Finance* **20**(4), 587–615.
- Markowitz, H.M. (1952). Portfolio selection, *The Journal of Finance* **7**(1), 77–91.
- Mehrling, P. (2005). *Fischer Black and the Revolutionary Idea of Finance*, Wiley, New York.
- Modigliani, F. & Miller, M.H. (1958). The cost of capital, corporation finance, and the theory of investment, *The American Economic Review* **48**, 261–297.
- Sharpe, W.F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk, *The Journal of Finance* **19**(3), 425–442.
- Treynor, J.L. (1961). *Market Value, Time, and Risk*. Unpublished manuscript. Dated 8/8/1961, #95-209.
- Treynor, J.L. (1962). *Toward a Theory of Market Value of Risk Assets*. Unpublished manuscript. Dated Fall of 1962.
- Treynor, J.L. (1963). *Implications for the Theory of Finance*. Unpublished manuscript. Dated Spring of 1963.
- Treynor, J.L. (1965). How to rate management of investment funds, *Harvard Business Review* **43**, 63–75.
- Treynor, J.L. (2007). *Treynor on Institutional Investing*, Wiley, New York.
- Treynor, J.L. & Mazuy, K. (1966). Can mutual funds outguess the market? *Harvard Business Review* **44**, 131–136.
- Treynor, J.L. & Vancil, R. (1956). *Machine Tool Leasing*, Management Analysis Center.

Related Articles

Black, Fischer; Capital Asset Pricing Model; Factor Models; Modigliani, Franco; Samuelson, Paul A.; Sharpe, William F.

ETHAN NAMVAR

Rubinstein, Edward Mark

Mark Rubinstein, the only child of Sam and Gladys Rubinstein of Seattle, Washington, was born on June 8, 1944. He attended the Lakeside School in Seattle and graduated in 1962 as one of the two graduation speakers. He earned an A.B. in Economics, magna cum laude, from Harvard College in 1966 and an MBA with a concentration in finance from the Graduate School of Business at Stanford University in 1968. In 1971, Rubinstein earned his PhD. in Finance from the University of California, Los Angeles (UCLA). During this time at UCLA, he was heavily influenced by the microeconomist Jack Hirshleifer. In July 1972, he became an assistant professor in finance at the University of Californian at Berkeley, where he remained for his entire career. He was advanced to tenure unusually early in 1976 and became a full professor in 1980.

Rubinstein's early work concentrated on asset pricing. Specifically, between 1971 and 1973, his research centered on the mean–variance capital asset pricing model and came to include skewness as a measure of risk [3–5]. Rubinstein's extension has new relevance as several researchers have since determined its predictive power in explaining realized security returns. In 1974, Rubinstein's research turned to more general models of asset pricing. He developed an extensive example of multiperiod security market equilibrium, which later became the dominant model used by academics in their theoretical papers on asset pricing. Unlike earlier work, he left the intertemporal process of security returns to be determined in equilibrium rather than as datum (although as special cases he assumed a random walk and constant interest rates). Rubinstein was thus able to derive conditions for the existence of a random walk and an unbiased term structure of interest rates. He also was the first to derive a simple equation in equilibrium for valuing a risky stream of income received over time. He published the first paper to show explicitly how and why in equilibrium investors would want to hold long-term bonds in their portfolios, and in particular would want to hold a riskless (in terms of income) annuity maturing at their death, foreshadowing several strands of later research.

In 1975, Rubinstein began developing theoretical models of “efficient markets.” In 1976, he published a paper showing that the same formula derived by Black and Scholes for valuing options could come from an alternative set of assumptions based on risk aversion and discrete-time trading opportunities. (Black and Scholes had required continuous trading and continuous price movements.)

Working together with Cox *et al.* [1], Rubinstein published the popular and original paper developing the binomial option pricing model, one of the most widely cited papers in financial economics and now probably the most widely used model by professional traders to value derivatives. The model is often referred to as the *Cox–Ross–Rubinstein option pricing (CRR)* model. At the same time, Rubinstein began work with Cox [2] on their own text, *Options Markets*, which was eventually published in 1985 and won the biennial award of the University of Chicago for the best work by professors of business concerning any area of business.

He supplemented his academic work with firsthand experience as a market maker in options when he became a member of the Pacific Stock Exchange. In 1981, together with Hayne E. Leland and John W. O'Brien, Rubinstein founded the Leland O'Brien Rubinstein (LOR) Associates, the original portfolio insurance firm. At the time, the novel idea of portfolio insurance had been put forth by Leland, later fully developed together with Rubinstein, and successfully marketed among large institutional investors by O'Brien. Their business grew extremely rapidly, only to be cut short when they had to share the blame for the October 1987 stock market crash. Not admitting defeat, LOR invented another product that became the first exchange-traded fund (ETF), the SuperTrust, listed on the American Stock Exchange in 1992. Rubinstein also published a related article examining alternative basket vehicles.

In the early 1990s, Rubinstein published a series of eight articles in the *Risk Magazine* showing how option pricing tools could easily be applied to value a host of so-called exotic derivatives, which were just becoming popular.

Motivated by the failure after 1987 of index options to be priced anywhere close to the predictions of the Black–Scholes formula, in an article

published in the *Journal of Finance* [8], he developed an important generalization of the original binomial model, which he called *implied binomial trees*. The article included new techniques for inferring risk-neutral probability distributions from options on the same underlying asset. Rubinstein's revisions of the model provide the natural generalization of the standard binomial model to accommodate arbitrary expiration date risk-neutral probability distributions. This paper, in turn, spurred new academic work on option pricing in the latter half of the 1990s and found immediate application among various professionals. In 1998 and 1999, Rubinstein rounded out his work on derivatives by publishing a second text titled "Rubinstein on Derivatives," which expanded its domain from calls and puts to futures and more general types of derivatives. The book also pioneered new ways to integrate computers as an aid to learning.

After a 1999 debate about the empirical rationality of financial markets with the key behavioral finance theorist, Richard Thaler, Rubinstein began to rethink the concept of efficient markets. In 2001, he published a version of his conference argument in the *Financial Analysts Journal* [6, 7], titled "Rational Markets? Yes or No: The Affirmative Case," which won the Graham and Dodd Plaque award in 2002.

He then returned to the more general theory of investments with which he had begun his research career as a doctoral student. In 2006, Rubinstein [11] published "A History of the Theory of Investments: My Annotated Bibliography"—an academic history of the theory of investments from the thirteenth to the beginning of the twenty-first century, systematizing the knowledge, and identifying the relations between apparently disparate lines of research. No other book has so far been written that comes close to examining in detail the intellectual path that has led to modern financial economics (particularly, in the subarea of investments). Rubinstein shows that the discovery of key ideas in finance is much more complex and multistaged than anyone had realized. Too few are given too much credit, and sometimes original work has been forgotten.

Rubinstein has taught and lectured widely. During his career, he has given 303 invited lectures, including conference presentations, full course seminars, and honorary addresses all over the United

States and around the world. He has served as chairman of the Berkeley finance group, and as director of the Berkeley Program in Finance; he is the founder of the Berkeley Options Database (the first large transaction-level database ever assembled with respect to options and stocks). He has served on the editorial boards of numerous finance journals. He has authored 62 journal articles, published 3 books, and developed several computer programs dealing with derivatives.

Rubinstein is currently a professor of finance at the Haas School of Business at the University of California, Berkeley. Many of his papers are frequently reprinted in survey publications, and he has won numerous prizes and awards for his research and writing on financial economics. He was named "Businessman of the Year" (one of 12) in 1987 by *Fortune* magazine. In 1995, the International Association of Financial Engineers (IAFE) named him the 1995 IAFE/SunGard Financial Engineer of the Year. In 2000, he was elected to Derivatives Strategy Magazine's "Derivatives Hall of Fame" and named in the "RISK Hall of Fame" by *Risk Magazine* in 2002. Of all his awards, the one he cherishes the most is the 2003 Earl F. Cheit Teaching award in the Masters of Financial Engineering Program at the University of California, Berkeley [10] (Rubinstein, M.E. (2003). *A Short Career Biography*. Unpublished.)

Rubinstein has two grown-up children, Maisie and Judd. He lives with Diane Rubinstein in the San Francisco Bay Area.

References

- [1] Cox, J.C., Ross, S.A. & Rubinstein, M.E. (1979). Optional pricing: a simplified approach, *Journal of Financial Economics* September, 229–263.
- [2] Cox, J.C. & Rubinstein, M.E. (1985). *Options Markets*, Prentice-Hall.
- [3] Rubinstein, M.E. (1973). The fundamental theorem of parameter-preference security valuation, *Journal of Financial and Quantitative Analysis* January, 61–69.
- [4] Rubinstein, M.E. (1973). A comparative statics analysis of risk premiums, *Journal of Business* October.
- [5] Rubinstein, M.E. (1973). A mean-variance synthesis of corporate financial theory, *Journal of Finance* March.
- [6] Rubinstein, M.E. (1989). Market basket alternatives, *Financial Analysts Journal* September/October.
- [7] Rubinstein, M.E. (1989). Rational markets? Yes or No: the affirmative case, *Financial Analysts Journal* May/June.

- [8] Rubinstein, M.E. (1994). Implied binomial trees, *Journal of Finance* July, 771–818.
- [9] Rubinstein, M.E. (2000). *Rubinstein on Derivatives, Risk Books*.
- [10] Rubinstein, M.E. (2003). All in All, it's been a Good Life, *The Growth of Modern Risk Management: A History* July, 581–585.
- [11] Rubinstein, M.E. (2006). *A History of the Theory of Investments: My Annotated Bibliography*, John Wiley & Sons, New York.

ETHAN NAMVAR

Infinite Divisibility

We say that a random variable X has an infinitely divisible (ID) *distribution* (in short X is ID) if for all the integers $n \geq 1$ there exist n independent identically distributed (i.i.d) random variables $X_1, \dots, X_n$, such that $X_1 + \dots + X_n \stackrel{d}{=} X$, where $\stackrel{d}{=}$ is equality in distribution. Alternatively, X (or its distribution μ) is ID if for all $n \geq 1$, μ is the n th convolution $\mu_n * \dots * \mu_n$, where μ_n is a probability distribution.

There are several advantages in using infinitely divisible distributions and processes in financial modeling. First, they offer wide possibilities for modeling alternatives to the Gaussian and stable distributions, while maintaining a link with the central limit theorem and a rich probabilistic structure. Second, they are closely linked to Lévy processes: for each ID distribution μ there is a Lévy process (see **Lévy Processes**) $\{X_t : t \geq 0\}$ with X_1 having distribution μ . Third, every stationary distribution of an Ornstein–Uhlenbeck process (see **Ornstein–Uhlenbeck Processes**) belongs to the class L of ID distributions, which are self-decomposable (SD). We say that a random variable X is SD if it has the linear autoregressive property: for any $\theta \in (0, 1)$, there is a random variable ε_θ independent of X such that $X \stackrel{d}{=} \theta X + \varepsilon_\theta$.

The concept of infinite divisibility in probability was introduced in 1929 by de Fenneti. Its theory was established in the 1930s by Khintchine, Kolmogorov, and Lévy. Motivated by applications arising in different fields, from the 1960s on there was a renewed interest in the subject, in particular, among many other topics, in the study of concrete examples and subclasses of ID distributions. Historical notes and references are found in [3, 6, 8, 9].

Link with the Central Limit Theorem

The class of ID distributions is characterized as the class of possible limit laws for triangular arrays of the form $X_{n,1} + \dots + X_{n,k_n} - a_n$, where $k_n > 0$ is an increasing sequence, $X_{n,1}, \dots, X_{n,k_n}$ are independent random variable for every $n \geq 1$, a_n are normalized constants, and $\{X_{n,j}\}$ is *infinitesimal*: $\lim_{n \rightarrow \infty} \max_{1 \leq j \leq k_n} P(|X_{n,j}| > \epsilon) = 0$, for each

$\epsilon > 0$. On the other hand, the class L of SD distributions is characterized as the class of possible limit laws for normalized sequences of the form $(X_1 + \dots + X_n - a_n)/b_n$, where $X_1, X_2, \dots$ are independent random variables and a_n and $b_n > 0$ are sequences of numbers with $\lim_{n \rightarrow \infty} b_n = \infty$ and $\lim_{n \rightarrow \infty} b_{n+1}/b_n = 1$.

Lévy–Khintchine Representation

In terms of characteristic functions (see **Filtering**), a random variable X is ID if $\varphi(u) = E[e^{iuX}]$ is represented by $\varphi = (\varphi_n)^n$, where φ_n is the characteristic function of a probability distribution for every $n \geq 1$. We define the *characteristic exponent* or *cumulant function* of X by $\Psi(u) = \log \varphi(u)$. The *Lévy–Khintchine representation* establishes that a distribution function μ is ID if and only if its characteristic exponent is represented by

$$\begin{aligned} \Psi(u) = & iau - \frac{1}{2}u^2\sigma^2 \\ & + \int_{\mathbb{R}} (e^{iux} - 1 - iux1_{|x| \leq 1}) \Pi(dx), \quad u \in \mathbb{R} \end{aligned} \quad (1)$$

where $\sigma^2 \geq 0$, $a \in \mathbb{R}$ and Π is a positive measure on $\mathbb{R}$ with no atom at zero and $\int_{\mathbb{R}} \min(1, |x|^2) \Pi(dx) < \infty$. The triplet (a, σ^2, Π) is unique and is called the *generating triplet* of μ , while Π is its *Lévy measure*. When Π is zero, we have the Gaussian distribution. We speak of the purely non-Gaussian case when $\sigma^2 = 0$. When $\Pi(dx) = h(x)dx$ is absolutely continuous, we call the nonnegative function h the Lévy density of Π . Distributions in the class L are also characterized by having Lévy densities of the form $h(x) = |x|^{-1}g(x)$, where g is nondecreasing in $x < 0$ and nonincreasing in $x > 0$.

A nonnegative ID random variable is characterized by a special form of its Lévy–Khintchine representation: it is purely non-Gaussian, $\Pi(-\infty, 0) = 0$, $\int_{|x| \leq 1} |x| \Pi(dx) < \infty$, and

$$\Psi(u) = ia_0u + \int_{\mathbb{R}_+} (e^{iux} - 1) \Pi(dx) \quad (2)$$

where $a_0 \geq 0$ is called the *drift*. The associated Lévy process $\{X_t : t \geq 0\}$ is called a *subordinator*. It is a

nonnegative increasing process having characteristic exponent (2). Subordinators are useful models for random time evolutions.

Several properties of an ID random variable X are related to corresponding properties of its Lévy measure Π . For example, the k th moment $E|X|^k$ is finite if and only if $\int_{|x|>1} |x|^k \Pi(dx)$ is finite. Likewise, for the $\text{ID}_{\log}$ condition: $\int_{|x|>2} \ln|x| \Pi(dx) < \infty$ if and only if $\int_{|x|>2} \ln|x| \mu(dx) < \infty$.

The monograph [8] has a detailed study of multivariate ID distributions and their associated Lévy processes.

Classical Examples and Criteria

The Poisson distribution with mean $\lambda > 0$ is ID with Lévy measure $\Pi(B) = \lambda 1_{\{1\}}(B)$, but is not SD. A compound Poisson distribution is the law of $X = \sum_{i=1}^N Y_i$, where $N, Y_1, Y_2, \dots$ are independent random variables, N having Poisson distribution with mean λ and the Y_i 's have the same distribution G , with $G(\{0\}) = 0$. Any compound Poisson distribution is ID with Lévy measure $\Pi(B) = \lambda G(B)$. This distribution is a building block for all other ID laws, since every ID distribution is the limit of a sequence of compound Poisson distributions.

An important example of an SD law is the gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$. It has Lévy density $h(x) = \alpha x^{-1} e^{-\beta x}$, $x > 0$. The α -stable distribution, with $0 < \alpha < 2$ and purely non Gaussian, is also SD. Its Lévy density is $h(x) = c_1 x^{-1-\alpha} dx$ on $(0, \infty)$ and $h(dx) = c_2 |x|^{-1-\alpha}$ on $(-\infty, 0)$, with $c_1 \geq 0$, $c_2 \geq 0$ and $c_1 + c_2 > 0$.

There is no explicit characterization of infinite divisibility in terms of densities or distributions. However, there are some sufficient or necessary conditions to test for infinite divisibility. A nonnegative random variable with density f is ID in any of the following cases: (i) $\log f$ is convex, (ii) f is completely monotone, or (iii) f is hyperbolically completely monotone [9]. If X is symmetric around zero, it is ID if it has a density that is completely monotone on $(0, \infty)$. For a non-Gaussian ID distribution F , its tail behavior is $-\log(1 + F(-x) - F(x)) = O(x \log x)$, when $x \rightarrow \infty$. Hence, no bounded random variable is ID and if a density has a decay of the type $c_1 \exp(-c_2 x^2)$ with some c_1, c_2 positive and if it is not Gaussian, then F is not ID. An important property

of SD distributions is that they always have densities that are unimodal.

Infinite divisibility is preserved under some mixtures of distributions. One has the surprising fact that any mixture of the exponential distribution is ID: $X \stackrel{d}{=} YV$ is ID whenever V has exponential distribution and Y is an arbitrary nonnegative random variable independent of V . The monograph [9] has a detailed study of ID mixtures.

Stochastic Integral Representations

Several classes of ID distributions are characterized by stochastic integrals (see **Stochastic Integrals**) of a nonrandom function with respect to a Lévy process [2]. The classical example is the class L that is also characterized as all the laws of $X \stackrel{d}{=} \int_0^\infty e^{-t} dZ_t$, where Z_t is a Lévy process having Lévy measure Π_Z with the $\text{ID}_{\log}$ condition. More generally, the stochastic integral $\int_0^1 \log t^{-1} dZ_t$ is well defined for every Lévy process Z_t . Denote by $B(\mathbb{R})$ the class of all the distributions of these stochastic integrals. The class $B(\mathbb{R})$ coincides with those ID laws with completely monotone Lévy density. It is also characterized as the smallest class that contains all mixtures of exponential distributions and is closed under convolution, convergence, and reflection. It is sometimes called the *Bondenson–Goldie–Steutel class of distributions*. Multivariate extensions are presented in [2].

Generalized Gamma Convolutions

The class of generalized gamma convolutions (GGCs) is the smallest class of probability distributions on $\mathbb{R}_+$ that contains all gamma distributions and is closed under convolution and convergence in distribution [6]. These laws are in the class L and have Lévy density of the form $h(x) = x^{-1} g(x)$, $x > 0$, with g a completely monotone function on $(0, \infty)$. Most of the classical distributions on $\mathbb{R}_+$ are GGC: gamma, lognormal, positive α -stable, Pareto, Student t -distribution, Gumbel, and F -distribution. Of special applicability in financial modeling is the family of generalized inverse Gaussian distributions [4, 7].

A distribution μ with characteristic exponent Ψ is GGC if and only if there exists a positive Radon

measure U on $(0, \infty)$ such that

$$\Psi(u) = ia_0u - \int_0^\infty \log\left(1 + \frac{iu}{s}\right) U(ds) \quad (3)$$

with $\int_0^1 |\log x| U(dx) < \infty$ and $\int_1^\infty U(dx)/x < \infty$. The measure U_μ is called the *Thorin measure* of μ . So, the triplet of μ is $(a_0, 0, \nu_\mu)$ where the Lévy measure is concentrated on $(0, \infty)$ and such that $\nu_\mu(dx) = dx/x \int_0^\infty e^{-xs} U_\mu(ds)$. Moreover, any GGC is the law of a Wiener-gamma integral $\int_0^\infty h(u) d\gamma_u$, where $(\gamma_t; t \geq 0)$ is the standard gamma process with Lévy measure $\nu(dx) = e^{-x}(dx/x)$ and h is a Borel function $h: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ with $\int_0^\infty \log(1 + h(t))dt < \infty$. The function h is called the *Thorin function* of μ and is obtained as follows. Let $F_U(x) = \int_0^x U(dy)$ for $x \geq 0$ and let $F_U^{-1}(s)$ be the right continuous inverse of $F_U^{-1}(s)$ in the sense of composition of functions, that is $F_U^{-1}(s) = \inf\{t > 0; F_U(t) \geq s\}$ for $s \geq 0$. Then, $h(s) = 1/F_U^{-1}(s)$ for $s \geq 0$. For the positive α -stable distributions, $0 < \alpha < 1$, $h(s) = \{s\theta\Gamma(\alpha + 1)\}^{-1/\alpha}$ for a $\theta > 0$.

For distributions on $\mathbb{R}$, Thorin also introduced the class $T(\mathbb{R})$ of extended generalized gamma convolutions as the smallest class that contains the GGC and is closed under convolution, convergence in distribution, and reflection. These distributions are in the class L and are characterized by the alternative representation of their characteristic exponents

$$\begin{aligned} \Psi(u) = iua - \frac{1}{2}u^2\sigma^2 \\ - \int_{\mathbb{R}_+} \left[\ln\left(1 - \frac{iu}{x}\right) + \frac{iux}{1+x^2} \right] U(dx) \end{aligned} \quad (4)$$

where $a \in \mathbb{R}$, $\sigma^2 \geq 0$ and $U: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is a nondecreasing function with $U(0) = 0$, $\int_0^1 |\ln(x)| U(dx) < \infty$ and $\int_1^\infty x^{-2} U(dx) < \infty$. Several examples of Thorin distributions are given in [6, 9]. Any member of this class is the law of a stochastic integral $\int_0^\infty g^*(t) dZ_t$, where Z_t is a Lévy process with Z_1 satisfying the ID_{log} condition and g^* is the inverse of the incomplete gamma function $g(t) = \int_t^\infty u^{-1} e^{-u} du$ [2].

Type G Distributions

A random variable X is of *type G* if $X \stackrel{d}{=} \sqrt{V}N$, where N and V are independent random variables

with V being nonnegative ID and N having the standard normal distribution. Any type G distribution is ID and it is interpreted as the law of a random time changed Brownian motion B_V , where $\{B_t: t \geq 0\}$ is a Brownian motion independent of V . When we know the Lévy measure ρ of V , we can compute the Lévy density of X as $h(x) = (2\pi)^{-1/2} \int_{\mathbb{R}_+} s^{-1/2} e^{-\frac{1}{2s}x^2} \rho(ds)$ as well as its characteristic exponent

$$\Psi_X(u) = \int_{\mathbb{R}_+} \left(e^{-(1/2)u^2s} - 1 \right) \rho(ds) \quad (5)$$

Many classical distributions are of type G and SD: the gamma variance distribution, where V has a gamma distribution; the Student t , where V has the distribution of the reciprocal chi-square distribution and the symmetric α -stable distributions, $0 < \alpha < 2$; here V is a positive $\alpha/2$ -stable random variable, including the Cauchy distribution case $\alpha = 1$. Of special relevance in financial modeling are the normal inverse Gaussian, with V following the inverse Gaussian law [1], and the zero-mean symmetric generalized hyperbolic distributions, where V has the generalized inverse Gaussian law [5, 7]; all their moments are finite and they can accommodate heavy tails.

Tempered Stable Distributions

Tempered stable distributions (*see Tempered Stable Process*) are useful in mathematical finance as an attractive alternative to stable distributions, since they can have moments and heavy tails at the same time. Their corresponding Lévy and Ornstein–Uhlenbeck processes combines both the stable and Gaussian trends. An ID distribution on $\mathbb{R}$ is *tempered stable* if it is purely non-Gaussian and if its Lévy measure is of the form

$$\Pi(B) = \int_{\mathbb{R}} \int_0^\infty 1_B(sx) s^{-1-\alpha} g(s) ds \tau(dx) \quad (6)$$

where $0 < \alpha < 2$, g is a completely monotone function on $(0, \infty)$ and τ is a finite Borel measure on $\mathbb{R}$ such that τ has no atom at zero and $\int_{\mathbb{R}} |x|^\alpha \tau(dx) < \infty$. These distributions are in class L and constitute a proper subclass of the class of Thorin distributions $T(\mathbb{R})$.

References

- [1] Barndorff-Nielsen, O.E. (1998). Processes of normal inverse Gaussian type, *Finance and Stochastics* **2**, 41–68.
- [2] Barndorff-Nielsen, O.E., Maejima, M. & Sato, K. (2006). Some classes of multivariate infinitely divisible distributions admitting stochastic integral representations, *Bernoulli* **12**, 1–33.
- [3] Barndorff-Nielsen, O.E., Mikosch, T. & Resnick, S. (eds) (2001). *Lévy Processes—Theory and Applications*, Birkhäuser, Boston.
- [4] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics (with Discussion), *Journal of the Royal Statistical Society Series B* **63**, 167–241.
- [5] Bibby, B.M. & Sørensen, M. (2003). Hyperbolic distributions in finance, in *Handbook of Heavy Tailed Distributions in Finance*, S.T. Rachev, ed, Elsevier, Amsterdam.
- [6] Bondesson, L. (1992). *Generalized Gamma Convolutions and Related Classes of Distributions and Densities*, *Lecture Notes in Statistics*, Springer, Berlin, Vol. 76.
- [7] Eberlein, E. & Hammerstein, E.V. (2004). Generalized hyperbolic and inverse Gaussian distributions: limiting cases and approximation of processes, in *Seminar on Stochastic Analysis, Random Fields and Applications IV, Progress in Probability*, R.C. Dalang, M. Dozzi & F. Russo, eds, Birkhäuser, Vol. 58, pp. 221–264.
- [8] Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge.
- [9] Steutel, F.W. & Van Harn, K. (2003). *Infinite Divisibility of Probability Distributions on the Real Line*, Marcel-Dekker, New York.

Further Reading

- James, L.F., Roynette, B. & Yor, M. (2008). Generalized gamma convolutions, Dirichlet means, Thorin measures, with explicit examples, *Probability Surveys* **8**, 346–415.
- Rosinski, J. (2007). Tempering stable processes, *Stochastic Processes and Their Applications* **117**, 677–707.

Related Articles

Exponential Lévy Models; Heavy Tails; Lévy Processes; Ornstein–Uhlenbeck Processes; Tempered Stable Process; Time-changed Lévy Process.

VÍCTOR PÉREZ-ABREU

Ornstein–Uhlenbeck Processes

There are several reasons why Ornstein–Uhlenbeck processes are of practical interest in financial stochastic modeling. These continuous-time stochastic processes offer the possibility of capturing important distributional deviations from Gaussianity and for flexible modeling of dependence structures, while retaining analytic tractability.

An *Ornstein–Uhlenbeck* (OU) process is defined as the solution X_t of a Langevin-type stochastic differential equation (SDE) $dX_t = -\lambda X_t dt + dZ_t$, where $\lambda > 0$ and Z_t is a Lévy process (*see Lévy Processes*). The process is named after L. S. Ornstein and G. E. Uhlenbeck who, in 1930, considered the classical Langevin equation when Z is a Brownian motion, and hence X_t is a Gaussian process. Historical notes, references, and details are found in [6, 7] while modeling aspects are found in [1]. At the time of writing, new extensions and applications of OU processes are thriving, many of them motivated by financial modeling.

The Gaussian OU Process

Let $\{B_t : t \geq 0\}$ be a standard Brownian motion, σ a positive constant, and x_0 a real constant. The classical OU process

$$X_t = e^{-\lambda t} x_0 + \sigma \int_0^t e^{-\lambda(t-s)} dB_s, \quad t \geq 0 \quad (1)$$

is the solution of the classical Langevin equation $dX_t = -\lambda X_t dt + \sigma dB_t$, $X_0 = x_0$. It was originally proposed as a model for the velocity of a Brownian motion and it is the continuous-time analog of the discrete-time autoregressive process AR(1). In mathematical finance, OU is used for modeling of the dynamics of interest rates and volatilities of asset prices. The process X_t is a Gaussian process with (almost surely) continuous sample paths, mean function $E(X_t) = x_0 \sigma e^{-\lambda t}$, and covariance

$$\text{Cov}(X_t, X_s) = \frac{\sigma^2}{2\lambda} (e^{-\lambda|t-s|} - e^{-\lambda(t+s)}) \quad (2)$$

For $t = s$, we obtain $\text{var}(X_t) = \frac{\sigma^2}{2\lambda} (1 - e^{-2\lambda t})$. Let N be a zero-mean Gaussian random variable with variance $\frac{\sigma^2}{2\lambda}$, independent of the Brownian motion $\{B_t : t \geq 0\}$. The process $X_t = \sigma e^{-\lambda t} \int_0^t e^{\lambda s} dB_s + N$ is a stationary Gaussian process with $\text{Cov}(X_t, X_s) = \frac{\sigma^2}{2\lambda} e^{-\lambda|t-s|}$. Moreover, X_t is a Markov process with stationary transition probability

$$P_t(x, B) = \frac{\sqrt{\lambda}}{\sigma \sqrt{\pi(1 - e^{-2\lambda t})}} \times \int_B \exp \left\{ -\frac{\lambda}{\sigma^2} \frac{(y - x e^{-\lambda t})^2}{1 - e^{-2\lambda t}} \right\} dy \quad (3)$$

Non-Gaussian OU Processes

Let $\{Z_t : t \geq 0\}$ be a Lévy process (*see Lévy Processes*). A solution of the Langevin-type SDE $dX_t = -\lambda X_t dt + dZ_t$ is a stochastic process $\{X_t : t \geq 0\}$ with right-continuous and left-limit paths satisfying the equation

$$X_t = X_0 - \lambda \int_0^t X_s ds + Z_t, \quad t \geq 0 \quad (4)$$

When X_0 is independent of $\{Z_t : t \geq 0\}$, the unique (almost surely) solution is the OU process

$$X_t = e^{-\lambda t} X_0 + \int_0^t e^{-\lambda(t-s)} dZ_s, \quad t \geq 0 \quad (5)$$

We call Z_t the *background driving Lévy process* (BDLP). Of special relevance in financial modeling is the case when Z_t is a nonnegative increasing Lévy process (a subordinator) and X_0 is nonnegative. The corresponding OU process is positive, moves up entirely by jumps, and then tails off exponentially. Hence it can be used as a variance process.

Every OU process is a time-homogeneous Markov process starting from X_0 and its transition probability $P_t(x, dy)$ is infinitely divisible (*see Infinite Divisibility*) with characteristic function (*see Filtering*)

$$\int_{\mathbb{R}} e^{iuy} P_t(x, dy) = \exp \left[i x u e^{-\lambda t} + \int_0^t \Psi(e^{-\lambda s} u) ds \right] \quad (6)$$

where Ψ is the *characteristic exponent* of the Lévy process Z_t given by the Lévy–Khintchine representation

$$\begin{aligned} \Psi(u) = iau - \frac{1}{2}u^2\sigma^2 \\ + \int_{\mathbb{R}} (e^{iux} - 1 - iux1_{|x|\leq 1})\Pi(dx), \quad u \in \mathbb{R} \end{aligned} \quad (7)$$

where $\sigma^2 \geq 0$, $a \in \mathbb{R}$, and Π , the Lévy measure, is a positive measure on $\mathbb{R}$ with $\Pi(\{0\}) = 0$ and $\int_{\mathbb{R}} \min(1, |x|^2)\Pi(dx) < \infty$. For each $t > 0$, the probability distribution of Z_t has characteristic function $\varphi_t(u) = E[e^{iuX_t}] = \exp(t\Psi(u))$. When the Lévy measure is zero, Z_t is a Brownian motion with variance σ^2 and drift a .

The Integrated OU Process

A non-Gaussian OU process X_t has the same jump times of Z_t , as one sees from equation (4). However, X_t and Z_t cobreak in the sense that a linear combination of the two does not jump. We see this by considering the continuous *integrated OU process* $I_t^X = \int_0^t X_s ds$, which has two alternative representations

$$\begin{aligned} I_t^X = \lambda^{-1} \{X_0 - X_t + Z_t\} = \lambda^{-1}(1 - e^{-\lambda t})X_0 \\ + \lambda^{-1} \int_0^t \{1 - e^{-\lambda(t-s)}\} dZ_s \end{aligned} \quad (8)$$

In the Gaussian case, the process I_t^X is interpreted as the displacement of the Brownian particle. In financial applications, I_t^X is used to model integrated variance [1].

Stationary Distribution and the Stationary OU Process

An OU process has an asymptotic distribution μ when $t \rightarrow \infty$ if it does not have too many big jumps. This is achieved if Z_1 is $\text{ID}_{\log}$: $\int_{|x|>2} \ln|x| \Pi(dx) < \infty$, where Π is the Lévy measure of Z_1 . In this case, μ does not depend on X_0 and we call μ the *stationary distribution* of X_t . Moreover, μ is a self-decomposable (SD) distribution (and hence infinitely

divisible): for any $\theta \in (0, 1)$, there is a random variable ε_θ independent of X such that $X \stackrel{d}{=} \theta X + \varepsilon_\theta$. Conversely, for every SD distribution μ there exists a Lévy process Z_t with Z_1 being $\text{ID}_{\log}$ and such that μ is the stationary distribution of the OU process driven by Z_t .

The strictly stationary OU process is defined as

$$X_t = e^{-\lambda t} \int_{-\infty}^t e^{\lambda s} dZ_s, \quad t \in \mathbb{R} \quad (9)$$

where $\{Z_t : t \in \mathbb{R}\}$ is a Lévy process constructed as follows: let $\{Z_t^1 : t \geq 0\}$ be a Lévy process with characteristic exponent Ψ_1 and let $\{Z_t^2 : t \geq 0\}$ be a Lévy process with characteristic exponent $\Psi_2(u) = \Psi_1(-u)$ and independent of Z^1 . Then $Z_t = Z_t^1$ for $t \geq 0$ and $Z_t = Z_{-t}^2$ for $t < 0$. In this case, the law of X_t is SD and conversely, for any SD law μ there exists a BDLP Z_t such that equation (9) determines a stationary OU process with distribution μ . As a result, taking $X_0 = \int_{-\infty}^0 e^{\lambda s} dZ_s$, we can always consider (5) as a strictly stationary OU process with a prescribed SD distribution μ . It is an important example of a continuous-time moving average process.

Generalizations

The monographs [6, 7] contain a detailed study of multivariate OU process, while matrix extensions are considered in [2]. Another extension is the *generalized OU process*, which has arisen in several financial applications [4, 8]. It is defined as

$$X_t = e^{-\xi_t} X_0 + e^{-\xi_t} \int_0^t e^{-\xi_{s-}} d\eta_s, \quad t \geq 0 \quad (10)$$

where $\{(\xi_t, \eta_t) : t \geq 0\}$ is a bivariate Lévy process, independent of X_0 . This process is a homogeneous Markov process starting from X_0 , and, in general, the existence of the stationary solution depends on the convergence of integrals of exponentials of Lévy processes. For example, when ξ and η are independent, and if $\xi_t \rightarrow \infty$ and $V_\infty = \int_0^\infty e^{-\xi_{s-}} d\eta_s$ is defined and finite, then the law of V_∞ is the unique stationary solution of X_t . In the dependent case, the generalized OU process admits a stationary solution that does not degenerate to a constant process if and only if $V_\infty = \lim_{t \rightarrow \infty} \int_0^t e^{-\xi_{s-}} dL_s$ exists and is finite almost surely and does not

degenerate to constant random variable, and where L_t is the accompanying Lévy process $L_t = \eta_t + \sum_{0 < s \leq t} (e^{-\Delta \xi_s} - 1) \Delta \eta_s - tE(B_1^\xi, B_1^\eta)$, where $\Delta \xi_s = \xi_s - \xi_{s-}$, with B_1^ξ, B_1^η the Gaussian parts of ξ and η respectively [3, 5].

References

- [1] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics (with discussion), *The Journal of the Royal Statistical Society B* **63**, 167–241.
- [2] Barndorff-Nielsen, O.E. & Stelzer, R. (2007). Positive-definite matrix processes of finite variation, *Probability and Mathematical Statistics* **27**, 3–43.
- [3] Carmona, P., Petit, F. & Yor, M. (2001). Exponential functionals of Lévy processes, in *Lévy Processes. Theory and Applications*, O.E. Barndorff-Nielsen, T. Mikosch & S.I. Resnick, eds, Birkhäuser, pp. 41–55.
- [4] Klüppelberg, C., Linder, A. & Maller, R. (2006). Continuous time volatility modelling: COGARCH versus Ornstein-Uhlenbeck models, in *The Shiryaev Festschrift: From Stochastic Calculus to Mathematical Finance*, Y. Kabanov, R. Lipster & J. Stoyanov, eds, Springer, pp. 392–419.
- [5] Linder, A. & Maller, R. (2005). Lévy processes and the stationarity of generalised Ornstein-Uhlenbeck processes, *Stochastic Processes and Their Applications* **115**, 1701–1722.
- [6] Rocha-Arteaga, A. & Sato, K. (2003). *Topics in Infinitely Divisible Distributions and Lévy Processes*, *Aportaciones Matemáticas Investigación*, Mexican Mathematical Society, 17.
- [7] Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge.
- [8] Yor, M. (2001). *Exponential Functionals of Brownian Motion and Related Processes*, Springer, New York.

Related Articles

Infinite Divisibility; Lévy Processes; Stochastic Integrals.

VÍCTOR PÉREZ-ABREU

Fractional Brownian Motion

A fractional Brownian motion (fBm) is a self-similar Gaussian process, defined as follows:

Definition 1 Let $0 < H < 1$. The Gaussian stochastic process $\{B_H(t)\}_{t \geq 0}$ satisfying the following three properties

- (i) $B_H(0) = 0$
- (ii) $E[B_H(t)] = 0$ for all $t \geq 0$,
- (iii) for all $s, t \geq 0$,

$$E[B_H(t)B_H(s)] = \frac{1}{2} (|t|^{2H} - |t-s|^{2H} + |s|^{2H}) \quad (1)$$

is called the (standard) fBm with parameter H .

The fBm has been the subject of numerous investigations, in particular, in the context of long-range dependence (often referred to as long memory). fBm was first introduced in 1940 by Kolmogorov (see **Kolmogorov, Andrei Nikolaevich**) [11], but its main properties and its relevance in many fields of application such as economics, finance, turbulence, and telecommunications were first discussed in the seminal paper of Mandelbrot (see **Mandelbrot, Benoit**) and Van Ness [12].

For historical reasons, the parameter H is also referred to as the *Hurst coefficient*. In fact, in 1951, while he was investigating the flow of the river Nile, the British hydrologist H. E. Hurst [10] noticed that his measurements showed dependence properties and, in particular, long memory behavior in the sense that they seemed to require models, whose auto-correlation functions exhibit a power law decay at large timescales. This index of dependence H always takes values between 0 and 1 and indicates relatively long-range dependence if $H > 0.5$, for example, Hurst observed $H = 0.91$ in the case of Nile level data.

If $H = 0.5$, it is obvious from equation (1) that the increments of fBm are independent and $\{B_{0.5}(t)\}_{t \in \mathbb{R}} = \{B(t)\}_{t \in \mathbb{R}}$ is ordinary Brownian motion. Moreover, fBm has *stationary increments* which, for $H \neq 0.5$, are not independent.

One can define a parametric family of fBms in terms of the *stochastic Weyl integral* (see e.g. [16], chapter 7.2). In fact, for any $a, b \in \mathbb{R}$,

$$\{B_H(t)\}_{t \in \mathbb{R}} \stackrel{d}{=} \left\{ \int_{\mathbb{R}} \left\{ a [(t-s)_+^{H-\frac{1}{2}} - (-s)_+^{H-\frac{1}{2}}] + b [(t-s)_-^{H-\frac{1}{2}} - (-s)_-^{H-\frac{1}{2}}] \right\} dB(s) \right\}_{t \in \mathbb{R}} \quad (2)$$

where $u_+ = \max(u, 0)$, $u_- = \max(-u, 0)$, and $\{B(t)\}_{t \in \mathbb{R}}$ is a two-sided standard Brownian motion constructed by taking a Brownian motion B_1 and an independent copy B_2 and setting $B(t) = B_1(t)1_{\{t \geq 0\}} - B_2(-t)1_{\{t < 0\}}$.

If we choose $a = \sqrt{\Gamma(2H+1) \sin(\pi H)} / \Gamma(H+1/2)$ and $b = 0$ in equation (2) then $\{B_H(t)\}_{t \in \mathbb{R}}$ is an fBm satisfying equation (1).

fBm admits a *Volterra type representation* $B_H(t) = \int_0^t K_H(t, s) B(ds)$, where K_H is some square integrable kernel (see [13] or [1] for details).

Properties

Many properties of fBm, like self-similarity, are given by its fractional index H .

Definition 2 A real-valued stochastic process $\{X(t)\}_{t \in \mathbb{R}}$ is self-similar with index H if for all $c > 0$, $\{X(ct)\}_{t \in \mathbb{R}} \stackrel{d}{=} c^H \{X(t)\}_{t \in \mathbb{R}}$, where $\stackrel{d}{=}$ denotes equality in distribution.

Proposition 1 Fractional Brownian motion (fBm) is self-similar with index H . Moreover, fBm is the only self-similar Gaussian process with stationary increments.

Now, we consider the increments of fBm.

Definition 3 The stationary process $\{Y(t)\}_{t \in \mathbb{R}}$ given by

$$Y(t) = B_H(t) - B_H(t-1) \quad t \in \mathbb{R} \quad (3)$$

is called *fractional Gaussian noise*.

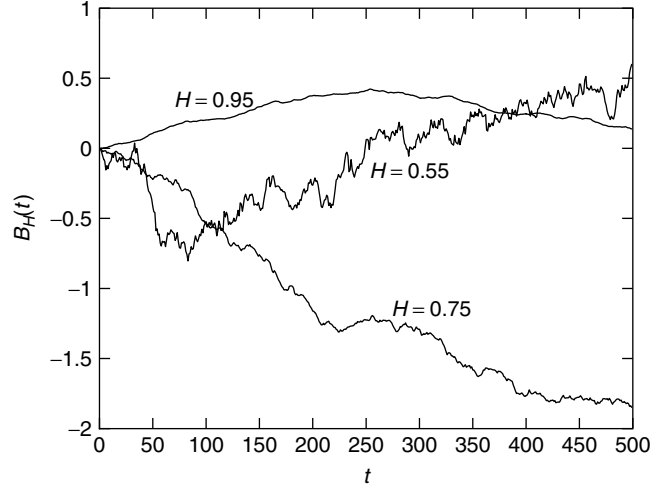


Figure 1 Various sample paths, each showing 500 points of fBm

For $n \in \mathbb{N}$, it follows by the stationarity of the increments of B_H , *such that*

$$|B_H(t) - B_H(s)| \leq c|t - s|^{H-\epsilon} \quad (6)$$

$$\begin{aligned} \rho_H(n) &:= \text{cov}(Y(k+n), Y(k)) \\ &= \frac{1}{2}(|n+1|^{2H} - 2|n|^{2H} - |n-1|^{2H}) \quad (4) \end{aligned}$$

Proposition 2

- (i) If $0 < H < 0.5$, ρ_H is negative and $\sum_{n=1}^{\infty} |\rho_H(n)| < \infty$.
- (ii) If $H = 0.5$, ρ_H equals 0, that is, the increments are independent.
- (iii) If $0.5 < H < 1$, ρ_H is positive, $\sum_{n=1}^{\infty} |\rho_H(n)| = \infty$,
and

$$\rho_H(n) \sim Cn^{2H-2}, \quad n \rightarrow \infty \quad (5)$$

Hence, for $0.5 < H < 1$ the increments of fBm are persistent or *long-range dependent*, whereas for $0 < H < 0.5$ they are said to be *antipersistent*.

Proposition 3 *The sample paths of fBm are continuous. In particular, for every $\tilde{H} < H$ there exists a modification of B_H whose sample paths are almost surely (a.s.) locally $\tilde{H}$ -Hölder continuous on $\mathbb{R}$, that is, for each trajectory, there exists a constant $c > 0$*

for any $\epsilon > 0$.

Figure 1 shows the sample paths of fBm for various values of the Hurst parameter H .

Proposition 4 *The sample paths of fBm are of finite p -variation for every $p > 1/H$ and of infinite p -variation if $p < 1/H$.*

Consequently, for $H < 0.5$ the quadratic variation is infinite. On the other hand, if $H > 0.5$ it is known that the quadratic variation of fBm is zero, whereas the total variation is infinite.

Corollary 1 *This shows that for $H \neq 1/2$, fBm cannot be a semimartingale.*

A proof of this well-known fact can be found in for example, [15] or [4].

However, since fBm is not a semimartingale one cannot use the Itô stochastic integral (*see Stochastic Integrals*) when considering integrals with respect to fBm. Recently, integration with respect to fBms has been studied extensively and various approaches have been made to define a stochastic integration theory for fBm (see e.g., [14] for a survey).

Applications in Finance

Many studies of financial time series point to long-range dependence (see **Long Range Dependence**), which indicates the potential usefulness of fBm in financial modeling (see [7] for a summary and references). One obstacle is that fBm is not a semimartingale (see **Semimartingale**), so the Itô integral cannot be used to define the gain of a self-financing portfolio as, for instance, in the Black–Scholes model (see **Black–Scholes Formula**). Various approaches have been developed for integrating fBm, some of which are as follows:

1. The *pathwise Riemann–Stieltjes fractional integral* defined by

$$\begin{aligned} & \int_0^T f(t) dB_H(t) \\ &= \lim_{|\pi| \rightarrow 0} \sum_{k=0}^{n-1} f(t_k) (B_H(t_{k+1}) - B_H(t_k)) \end{aligned} \quad (7)$$

where $\pi = \{t_k : 0 = t_0 < t_1 < \dots < t_n = T\}$ is a partition of the interval $[0, T]$ and f has bounded p -variation for some $p < 1/(1-H)$ a.s.

2. Under some regularity conditions on f , the *fractional Wick–Itô integral* has the form

$$\begin{aligned} & \int_0^T f(t) \delta B_H(t) \\ &= \lim_{|\pi| \rightarrow 0} \sum_{k=0}^{n-1} f(t_k) \diamond (B_H(t_{k+1}) - B_H(t_k)) \end{aligned} \quad (8)$$

where $\diamond$ represents the Wick product [18] and the convergence is the $L^2(\Omega)$ -convergence of random variables [2].

Whereas, the pathwise fractional integral mirrors a Stratonovich integral, the Wick–Itô–Skorohod calculus is similar to the Itô calculus, for example, integrals always have zero expectation.

The Wick–Itô integral was constructed by Duncan *et al.* [8] and later applied to finance by, for example,

Hu and Oksendal [9] in a fractional Black–Scholes pricing model in which the “gain” of a self-financing portfolio ϕ is replaced by $\int_0^T \phi(t) \delta S(t)$. However, results produced by this approach are controversial: indeed, for a piecewise constant strategy (represented by a simple predictable process) ϕ , this definition does not coincide with the capital gain of the portfolio, so the approach lacks economical interpretation [3]. An interesting study is [17], where the implications of different notions of integrals to the problem of arbitrage and self-financing condition in the fractional pricing model are considered.

An alternative is to use mixed Brownian motion, defined as the sum of a (regular) Brownian motion and an fBm with index H which, under some conditions on H , is a semimartingale [5]. Alternatively, Rogers [15] proposes to modify the behavior near zero of the kernel in equation (2) to obtain a semimartingale. In both the cases, one loses self-similarity, but conserves long-range dependence.

On the other hand, there is empirical evidence of long-range dependence in absolute returns [7], showing that it might be more interesting to use fractional processes as models of volatility rather than prices [6]. Fractional volatility processes are compatible with the semimartingale assumption for prices, so the technical obstacles discussed above do not necessarily arise when defining portfolio gain processes (see **Long Range Dependence**; **Multifractals**).

References

- [1] Baudoin, F. & Nualart, D. (2003). Equivalence of Volterra processes, *Stochastic Processes and their Applications* **107**, 327–350.
- [2] Bender, C. (2003). An Itô formula for generalized functionals of a fractional Brownian motion with arbitrary Hurst parameter, *Stochastic Processes and their Applications* **104**, 81–106.
- [3] Björk, T. & Hult, H. (2005). A note on Wick products and the fractional Black–Scholes model, *Finance and Stochastics* **9**, 197–209.
- [4] Cheridito, P. (2001). *Regularizing Fractional Brownian Motion with a View towards Stock Price Modelling*, PhD Dissertation, ETH Zurich.
- [5] Cheridito, P. (2003). Arbitrage in fractional Brownian motion models, *Finance and Stochastics* **7**, 533–553.
- [6] Comte, F. & Renault, E. (1998). Long memory in continuous time stochastic volatility models, *Mathematical Finance* **8**, 291–323.

- [7] Cont, R. (2005). Long range dependence in financial time series, in *Fractals in Engineering*, E. Lutton & J. Levy-Vehel, eds, Springer.
- [8] Duncan, T.E., Hu, Y. & Pasik-Duncan, B. (2000). Stochastic calculus for fractional Brownian motion I. Theory, *SIAM Journal of Control and Optimization* **28**, 582–612.
- [9] Hu, Y. & Oksendal, B. (2003). Fractional white noise calculus and applications to finance, *Infinite Dimensional Analysis, Quantum Probability and Related Topics* **6**, 1–32.
- [10] Hurst, H. (1951). Long term storage capacity of reservoirs, *Transactions of the American Society of Civil Engineers* **116**, 770–1299.
- [11] Kolmogorov, A.N. (1940). Wiener'sche Spiralen und einige andere interessante Kurven im Hilbertschen Raum, *Comptes Rendus (Doklady) Academic Sciences USSR (N.S.)* **26**, 115–118.
- [12] Mandelbrot, B.B. & Van Ness, J.W. (1968). Fractional Brownian motions, fractional noises and applications, *SIAM Review* **10**, 422–437.
- [13] Norros, I., Valkeila, E. & Virtamo, J. (1999). An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motion, *Bernoulli* **5**, 571–589.
- [14] Nualart, D. (2003). Stochastic calculus with respect to the fractional Brownian motion and applications, *Contemporary Mathematics* **336**, 3–39.
- [15] Rogers, L.C.G. (1997). Arbitrage with fractional Brownian motion, *Mathematical Finance* **7**, 95–105.
- [16] Samorodnitsky, G. & Taqqu, M. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, Chapman & Hall, New York.
- [17] Sottinen, T. & Valkeila, E. (2003). On arbitrage and replication in the fractional Black-Scholes pricing model, *Statistics and Decisions* **21**, 93–107.
- [18] Wick, G.-C. (1950). Evaluation of the collision matrix, *Physical Review* **80**, 268–272.

Further Reading

- Doukhan, P., Oppenheim, G. & Taqqu, M.S. (2003). *Theory and Applications of Long-Range Dependence*, Birkhäuser, Boston.
- Lin, S.J. (1995). Stochastic analysis of fractional Brownian motion, *Stochastics and Stochastics Reports* **55**, 121–140.

Related Articles

Long Range Dependence; Mandelbrot, Benoit; Multifractals; Semimartingale; Stylized Properties of Asset Returns.

TINA M. MARQUARDT

Lévy Processes

A *Lévy process* is a continuous-time stochastic process with independent and stationary increments. Lévy processes may be thought of as the continuous-time analogs of random walks. Mathematically, a Lévy process can be defined as follows.

Definition 1 An $\mathbb{R}^d$ -valued stochastic process $X = \{X_t : t \geq 0\}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be a *Lévy process* if it possesses the following properties:

1. The paths of X are $\mathbb{P}$ almost surely right continuous with left limits.
2. $\mathbb{P}(X_0 = 0) = 1$.
3. For $0 \leq s \leq t$, $X_t - X_s$ is equal in distribution to X_{t-s} .
4. For $0 \leq s \leq t$, $X_t - X_s$ is independent of $\{X_u : u \leq s\}$.

Historically, Lévy processes have always played a central role in the study of stochastic processes with some of the earliest work dating back to the early 1900s. The reason for this is that, mathematically, they represent an extremely robust class of processes, which exhibit many of the interesting phenomena that appear in, for example, the theories of stochastic and potential analysis. Moreover, this in turn, together with their elementary definition, has made Lévy processes an extremely attractive class of processes for modeling in a wide variety of physical, biological, engineering, and economical scenarios. Indeed, the first appearance of particular examples of Lévy processes can be found in the foundational works of Bachelier [1, 2], concerning the use of Brownian motion, within the context of financial mathematics, and Lundberg [9], concerning the use of Poisson processes within the context of insurance mathematics.

The term *Lévy process* honors the work of the French mathematician Paul Lévy who, although not alone in his contribution, played an instrumental role in bringing together an understanding and characterization of processes with stationary and independent increments. In earlier literature, Lévy processes have been dealt with under various names. In the 1940s, Lévy himself referred to them as a subclass of *processus additifs* (additive processes), that is, processes

with independent increments. For the most part, however, research literature through the 1960s and 1970s refers to Lévy processes simply as *processes with stationary and independent increments*. One sees a change in language through the 1980s and by the 1990s the use of the term *Lévy process* had become standard.

Judging by the volume of published mathematical research articles, the theory of Lévy processes can be said to have experienced a steady flow of interest from the time of the foundational works, for example, of Lévy [8], Kolmogorov [7], Khintchine [6], and Itô [5]. However, it was arguably in the 1990s that a surge of interest in this field of research occurred, drastically accelerating the breadth and depth of understanding and application of the theory of Lévy processes. While there are many who made prolific contributions during this period, as well as thereafter, the general progression of this field of mathematics was enormously encouraged by the monographs of Bertoin [3] and Sato [10]. It was also the growing research momentum in the field of financial and insurance mathematics that stimulated a great deal of the interest in Lévy processes in recent times, thus entwining the modern theory of Lévy processes ever more with its historical roots.

Lévy Processes and Infinite Divisibility

The properties of stationary and independent increments imply that a Lévy process is a Markov process. One may show in addition that Lévy processes are strong Markov processes. From Definition 1 alone it is otherwise difficult to understand the richness of the class of Lévy processes. To get a better impression in this respect, it is necessary to introduce the notion of an *infinitely divisible* distribution. Generally, an $\mathbb{R}^d$ -valued random variable Θ has an infinitely divisible distribution if for each $n = 1, 2, \dots$ there exists a sequence of i.i.d. random variables $\Theta_{1,n}, \dots, \Theta_{n,n}$ such that

$$\Theta \stackrel{d}{=} \Theta_{1,n} + \dots + \Theta_{n,n} \quad (1)$$

where $\stackrel{d}{=}$ is equality in distribution. Alternatively, this relation can be expressed in terms of characteristic exponents. That is to say, if Θ has characteristic exponent $\Psi(u) := -\log \mathbb{E}(e^{iu \cdot \Theta})$, then Θ is infinitely divisible if and only if for all $n \geq 1$ there exists a characteristic exponent of a probability distribution, say Ψ_n , such that $\Psi(u) = n\Psi_n(u)$ for all $u \in \mathbb{R}^d$.

It turns out that Θ has an infinitely divisible distribution if and only if there exists a triple (a, Σ, Π) , where $a \in \mathbb{R}^d$, Σ is a $d \times d$ matrix whose eigenvalues are all nonnegative, and Π is a measure concentrated on $\mathbb{R}^d \setminus \{0\}$ satisfying $\int_{\mathbb{R}^d} (1 \wedge |x|^2) \Pi(dx) < \infty$, such that

$$\Psi(u) = ia \cdot u + \frac{1}{2} u \cdot \Sigma u + \int_{\mathbb{R}^d} (1 - e^{iu \cdot x} + iu \cdot x \mathbf{1}_{(|x| < 1)}) \Pi(dx) \quad (2)$$

for every $\theta \in \mathbb{R}^d$. Here, we use the notation $u \cdot x$ for the Euclidian inner product and $|x|$ for Euclidian distance. The measure Π is called the *Lévy (characteristic) measure* and it is unique. The identity in equation (2) is known as the *Lévy–Khintchine formula*.

The link between a Lévy processes and infinitely divisible distributions becomes clear when one notes that for each $t > 0$ and any $n = 1, 2, \dots$,

$$X_t = X_{t/n} + (X_{2t/n} - X_{t/n}) + \dots + (X_t - X_{(n-1)t/n}) \quad (3)$$

As a result of the fact that X has stationary independent Increments, it follows that X_t is infinitely divisible.

It can be deduced from the above observation that any Lévy process has the property that for all $t \geq 0$

$$\mathbb{E}(e^{iu \cdot X_t}) = e^{-t\Psi(u)} \quad (4)$$

where $\Psi(\theta) := \Psi_1(\theta)$ is the characteristic exponent of X_1 , which has an infinitely divisible distribution. The converse of this statement is also true, thus constituting the Lévy–Khintchine formula for Lévy processes.

Theorem 1 (Lévy–Khintchine formula for Lévy processes). *$a \in \mathbb{R}^d$, Σ is a $d \times d$ matrix whose eigenvalues are all nonnegative, and Π is a measure concentrated on $\mathbb{R}^d \setminus \{0\}$ satisfying $\int_{\mathbb{R}^d} (1 \wedge |x|^2) \Pi(dx) < \infty$. Then there exists a Lévy process having characteristic exponent*

$$\Psi(u) = ia \cdot u + \frac{1}{2} u \cdot \Sigma u + \int_{\mathbb{R}^d} (1 - e^{iu \cdot x} + iu \cdot x \mathbf{1}_{(|x| < 1)}) \Pi(dx) \quad (5)$$

Two fundamental examples of Lévy processes, which are shown in the next section to form the “building blocks” of all the other Lévy processes, are Brownian motion and compound Poisson processes. A Brownian motion is the Lévy process associated with the characteristic exponent

$$\Psi(u) = \frac{1}{2} u \cdot \Sigma u \quad (6)$$

and therefore has increments over time periods of length t , which are Gaussian distributed with covariance matrix Σt . It can be shown that, up to the addition of a linear drift, Brownian motions are the only Lévy processes that have continuous paths.

A compound Poisson process is the Lévy process associated with the characteristic exponent:

$$\Psi(u) = \int_{\mathbb{R}^d} (1 - e^{iu \cdot x}) \lambda F(dx) \quad (7)$$

where $\lambda > 0$ and F is a probability distribution. Such processes may be described pathwise by the piecewise linear process:

$$\sum_{i=1}^{N_t} \xi_i, \quad t \geq 0 \quad (8)$$

where $\{\xi_i : i \geq 1\}$ are a sequence of i.i.d. random variables with common distribution F , and $\{N_t : t \geq 0\}$ is a Poisson process with rate λ ; the latter is the process with initial value zero and with unit increments whose interarrival times are independent and exponentially distributed with parameter λ .

It is a straightforward exercise to show that the sum of any finite number of independent Lévy processes is also a Lévy process. Under some circumstances, one may show that a countably infinite sum of Lévy processes also converges in an appropriate sense to a Lévy process. This idea forms the basis of the Lévy–Itô decomposition, discussed in the next section, where, as alluded to above, the Lévy processes that are summed together are either a Brownian motion with drift or a compound Poisson process with drift.

The Lévy–Itô Decomposition

Hidden in the Lévy–Khintchine formula is a representation of the path of a given Lévy process. Every

Lévy process may always be written as the independent sum of up to a countably infinite number of other Lévy processes, at most one of which will be a linear Brownian motion and the remaining processes will be compound Poisson processes with drift.

Let Ψ be the characteristic exponent of some infinitely divisible distribution with associated triple (a, Σ, Π) . The necessary assumption that $\int_{\mathbb{R}^d} (1 \wedge |x|^2) \Pi(dx) < \infty$ implies that $\Pi(A) < \infty$ for all Borel A such that 0 is in the interior of A^c and, in particular, that $\Pi(\{x : |x| \geq 1\}) \in [0, \infty)$. With this in mind, it is not difficult to see that, after some simple reorganization, for $u \in \mathbb{R}^d$, the Lévy–Khintchine formula can be written in the form

$$\begin{aligned} \Psi(\theta) = & \left\{ iu \cdot a + \frac{1}{2} u \cdot \Sigma u \right\} \\ & + \left\{ \lambda_0 \int_{|x| \geq 1} (1 - e^{iu \cdot x}) F_0(dx) \right\} \\ & + \sum_{n \geq 1} \left\{ \lambda_n \int_{2^{-n} \leq |x| < 2^{-(n-1)}} (1 - e^{iu \cdot x}) F_n(dx) \right. \\ & \left. + i \lambda_n u \cdot \left(\int_{2^{-n} \leq |x| < 2^{-(n-1)}} x F_n(dx) \right) \right\} \quad (9) \end{aligned}$$

where $\lambda_0 = \Pi(\{x : |x| \geq 1\})$, $F_0(dx) = \Pi(dx)/\lambda_0$, and for $n = 1, 2, 3, \dots$, $\lambda_n = \Pi(\{x : 2^{-n} \leq |x| < 2^{-(n-1)}\})$ and $F_n(dx) = \Pi(dx)/\lambda_n$ (with the understanding that the n th integral is absent if $\lambda_n = 0$). This decomposition suggests that the Lévy process $X = \{X_t : t \geq 0\}$ associated with Ψ may be written as the independent sum:

$$X_t = Y_t + X_t^{(0)} + \lim_{k \uparrow \infty} \sum_{n=1}^k X_t^{(n)}, \quad t \geq 0 \quad (10)$$

where

$$Y_t = B_t^\Sigma - at, \quad t \geq 0 \quad (11)$$

with $\{B_t^\Sigma : t \geq 0\}$ a d -dimensional Brownian motion with covariance matrix Σ ,

$$X_t^{(0)} = \sum_{i=1}^{N_t^{(0)}} \xi_i^{(0)}, \quad t \geq 0 \quad (12)$$

with $\{N_t^{(0)} : t \geq 0\}$ as a Poisson process with rate λ_0 and $\{\xi_i^{(0)} : i \geq 1\}$ are independent and identically

distributed with common distribution $F_0(dx)$ concentrated on $\{x : |x| \geq 1\}$ and for $n = 1, 2, 3, \dots$

$$X_t^{(n)} = \sum_{i=1}^{N_t^{(n)}} \xi_i^{(n)} - \lambda_n t \int_{2^{-n} \leq |x| < 2^{-(n-1)}} x F_n(dx), \quad t \geq 0 \quad (13)$$

with $\{N_t^{(n)} : t \geq 0\}$ as a Poisson process with rate λ_n and $\{\xi_i^{(n)} : i \geq 1\}$ are independent and identically distributed with common distribution $F_n(dx)$ concentrated on $\{x : 2^{-n} \leq |x| < 2^{-(n-1)}\}$. The limit in equation (10) needs to be understood in the appropriate context, however.

It is a straightforward exercise to deduce that $X_t^{(n)}$ is a square integrable martingale on account of the fact that it is a centered compound Poisson process together with the fact that x^2 is integrable in the neighborhood of the origin against the measure Π . It is not difficult to see that $\sum_{n=1}^k X_t^{(n)}$ is also a square integrable martingale. The convergence of $\sum_{n=1}^k X_t^{(n)}$ as $k \uparrow \infty$ can happen in one of the two ways. The two quantities

$$\begin{aligned} \lim_{k \uparrow \infty} \sum_{n=1}^k \sum_{i=1}^{N_t^{(n)}} |\xi_i^{(n)}| \quad \text{and} \\ \lim_{k \uparrow \infty} \sum_{n=1}^k \int_{2^{-n} \leq |x| < 2^{-(n-1)}} |x| \lambda_n F_n(dx) \quad (14) \end{aligned}$$

are either simultaneously finite or infinite (for all $t > 0$), where the random limit is understood in the almost sure sense. When both are finite, that is to say, when $\int_{|x| < 1} |x| \Pi(dx) < \infty$, then $\sum_{n=1}^\infty X_t^{(n)}$ is well defined as the difference of a stochastic processes with jumps and a linear drift. Conversely when $\int_{|x| < 1} |x| \Pi(dx) = \infty$, it can be shown that, thanks to the assumption, $\int_{|x| < 1} |x|^2 \Pi(dx) < \infty$, $\sum_{n=1}^k X_t^{(n)}$ converges uniformly over finite time horizons in the L^2 norm as $k \uparrow \infty$. In that case, the two exploding limits in equation (14) compensate each other in the right way for their difference to converge in the prescribed sense.

Either way, the properties of stationary and independent increments and almost surely right continuous paths with left limits that belong to $\sum_{n=1}^k X_t^{(n)}$ as a finite sum of Lévy processes are also inherited by the limiting process as $k \uparrow \infty$. It is also the case that the limiting Lévy process is also a square integrable

martingale just as the elements of the approximating sequence are.

Path Variation

Consider any function $f : [0, \infty) \rightarrow \mathbb{R}^d$. Given any partition $\mathcal{P} = \{a = t_0 < t_1 < \dots < t_n = b\}$ of the bounded interval $[a, b]$, define the variation of f over $[a, b]$ with partition $\mathcal{P}$ by

$$V_{\mathcal{P}}(f, [a, b]) = \sum_{i=1}^n |f(t_i) - f(t_{i-1})| \quad (15)$$

The function f is said to be of bounded variation over $[a, b]$ if

$$V(f, [a, b]) := \sup_{\mathcal{P}} V_{\mathcal{P}}(f, [a, b]) < \infty \quad (16)$$

where the supremum is taken over all partitions of $[a, b]$. Moreover, f is said to be of *bounded variation* if the above inequality is valid for all bounded intervals $[a, b]$. If $V(f, [a, b]) = \infty$ for all bounded intervals $[a, b]$, then f is said to be of *unbounded variation*.

For any given stochastic process $X = \{X_t : t \geq 0\}$, we may adopt these notions in the almost sure sense. So, for example, the statement “ X is a process of bounded variation” (or “has paths of bounded variation”) simply means that as a random mapping, $X : [0, \infty) \rightarrow \mathbb{R}^d$ is of bounded variation almost surely.

In the case that X is a Lévy process, the Lévy–Itô decomposition also gives the opportunity to establish a precise characterization of the path variation of a Lévy process. Since any Lévy process may be written as the independent sum as in equation (10) and any d -dimension Brownian motion is known to have paths of unbounded variation, it follows that any Lévy process for which $\Sigma \neq 0$ has unbounded variation. In the case that $\Sigma = 0$, since the paths of the component $X^{(0)}$ in equation (10) are independent and clearly of bounded variation (they are piecewise linear), the path variation of X is characterized by the way in which the component $\sum_{n=1}^k X_t^{(n)}$ converges. In the case that

$$\int_{|x|<1} |x| \Pi(dx) < \infty \quad (17)$$

the Lévy process X will thus be of bounded variation and otherwise, when the above integral is infinite, the paths are of unbounded variation.

In the case that $d = 1$, as an extreme case of a Lévy process with bounded variation, it is possible that the process X has nondecreasing paths, in which case it is called a *subordinator*. As is apparent from the Lévy–Itô decomposition (9), this will necessarily occur when $\Pi(-\infty, 0) = 0$,

$$\int_{(0,1)} x \Pi(dx) < \infty \quad (18)$$

and $\Sigma = 0$. In that case, reconsidering the decomposition (10), one may identify

$$X_t = \left(-a - \int_{(0,1)} x \Pi(dx) \right) t + \lim_{k \uparrow \infty} \sum_{n=1}^k \sum_{i=1}^{N_t^{(n)}} \xi_i^{(n)} \quad (19)$$

On account of the assumption $\Pi(-\infty, 0) = 0$, all the jumps $\xi_i^{(n)}$ are nonnegative. Hence, it is also a necessary condition that

$$-a - \int_{(0,1)} x \Pi(dx) \geq 0 \quad (20)$$

for X to have nondecreasing paths. These necessary conditions are also sufficient.

Lévy Processes as Semimartingales

Recall that a semimartingale with respect to a given filtration $\mathbb{F} := \{\mathcal{F}_t : t \geq 0\}$ is defined as the sum of an $\mathbb{F}$ -local martingale and an $\mathbb{F}$ -adapted process of bounded variation. The importance of semimartingales is that they form a natural class of stochastic processes with respect to which one may construct a stochastic integral and thereafter perform calculus. Moreover, the theory of stochastic calculus plays a significant role in mathematical finance as it can be used as a key ingredient in justifying the pricing and hedging of derivatives in markets where risky assets are modeled as positive semimartingales.

A popular choice of model for risky assets in recent years has been the exponential of a Lévy process (see **Exponential Lévy Models**). Lévy processes have also been used as building blocks in more complex stochastic models for prices, such as stochastic

volatility models with jumps (*see* **Barndorff-Nielsen and Shephard (BNS) Models**) and time-changed Lévy models (*see* **Time-changed Lévy Process**). The monograph of Cont and Tankov [4] gives an extensive exposition on these types of models. Thanks to Itô's formula for semimartingales, the exponential of a Lévy process is a semimartingale when it can be shown that a Lévy process is a semimartingale. However, reconsidering equation (10) and recalling that B^Σ and $\lim_{k \uparrow \infty} \sum_{n=1}^k X^{(n)}$ are martingales and that $X^{(0)} - a \cdot$ is an adapted process with bounded variation paths, it follows immediately that any Lévy process is a semimartingale.

References

- [1] Bachelier, L. (1900). Théorie de la spéculation, *Annales Scientifiques de l'École Normale Supérieure* **17**, 21–86.
- [2] Bachelier, L. (1901). Théorie mathématique du jeu, *Annales Scientifiques de l'École Normale Supérieure* **18**, 143–210.
- [3] Bertoin, J. (1996). *Lévy Processes*, Cambridge University Press, Cambridge.
- [4] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, *Financial Mathematics Series*, Chapman & Hall/CRC.
- [5] Itô, K. (1942). On stochastic processes. I. (Infinitely divisible laws of probability), *Japanese Journal of Mathematics* **18**, 261–301.
- [6] Khintchine, A. (1937). A new derivation of one formula by Levy P., *Bulletin of Moscow State University* **I**(1), 1–5.
- [7] Kolmogorov, N.A. (1932). Sulla forma generale di un processo stocastico omogeneo (un problema di B. de Finetti), *Atti Reale Accademia Nazionale dei Lincei Rend* **15**, 805–808.
- [8] Lévy, P. (1934). Sur les intégrales dont les éléments sont des variables aléatoires indépendantes, *Annali della Scuola Normale Superiore di Pisa* **3–4**, 217–218, 337–366.
- [9] Lundberg, F. (1903). *Approximerad framställning av sannolikhetsfunktionen, Återförsäkring av kollektiv-risker*, Akademisk Afhandling Almqvist och Wiksell, Uppsala.
- [10] Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge.

Related Articles

Generalized Hyperbolic Models; Infinite Divisibility; Jump Processes; Lévy Copulas; Normal Inverse Gaussian Model; Poisson Process; Stochastic Exponential; Tempered Stable Process; Time-changed Lévy Process; Variance-gamma Model.

ANDREAS E. KYPRIANOU

Wiener–Hopf Decomposition

A fundamental part of the theory of random walks and Lévy processes is a set of conclusions, which, in modern times, are loosely referred to as *the Wiener–Hopf factorization*. Historically, the identities around which the Wiener–Hopf factorization is centered are the culmination of a number of works that include [2–4, 6–8, 14–17], and many others; although the analytical roots of the so-called Wiener–Hopf method go much further back than these probabilistic references; see, for example, [9, 13]. The importance of the Wiener–Hopf factorization for either a random walk or a Lévy process is that it characterizes the range of the running maximum of the process as well as the times at which new maxima are attained. We deal with the Wiener–Hopf factorization for random walks before moving to the case of Lévy processes. The discussion very closely follows the ideas of [6, 7]. Indeed, for the case of random walks, we shall not deter from providing proofs as their penetrating and yet elementary nature reveals a simple path decomposition that is arguably more fundamental than the Wiener–Hopf factorization itself. The Wiener–Hopf factorization for Lévy processes is essentially a technical variant of the case for random walks and we only state it without proof.

Random Walks and Infinite Divisibility

Suppose that $\{\xi_i : i = 1, 2, \dots\}$ are a sequence of $\mathbb{R}$ -valued independent and identically distributed (i.i.d.) random variables defined on the common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with common distribution function F . Let

$$S_0 = 0 \text{ and } S_n = \sum_{i=1}^n \xi_i \quad (1)$$

The process $S = \{S_n : n \geq 0\}$ is called a (*real valued*) random walk. For convenience, we make a number of assumptions on F . First,

$$\min\{F(0, \infty), F(-\infty, 0)\} > 0 \quad (2)$$

meaning that the random walk may experience both positive and negative jumps, and second, F has no atoms. In the prevailing analysis, we repeatedly refer

to general and specific classes of infinitely divisible random variables (see **Infinite Divisibility**). An $\mathbb{R}^d$ -valued random variable X is infinitely divisible if for each $n = 1, 2, 3, \dots$

$$X \stackrel{d}{=} X_{(1,n)} + \dots + X_{(n,n)} \quad (3)$$

where $\{X_{(i,n)} : i = 1, \dots, n\}$ are i.i.d. distributed and the equality is in the distribution. In other words, if μ is the characteristic function of X , then for each $n = 1, 2, 3, \dots$ we have $\mu = (\mu_n)^n$, where μ_n is the characteristic function of some $\mathbb{R}^d$ -valued random variable.

In general, if X is any $\mathbb{R}^d$ -valued random variable that is also infinitely divisible, then for each $\theta \in \mathbb{R}^d$, $E(e^{i\theta \cdot X}) = e^{-\Psi(\theta)}$ where

$$\begin{aligned} \Psi(\theta) = & ia \cdot \theta + \frac{1}{2} Q(\theta) \\ & + \int_{\mathbb{R}^d} (1 - e^{i\theta \cdot x} + i\theta \cdot x \mathbf{1}_{(|x| < 1)}) \Pi(dx) \end{aligned} \quad (4)$$

where $a \in \mathbb{R}^d$, Q is a positive semidefinite quadratic form on $\mathbb{R}^d$ and Π is a measure supported in $\mathbb{R}^d \setminus \{0\}$ such that

$$\int_{\mathbb{R}^d} 1 \wedge |x|^2 \Pi(dx) < \infty \quad (5)$$

Here, $|\cdot|$ is Euclidean distance and, for $a, b \in \mathbb{R}^d$, $a \cdot b$ is the usual Euclidean inner product.

A special example of an infinitely divisible distribution is the geometric distribution. The symbol Γ_p always denotes a geometric distribution with parameter $p \in (0, 1)$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$. In particular,

$$P(\Gamma_p = k) = pq^k, \quad k = 0, 1, 2, \dots \quad (6)$$

where $q = 1 - p$. The geometric distribution has the following properties that are worth recalling for the forthcoming discussion. First,

$$P(\Gamma_p \geq k) = q^k \quad (7)$$

and, second, the lack-of-memory property:

$$\begin{aligned} P(\Gamma_p \geq n + m | \Gamma_p \geq m) &= P(\Gamma_p \geq n), \\ n, m &= 0, 1, 2, \dots \end{aligned} \quad (8)$$

A more general class of infinitely divisible distributions than the latter, which will shortly be of use,

are those that may be expressed as the distribution of a random walk sampled at an independent and geometrically distributed time; $S_{\Gamma_p} = \sum_{i=1}^{\Gamma_p} \xi_i$. (Note, we interpret $\sum_{i=1}^0$ as the empty sum). To justify the previous claim, a straightforward computation shows that for each $n = 1, 2, 3, \dots$

$$\begin{aligned} \mathbb{E}(e^{i\theta S_{\Gamma_p}}) &= \left(\left(\frac{p}{1 - q \mathbb{E}(e^{i\theta \xi_1})} \right)^{\frac{1}{n}} \right)^n \\ &= \mathbb{E}(e^{i\theta S_{\Lambda_{1/n,p}}}) \end{aligned} \quad (9)$$

where $\Lambda_{1/n,p}$ is a negative binomial random variable with parameters $1/n$ and p , which is independent of S . The latter has distribution mass function

$$\mathbb{P}(\Lambda_{1/n,p} = k) = \frac{1}{k!} \frac{\Gamma(k+1/n)}{\Gamma(1/n)} p^{1/n} q^k \quad (10)$$

for $k = 0, 1, 2, \dots$

Wiener–Hopf Factorization for Random Walks

We now turn our attention to the Wiener–Hopf factorization. Fix $0 < p < 1$ and define

$$G = \inf \left\{ k = 0, 1, \dots, \Gamma_p : S_k = \max_{j=0,1,\dots,\Gamma_p} S_j \right\} \quad (11)$$

where Γ_p is a geometrically distributed random variable with parameter p , which is independent of the random walk S , that is, G is the first visit of S to its maximum over the time period $\{0, 1, \dots, \Gamma_p\}$. Now define

$$N = \inf\{n > 0 : S_n > 0\} \quad (12)$$

In other words, the first visit of S to $(0, \infty)$ after time 0.

Theorem 1 (Wiener–Hopf Factorization for Random Walks) Assume all of the notation and conventions above.

- (i) (G, S_G) is independent of $(\Gamma_p - G, S_{\Gamma_p} - S_G)$ and both pairs are infinitely divisible.

- (ii) For $0 < s \leq 1$ and $\theta \in \mathbb{R}$

$$\begin{aligned} E(s^G e^{i\theta S_G}) \\ = \exp \left\{ - \int_{(0,\infty)} \sum_{n=1}^{\infty} (1-s^n e^{i\theta x}) q^n \frac{1}{n} F^{*n}(dx) \right\} \end{aligned} \quad (13)$$

- (iii) For $0 < s \leq 1$ and $\theta \in \mathbb{R}$

$$\begin{aligned} E(s^N e^{i\theta S_N}) \\ = 1 - \exp \left\{ - \int_{(0,\infty)} \sum_{n=1}^{\infty} s^n e^{i\theta x} \frac{1}{n} F^{*n}(dx) \right\} \end{aligned} \quad (14)$$

Note that the third part of the Wiener–Hopf factorization characterizes what is known as the *ladder height process* of the random walk S . The latter is the bivariate random walk $(T, H) := \{(T_n, H_n) : n = 0, 1, 2, \dots\}$ where $(T_0, H_0) = (0, 0)$, and otherwise for $n = 1, 2, 3, \dots$,

$$T_n = \begin{cases} \min \{k \geq 1 : S_{T_{n-1}+k} > H_{n-1}\} & \text{if } T_{n-1} < \infty \\ \infty & \text{if } T_{n-1} = \infty \end{cases}$$

and

$$H_n = \begin{cases} S_{T_n} & \text{if } T_n < \infty \\ \infty & \text{if } T_n = \infty \end{cases} \quad (15)$$

That is to say, the process (T, H) , until becoming infinite in value, represents the times and positions of the running maxima of S , the so-called ladder times and ladder heights. It is not difficult to see that T_n is a stopping time for each $n = 0, 1, 2, \dots$ and hence thanks to the i.i.d. increments of S , the increments of (T, H) are i.i.d. with the same law as the pair (N, S_N) .

Proof (i) The path of the random walk may be broken into $\nu \in \{0, 1, 2, \dots\}$ finite (or completed) excursions from the maximum followed by an additional excursion, which straddles the random time Γ_p . Here, we understand the use of the word straddle to mean that if ℓ is the index of the left end point of the straddling excursion then $\ell \leq \Gamma_p$. By the strong Markov property for random walks and lack of memory, the completed excursions must have the same law, namely, that of a random walk sampled on the time points $\{1, 2, \dots, N\}$ conditioned on the

event that $\{N \leq \Gamma_p\}$ and hence ν is geometrically distributed with parameter $1 - P(N \leq \Gamma_p)$. Mathematically, we express

$$(G, S_G) = \sum_{i=1}^{\nu} (N^{(i)}, H^{(i)}) \quad (16)$$

where the pairs $\{(N^{(i)}, H^{(i)}) : i = 1, 2, \dots\}$ are independent having the same distribution as (N, S_N) conditioned on $\{N \leq \Gamma_p\}$. Note also that G is the sum of the lengths of the latter conditioned excursions and S_G is the sum of the respective increment of the terminal value over the initial value of each excursion. In other words, (G, S_G) is the component-wise sum of ν independent copies of (N, S_N) (with $(G, S_G) = (0, 0)$ if $\nu = 0$). Infinite divisibility follows as a consequence of the fact that (G, S_G) is a geometric sum of i.i.d. random variables. The independence of (G, S_G) and $(\Gamma_p - G, S_{\Gamma_p} - S_G)$ is immediate from the decomposition described above.

Feller's classic duality lemma (cf [3]) for random walks says that for any $n = 0, 1, 2, \dots$ (which may later be randomized with an independent geometric distribution), the independence and common distribution of increments implies that $\{S_{n-k} - S_n : k = 0, 1, \dots, n\}$ has the same law as $\{-S_k : k = 0, 1, \dots, n\}$. In the current context, the duality lemma also implies that the pair $(\Gamma_p - G, S_{\Gamma_p} - S_G)$ is equal in distribution to (D, S_D) where

$$D := \sup \left\{ k = 0, 1, \dots, \Gamma_p : S_k = \min_{j=0,1,\dots,\Gamma_p} S_j \right\} \quad (17)$$

(ii) Note that, as a geometric sum of i.i.d. random variables, the pair (Γ_p, S_{Γ_p}) is infinitely divisible for $s \in (0, 1)$ and $\theta \in \mathbb{R}$, let $q = 1 - p$ and also that, on one hand,

$$\begin{aligned} E(s^{\Gamma_p} e^{i\theta S_{\Gamma_p}}) &= E\left(E(s e^{i\theta S_1})^{\Gamma_p}\right) \\ &= \sum_{k \geq 0} p (q s E(e^{i\theta S_1}))^k \\ &= \frac{p}{1 - q s E(e^{i\theta S_1})} \end{aligned} \quad (18)$$

and, on the other hand, with the help of Fubini's Theorem,

$$\begin{aligned} &\exp \left\{ - \int_{\mathbb{R}} \sum_{n=1}^{\infty} (1 - s^n e^{i\theta x}) q^n \frac{1}{n} F^{*n}(dx) \right\} \\ &= \exp \left\{ - \sum_{n=1}^{\infty} (1 - s^n E(e^{i\theta S_n})) q^n \frac{1}{n} \right\} \\ &= \exp \left\{ - \sum_{n=1}^{\infty} (1 - s^n E(e^{i\theta S_1})^n) q^n \frac{1}{n} \right\} \\ &= \exp \{ \log(1 - q) - \log(1 - q s E(e^{i\theta S_1})) \} \\ &= \frac{p}{1 - q s E(e^{i\theta S_1})} \end{aligned} \quad (19)$$

where, in the last equality, we have applied the Mercator–Newton series expansion of the logarithm. Comparing the conclusions of the last two series of equalities, the required expression for $E(s^{\Gamma_p} e^{i\theta S_{\Gamma_p}})$ follows. The Lévy measure mentioned in equation (4) is thus identifiable as

$$\Pi(dy, dx) = \sum_{n=1}^{\infty} \delta_{\{n\}}(dy) F^{*n}(dx) \frac{1}{n} q^n \quad (20)$$

for $(y, x) \in \mathbb{R}^2$.

We know that (Γ_p, S_{Γ_p}) may be written as the independent sum of (G, S_G) and $(\Gamma_p - G, S_{\Gamma_p} - S_G)$, where both are infinitely divisible. Further, the former has Lévy measure supported on $\{1, 2, \dots\} \times (0, \infty)$ and the latter has Lévy measure supported on $\{1, 2, \dots\} \times (-\infty, 0)$. In addition, $E(s^G e^{i\theta S_G})$ extends to the upper half of the complex plane in θ (and is continuous on the real axis) and $E(s^{\Gamma_p - G} e^{i\theta(S_{\Gamma_p} - S_G)})$ extends to the lower half of the complex plane in θ (and is continuous on the real axis).^a Taking account of equation (4), this forces the factorization of the expression for $E(s^{\Gamma_p} e^{i\theta S_{\Gamma_p}})$ in such a way that

$$E(s^G e^{i\theta S_G}) = e^{- \int_{(0, \infty)} \sum_{n=1}^{\infty} (1 - s^n e^{i\theta x}) q^n \frac{1}{n} F^{*n}(dx) / n} \quad (21)$$

(iii) Note that the path decomposition given in part (i) shows that

$$E(s^G e^{i\theta S_G}) = E\left(s^{\sum_{i=1}^{\nu} N^{(i)}} e^{i\theta \sum_{i=1}^{\nu} H^{(i)}}\right) \quad (22)$$

where the pairs $\{(N^{(i)}, H^{(i)}) : i = 1, 2, \dots\}$ are independent having the same distribution as (N, S_N) conditioned on $\{N \leq \Gamma_p\}$. Hence, we have

$$\begin{aligned}
 & E(s^G e^{i\theta S_G}) \\
 &= \sum_{k \geq 0} P(N > \Gamma_p) P(N \leq \Gamma_p)^k \\
 &\quad \times E\left(s^{\sum_{i=1}^k N^{(i)}} e^{i\theta \sum_{i=1}^k H^{(i)}}\right) \\
 &= \sum_{k \geq 0} P(N > \Gamma_p) P(N \leq \Gamma_p)^k E(s^N e^{i\theta S_N} | N \leq \Gamma_p)^k \\
 &= \sum_{k \geq 0} P(N > \Gamma_p) E(s^N e^{i\theta S_N} \mathbf{1}_{(N \leq \Gamma_p)})^k \\
 &= \sum_{k \geq 0} P(N > \Gamma_p) E((qs)^N e^{i\theta S_N})^k \\
 &= \frac{P(N > \Gamma_p)}{1 - E((qs)^N e^{i\theta S_N})} \tag{23}
 \end{aligned}$$

Note that in the fourth equality we use the fact that $P(\Gamma_p \geq n) = q^n$.

The required equality to be proved follows by setting $s = 0$ in equation (21) to recover

$$P(N > \Gamma_p) = \exp \left\{ - \int_{(0, \infty)} \sum_{n=1}^{\infty} \frac{q^n}{n} F^{*n}(\mathrm{d}x) \right\} \tag{24}$$

and then plugging this back into the right-hand side of equation (23) and rearranging.

Lévy Processes and Infinite Divisibility

A (one-dimensional) stochastic process $X = \{X_t : t \geq 0\}$ is called a *Lévy process* (see **Lévy Processes**) on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ if

1. X has paths that are $\mathbb{P}$ -almost surely right continuous with left limits;
2. given $0 \leq s \leq t < \infty$, $X_t - X_s$ is independent of $\{X_u : u \leq s\}$;
3. given $0 \leq s \leq t < \infty$, $X_t - X_s$ is equal in distribution to X_{t-s} ; and

$$\mathbb{P}(X_0 = 0) = 1 \tag{25}$$

It is easy to deduce that if X is a Lévy process, then for each $t > 0$ the random variable X_t is infinitely divisible. Indeed, one may also show *via* a straightforward computation that

$$\mathbb{E}(e^{i\theta X_t}) = e^{-\Psi(\theta)t} \text{ for all } \theta \in \mathbb{R}, \quad t \geq 0 \tag{26}$$

where, in its most general form, Ψ takes the form given in equation (4). Conversely, it can also be shown that given a Lévy–Khintchine exponent (4) of an infinitely divisible random variable, there exists a Lévy process that satisfies equation (26). In the special case that the Lévy–Khintchine exponent Ψ belongs to that of a positive-valued infinitely divisible distribution, it follows that the increments of the associated Lévy process must be positive and hence its paths are necessarily monotone increasing. In full generality, a Lévy process may be naively thought of as the independent sum of a linear Brownian motion plus an independent process with discontinuities in its path, which, in turn, may be seen as the limit (in an appropriate sense) of the partial sums of a sequence of compound Poisson processes with drift. The book by Bertoin [1] gives a comprehensive account of the above details.

The definition of a Lévy process suggests that it may be thought of as a continuous-time analog of a random walk. Let us introduce the exponential random variable with parameter p , denoted by $\mathbf{e}_p$, which henceforth is assumed to be independent of all other random quantities under discussion and defined on the same probability space. Like the geometric distribution, the exponential distribution also has a lack-of-memory property in the sense that for all $0 \leq s, t < \infty$ we have $\mathbb{P}(\mathbf{e}_p > t + s | \mathbf{e}_p > t) = \mathbb{P}(\mathbf{e}_p > s) = e^{-ps}$. Moreover, $\mathbf{e}_p$, and, more generally, $X_{\mathbf{e}_p}$, is infinitely divisible. Indeed, straightforward computations show that for each $n = 1, 2, 3, \dots$

$$\mathbb{E}(e^{i\theta X_{\mathbf{e}_p}}) = \left(\left(\frac{p}{p + \Psi(\theta)} \right)^{\frac{1}{n}} \right)^n = \mathbb{E}(e^{i\theta X_{\gamma_{1/n, p}}})^n \tag{27}$$

where $\gamma_{1/n, p}$ is a gamma distribution with parameters $1/n$ and p , which is independent of X . The latter has distribution

$$\mathbb{P}(\gamma_{1/n, p} \in \mathrm{d}x) = \frac{p^{1/n}}{\Gamma(1/n)} x^{-1+1/n} e^{-px} \mathrm{d}x \tag{28}$$

for $x > 0$.

Wiener–Hopf Factorization for Lévy Processes

The Wiener–Hopf factorization for a one-dimensional Lévy processes is slightly more technical than for random walks but, in principle, appeals to essentially the same ideas that have been exhibited in the above exposition of the Wiener–Hopf factorization for random walks. In this section, therefore, we give only a statement of the Wiener–Hopf factorization. The reader who is interested in the full technical details is directed primarily to the article by Greenwood and Pitman [6] for a natural and insightful probabilistic presentation (in the author’s opinion). Alternative accounts based on the aforementioned article can be found in the books by Bertoin [1] and Kyprianou [12], and derivation of the Wiener–Hopf factorization for Lévy processes from the Wiener–Hopf factorization for random walks can be found in [18].

Before proceeding to the statement of the Wiener–Hopf factorization, we first need to introduce the *ladder process* associated with any Lévy process X . Here, we encounter more subtleties than for the random walk. Consider the range of the times and positions at which the process X attains new maxima. That is to say, the random set $\{(t, \bar{X}_t) : \bar{X}_t = X_t\}$ where $\bar{X}_t = \sup_{s \leq t} X_s$ is the running maximum. It turns out that this range is equal in law to the range of a killed bivariate subordinator $(\tau, H) = \{(\tau_t, H_t) : t < \zeta\}$, where the killing time ζ is an independent and exponentially distributed random variable with some rate $\lambda \geq 0$. In the case that $\lim_{t \uparrow \infty} \bar{X}_t = \infty$, there should be no killing in the process (τ, H) and hence $\lambda = 0$ and we interpret $\mathbb{P}(\zeta = \infty) = 1$. Note that we may readily define the Laplace exponent of the killed process (τ, H) by

$$\mathbb{E}(e^{-\alpha\tau_t - \beta H_t} \mathbf{1}_{(t < \zeta)}) = e^{-\kappa(\alpha, \beta)t} \quad (29)$$

for all $\alpha, \beta \geq 0$ where, necessarily, $\kappa(\alpha, \beta) = \lambda + \phi(\alpha, \beta)$ is the rate of ζ , and ϕ is the bivariate Laplace exponent of the unkilld process $\{(\tau_t, H_t) : t \geq 0\}$. Analogous to the role played by joint probability generating and characteristic exponent of the pair (N, S_N) in Theorem 1 (iii), the quantity $\kappa(\alpha, \beta)$ also is prominent in the Wiener–Hopf factorization for Lévy processes, which we state below. To do so, we give one final definition. For each $t > 0$, let

$$\bar{G}_{\mathbf{e}_p} = \sup\{s < \mathbf{e}_p : X_s = \bar{X}_s\} \quad (30)$$

Theorem 2 (The Wiener–Hopf Factorization for Lévy Processes) *Suppose that X is any Lévy process other than a compound Poisson process. As usual, denote by $\mathbf{e}_p$ an independent and exponentially distributed random variable.*

(i) *The pairs*

$$(\bar{G}_{\mathbf{e}_p}, \bar{X}_{\mathbf{e}_p}) \text{ and } (\mathbf{e}_p - \bar{G}_{\mathbf{e}_p}, \bar{X}_{\mathbf{e}_p} - X_{\mathbf{e}_p}) \quad (31)$$

are independent and infinitely divisible.

(ii) *For $\alpha, \beta \geq 0$*

$$\mathbb{E} \left(e^{-\alpha \bar{G}_{\mathbf{e}_p} - \beta \bar{X}_{\mathbf{e}_p}} \right) = \frac{\kappa(p, 0)}{\kappa(p + \alpha, \beta)} \quad (32)$$

(iii) *The Laplace exponent $\kappa(\alpha, \beta)$ may be identified in terms of the law of X in the following way,*

$$\begin{aligned} \kappa(\alpha, \beta) &= k \exp \left(\int_0^\infty \int_0^\infty (e^{-t} - e^{-\alpha t - \beta x}) \right. \\ &\quad \left. \times \mathbb{P}(X_t \in dx) \frac{dt}{t} \right) \end{aligned} \quad (33)$$

where $\alpha, \beta \geq 0$ and k is a dimensionless strictly positive constant.

The First Passage Problem and Mathematical Finance

There are many applications of the Wiener–Hopf factorization in applied probability, and mathematical finance is no exception in this respect. One of the most prolific links is the relationship between the information contained in the Wiener–Hopf factorization and the distributions of the first passage times

$$\begin{aligned} \tau_x^+ &:= \inf\{t > 0 : X_t > x\} \text{ and} \\ \tau_x^- &:= \inf\{t > 0 : X_t < x\} \end{aligned} \quad (34)$$

together with the overshoots $X_{\tau_x^+} - x$ and $x - X_{\tau_x^-}$, where $x \in \mathbb{R}$. In turn, this is helpful for the pricing of certain types of exotic options.

For example, in a simple market model for which there is one risky asset modeled by an exponential Lévy process and one riskless asset with a fixed rate of return, say $r > 0$, the value of a perpetual American put, or indeed a perpetual down-and-in

put, boils down to the computation of the following quantity:

$$v_y(x) := \mathbb{E} \left(e^{-r\tau_y^-} \left(K - e^{X_{\tau_y^-}} \right)^+ | X_0 = x \right) \quad (35)$$

where $y \in \mathbb{R}$ and $z^+ = \max\{0, z\}$ and the expectation is taken with respect to an appropriate risk-neutral measure that keeps X in the class of Lévy processes (e.g., the measure that occurs as a result of the Escher transform). To see the connection with the Wiener–Hopf factorization consider the following lemma and its corollary:

Lemma 1 For all $\alpha > 0$, $\beta \geq 0$ and $x \geq 0$ we have

$$\mathbb{E} \left(e^{-\alpha\tau_x^+ - \beta X_{\tau_x^+}} \mathbf{1}_{(\tau_x^+ < \infty)} \right) = \frac{\mathbb{E} \left(e^{-\beta \bar{X}_{e_\alpha}} \mathbf{1}_{(\bar{X}_{e_\alpha} > x)} \right)}{\mathbb{E} \left(e^{-\beta \bar{X}_{e_\alpha}} \right)} \quad (36)$$

Proof First, assume that $\alpha, \beta, x > 0$ and note that

$$\begin{aligned} & \mathbb{E} \left(e^{-\beta \bar{X}_{e_\alpha}} \mathbf{1}_{(\bar{X}_{e_\alpha} > x)} \right) \\ &= \mathbb{E} \left(e^{-\beta \bar{X}_{e_\alpha}} \mathbf{1}_{(\tau_x^+ < e_\alpha)} \right) \\ &= \mathbb{E} \left(\mathbf{1}_{(\tau_x^+ < e_\alpha)} e^{-\beta X_{\tau_x^+}} \mathbb{E} \left(e^{-\beta (\bar{X}_{e_\alpha} - X_{\tau_x^+})} \middle| \mathcal{F}_{\tau_x^+} \right) \right) \end{aligned} \quad (37)$$

Now, conditionally on $\mathcal{F}_{\tau_x^+}$ and on the event $\tau_x^+ < e_\alpha$, the random variables $\bar{X}_{e_\alpha} - X_{\tau_x^+}$ and $\bar{X}_{e_\alpha}$ have the same distribution, thanks to the lack-of-memory property of e_α and the strong Markov property. Hence, we have the factorization

$$\mathbb{E} \left(e^{-\beta \bar{X}_{e_\alpha}} \mathbf{1}_{(\bar{X}_{e_\alpha} > x)} \right) = \mathbb{E} \left(e^{-\alpha\tau_x^+ - \beta X_{\tau_x^+}} \right) \mathbb{E} \left(e^{-\beta \bar{X}_{e_\alpha}} \right) \quad (38)$$

The case that β or x is equal to zero can be achieved by taking limits on both sides of the above equality.

By replacing X by $-X$ in Lemma 1, we get the following analogous result for first passage into the negative half line.

Corollary 1 For all $\alpha, \beta \geq 0$ and $x \geq 0$, we have

$$\mathbb{E} \left(e^{-\alpha\tau_{-x}^- + \beta X_{\tau_{-x}^-}} \mathbf{1}_{(\tau_{-x}^- < \infty)} \right) = \frac{\mathbb{E} \left(e^{\beta \underline{X}_{e_\alpha}} \mathbf{1}_{(-\underline{X}_{e_\alpha} > x)} \right)}{\mathbb{E} \left(e^{\beta \underline{X}_{e_\alpha}} \right)} \quad (39)$$

In that case, we may develop the expression in equation (35) by using Corollary 1 to obtain

$$v_y(x) = \frac{\mathbb{E} \left(\left(K \mathbb{E} \left[e^{\underline{X}_{e_r}} \right] - e^{x + \underline{X}_{e_r}} \right) \mathbf{1}_{(-\underline{X}_{e_r} > x - y)} \right)}{\mathbb{E} \left(e^{\underline{X}_{e_r}} \right)} \quad (40)$$

where $\underline{X}_t = \inf_{s \leq t} X_s$ is the running infimum. Ultimately, further development of the expression on the right-hand side above requires knowledge of the distribution of $\underline{X}_{e_r}$. This is information, which, in principle, can be extracted from the Wiener–Hopf factorization.

We conclude by mentioning the articles [5, 10] and [11] in which the Wiener–Hopf factorization is used for the pricing of barrier options (see **Lookback Options**).

End Notes

^aIt is this part of the proof that makes the connection with the general analytic technique of the Wiener–Hopf method of factorizing operators. This also explains the origin of the terminology *Wiener–Hopf factorization* for what is otherwise a path, and consequently distributional, decomposition.

References

- [1] Bertoin, J. (1996). *Lévy Processes*, Cambridge University Press.
- [2] Borovkov, A.A. (1976). *Stochastic Processes in Queueing Theory*, Springer-Verlag.
- [3] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, 2nd Edition, Wiley, Vol. II.
- [4] Fristedt, B.E. (1974). Sample functions of stochastic processes with stationary independent increments, *Advances in Probability* **3**, 241–396.
- [5] Fusai, G., Abrahams, I.D. & Sgarra, C. (2006). An exact analytical solution for discrete barrier options, *Finance and Stochastics* **10**, 1–26.
- [6] Greenwood, P.E. & Pitman, J.W. (1979). Fluctuation identities for Lévy processes and splitting at

- the maximum, *Advances in Applied Probability* **12**, 839–902.
- [7] Greenwood, P.E. & Pitman, J.W. (1980). Fluctuation identities for random walk by path decomposition at the maximum. Abstracts of the Ninth Conference on Stochastic Processes and Their Applications, Evanston, Illinois, 6–10 August 1979, *Advances in Applied Probability* **12**, 291–293.
 - [8] Gusak, D.V. & Korolyuk, V.S. (1969). On the joint distribution of a process with stationary independent increments and its maximum. *Theory of Probability* **14**, 400–409.
 - [9] Hopf, E. (1934). *Mathematical Problems of Radiative Equilibrium*. Cambridge tracts, No. 31.
 - [10] Jeannin, M. & Pistorius, M.R. (2007). *A Transform Approach to Calculate Prices and Greeks of Barrier Options Driven by a Class of Lévy*. Available at arXiv: <http://arxiv.org/abs/0812.3128>.
 - [11] Kudryavtsev, O. & Levendorski, S.Z. (2007). *Fast and Accurate Pricing of Barrier Options Under Levy Processes*. Available at SSRN: <http://ssrn.com/abstract=1040061>.
 - [12] Kyprianou, A.E. (2006). *Introductory Lectures on Fluctuations of Lévy Processes with Applications*, Springer.
 - [13] Payley, R. & Wiener, N. (1934). *Fourier Transforms in the Complex Domain*, American Mathematical Society. Colloquium Publications, New York, Vol. 19.
 - [14] Percheskii, E.A. & Rogozin, B.A. (1969). On the joint distribution of random variables associated with fluctuations of a process with independent increments, *Theory of Probability and its Applications* **14**, 410–423.
 - [15] Spitzer, E. (1956). A combinatorial lemma and its application to probability theory, *Transactions of the American Mathematical Society* **82**, 323–339.
 - [16] Spitzer, E. (1957). The Wiener-Hopf equation whose kernel is a probability density, *Duke Mathematical Journal* **24**, 327–343.
 - [17] Spitzer, E. (1964). *Principles of Random Walk*, Van Nostrand.
 - [18] Sato, K.-I. (1999). *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press.

Related Articles

Fractional Brownian Motion; Infinite Divisibility; Lévy Processes; Lookback Options.

ANDREAS E. KYPRIANOU

Poisson Process

In this article, we present the main results on Poisson processes, which are standard examples of jump processes. The reader can refer to the books [2, 5] for the study of standard Poisson processes, or [1, 3, 4, 6] for general Poisson processes.

Counting Processes and Stochastic Integrals

Let $(T_n, n \geq 0)$ be a strictly increasing sequence of random times (i.e., nonnegative random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$) such that $\lim_{n \rightarrow \infty} T_n = \infty$, with $T_0 = 0$. The counting process N associated with $(T_n, n \geq 0)$ is defined as

$$N_t = \begin{cases} n & \text{if } t \in [T_n, T_{n+1}[\\ +\infty & \text{otherwise} \end{cases} \quad (1)$$

or, equivalently,

$$N_t = \sum_{n \geq 1} \mathbb{1}_{\{T_n \leq t\}} = \sum_{n \geq 1} n \mathbb{1}_{\{T_n \leq t < T_{n+1}\}} \quad (2)$$

It is an increasing, right-continuous process. We denote by N_{t-} the left limit of N_s when $s \rightarrow t, s < t$ and by $\Delta N_s = N_s - N_{s-}$ the jump process of N . The stochastic integral of a real-valued process C with respect to the increasing process N is defined as

$$\begin{aligned} (C \star N)_t &:= \int_0^t C_s dN_s = \int_{[0, t]} C_s dN_s \\ &= \sum_{n=1}^{\infty} C_{T_n} \mathbb{1}_{\{T_n \leq t\}} \end{aligned} \quad (3)$$

The natural filtration of N (i.e., the smallest right-continuous and complete filtration that makes the process N adapted) is denoted by $\mathbf{F}^N$.

Standard Poisson Process

The standard Poisson process is a counting process $(N_t, t \geq 0)$ with stationary and independent increments, that is,

- for every $s, t \geq 0$, $N_{t+s} - N_t$ is independent of $\mathcal{F}_t^N$; and

- for every $s, t \geq 0$, the r.v. $N_{t+s} - N_t$ has the same law as N_s .

For any fixed $t \geq 0$, the random variable N_t has a Poisson law, with parameter λt , that is, $\mathbb{P}(N_t = n) = e^{-\lambda t} ((\lambda t)^n / n!)$ and, for every $x > 0, t > 0, u, \alpha \in \mathbb{R}$

$$\begin{aligned} \mathbb{E}(N_t) &= \lambda t, \quad \text{Var}(N_t) = \lambda t \\ \mathbb{E}(x^{N_t}) &= e^{\lambda t(x-1)}; \quad \mathbb{E}(e^{iuN_t}) = e^{\lambda t(e^{iu}-1)}; \\ \mathbb{E}(e^{\alpha N_t}) &= e^{\lambda t(e^{\alpha}-1)} \end{aligned} \quad (4)$$

From the property of independence and stationarity of the increments, it follows that the process $(M_t := N_t - \lambda t, t \geq 0)$ is a martingale. More generally, if H is an $\mathbf{F}^N$ -predictable^a bounded process, then the following processes are $\mathbf{F}^N$ -martingales:

$$\begin{aligned} (H \star M)_t &:= \int_0^t H_s dM_s = \int_0^t H_s dN_s - \lambda \int_0^t H_s ds \\ ((H \star M)_t)^2 &- \lambda \int_0^t H_s^2 ds \\ \exp \left(\int_0^t H_s dN_s - \lambda \int_0^t (e^{H_s} - 1) ds \right) \end{aligned} \quad (5)$$

In particular, the processes $(M_t^2 - \lambda t, t \geq 0)$ and $(M_t^2 - N_t, t \geq 0)$ are martingales. The process $(\lambda t, t \geq 0)$ is the predictable quadratic variation process of M (or the compensator of N), denoted by $\langle N \rangle$, the process $(N_t, t \geq 0)$ equals in this case its optional quadratic variation, denoted by $[N]$.

The above martingale properties do not extend to $\mathbf{F}^N$ -adapted processes H . For example, from the simple equality $\int_0^t (N_s - N_{s-}) dM_s = N_t$, it follows that $\int_0^t N_s dM_s$ is not a martingale.

Predictable Representation Property

Proposition 1 *Let N be a Poisson process, and $H_\infty \in L^2(\mathcal{F}_\infty^N)$, a square-integrable random variable. Then, there exists an $\mathbf{F}^N$ -predictable process $(h_s, s \geq 0)$ such that*

$$H_\infty = \mathbb{E}(H_\infty) + \int_0^\infty h_s dM_s \quad (6)$$

and $\mathbb{E}(\int_0^\infty h_s^2 ds) < \infty$, where $M_t = N_t - \lambda t$.

It follows that if X is a square-integrable $\mathbf{F}^N$ -martingale, there exists an $\mathbf{F}^N$ -predictable process $(x_s, s \geq 0)$ such that $X_t = X_0 + \int_0^t x_s dM_s$.

Independent Poisson Processes

Here, we assume that the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is endowed with a filtration $\mathbf{F}$.

A process $(N^1, \dots, N^d)$ is a d -dimensional $\mathbf{F}$ -Poisson process (with $d \geq 1$) if each $(N^j, j = 1, \dots, d)$ is a right-continuous $\mathbf{F}$ -adapted process such that $N_0^j = 0$, and if there exist constants $(\lambda_j, j = 1, \dots, d)$ such that for every $t \geq s \geq 0$, $\forall n_j \in \mathbb{N}$,

$$\begin{aligned} \mathbb{P} \left[\bigcap_{j=1}^d (N_t^j - N_s^j = n_j) | \mathcal{F}_s \right] \\ = \prod_{j=1}^d e^{-\lambda_j(t-s)} \frac{(\lambda_j(t-s))^{n_j}}{n_j!} \end{aligned} \quad (7)$$

Proposition 2 An $\mathbf{F}$ -adapted process N is a d -dimensional $\mathbf{F}$ -Poisson process if and only if

1. each N^j is an $\mathbf{F}$ -Poisson process
2. no two N^j 's jump simultaneously.

Inhomogeneous Poisson Processes

We assume that the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is endowed with a filtration $\mathbf{F}$.

Definition

Let λ be an $\mathbf{F}$ -adapted nonnegative process satisfying $\mathbb{E} \left(\int_0^t \lambda_s ds \right) < \infty, \forall t$, and $\int_0^\infty \lambda_s ds = \infty$.

An inhomogeneous Poisson process N with stochastic intensity λ is a counting process such that for every nonnegative $\mathbf{F}$ -predictable process $(\phi_t, t \geq 0)$, the following equality is satisfied:

$$\mathbb{E} \left(\int_0^\infty \phi_s dN_s \right) = \mathbb{E} \left(\int_0^\infty \phi_s \lambda_s ds \right) \quad (8)$$

Therefore $(M_t = N_t - \int_0^t \lambda_s ds, t \geq 0)$ is an $\mathbf{F}$ -martingale, and if ϕ is an $\mathbf{F}$ -predictable process such that $\forall t, \mathbb{E}(\int_0^t |\phi_s| \lambda_s ds) < \infty$, then $(\int_0^t \phi_s dM_s, t \geq 0)$ is an $\mathbf{F}$ -martingale. The process $\Lambda_t = \int_0^t \lambda_s ds$ is called the *compensator* of N .

An inhomogeneous Poisson process with stochastic intensity λ can be viewed as a time change of $\tilde{N}$, a standard Poisson process: indeed, the process $(N_t = \tilde{N}_{\Lambda_t}, t \geq 0)$ is an inhomogeneous Poisson process with stochastic intensity $(\lambda_t, t \geq 0)$.

For H an $\mathbf{F}$ -predictable process satisfying some integrability conditions, the following processes are martingales:

$$\begin{aligned} (H \star M)_t &= \int_0^t H_s dM_s = \int_0^t H_s dN_s - \int_0^t \lambda_s H_s ds \\ ((H \star M)_t)^2 &- \int_0^t \lambda_s H_s^2 ds \\ \exp \left(\int_0^t H_s dN_s - \int_0^t \lambda_s (e^{H_s} - 1) ds \right) \end{aligned} \quad (9)$$

Stochastic Calculus

Integration by Parts Formula. Let $dX_t = b_t dt + \varphi_t dM_t$ and $dY_t = c_t dt + \psi_t dM_t$, where φ and ψ are predictable processes, and b, c are adapted processes such that the processes X and Y are well defined. Then,

$$X_t Y_t = xy + \int_0^t Y_{s-} dX_s + \int_0^t X_{s-} dY_s + [X, Y]_t \quad (10)$$

where $[X, Y]_t$ is the quadratic covariation process, defined as

$$[X, Y]_t := \int_0^t \varphi_s \psi_s dN_s \quad (11)$$

In particular, if $dX_t = \varphi_t dM_t$ and $dY_t = \psi_t dM_t$ (i.e., X and Y are local martingales), the process $(X_t Y_t - [X, Y]_t, t \geq 0)$ is a martingale. It can be noted that, in that case, the process $(X_t Y_t - \langle X, Y \rangle_t, t \geq 0)$, where $\langle X, Y \rangle_t = \int_0^t \varphi_s \psi_s \lambda_s ds$ is also a martingale. The process $\langle X, Y \rangle$ is the compensator of $[X, Y]$ if $[X, Y]$ is integrable (see **Compensators**). The predictable process $(\langle X, Y \rangle_t, t \geq 0)$ is called the *predictable covariation process* of the pair (X, Y) , or the *compensator* of the product XY . If $dX_t^i = x_t^i dN_t^i$, where $N^i, i = 1, 2$ are independent inhomogeneous Poisson processes, the covariation processes $[X^1, X^1]$ and $\langle X^1, X^2 \rangle$ are null, and $X^1 X^2$ is a martingale.

Itô's Formula. Itô's formula is a special case of the general one; it is a bit simpler and is used for the

processes that are within bounded variation. Let b be an adapted process and φ a predictable process with adequate integrability conditions, and

$$dX_t = b_t dt + \varphi_t dM_t = (b_t - \varphi_t \lambda_t) dt + \varphi_t dN_t \quad (12)$$

and $F \in C^{1,1}(\mathbb{R}^+ \times \mathbb{R})$. Then, the process $(F(t, X_t), t \geq 0)$ is a semimartingale with decomposition

$$F(t, X_t) = Z_t + A_t \quad (13)$$

where Z is a local martingale given by

$$Z_t = F(0, X_0) + \int_0^t [F(s, X_{s-} + \varphi_s) - F(s, X_{s-})] dM_s \quad (14)$$

and A a bounded variation process

$$A_t = \int_0^t (\partial_t F(s, X_s) + b_s \partial_x F(s, X_s) + \lambda_s [F(s, X_{s-} + \varphi_s) - F(s, X_s) - \varphi_s \partial_x F(s, X_s)]) ds \quad (15)$$

Exponential Martingales

Proposition 3 Let N be an inhomogeneous Poisson process with stochastic intensity $(\lambda_t, t \geq 0)$, and $(\mu_t, t \geq 0)$ a predictable process such that $\int_0^t |\mu_s| \lambda_s ds < \infty$. Then, the process L defined by

$$L_t = \begin{cases} \exp\left(-\int_0^t \mu_s \lambda_s ds\right) & \text{if } t < T_1 \\ \prod_{n, T_n \leq t} (1 + \mu_{T_n}) \times \exp\left(-\int_0^t \mu_s \lambda_s ds\right) & \text{if } t \geq T_1 \end{cases} \quad (16)$$

is a local martingale solution of

$$dL_t = L_t - \mu_t dM_t, \quad L_0 = 1 \quad (17)$$

Moreover, if μ is such that $\forall s, \mu_s > -1$,

$$\begin{aligned} L_t &= \exp\left[-\int_0^t \mu_s \lambda_s ds + \int_0^t \ln(1 + \mu_s) dN_s\right] \\ &= \exp - \int_0^t (\mu_s - \ln(1 + \mu_s)) \lambda_s ds \\ &\quad + \int_0^t \ln(1 + \mu_s) dM_s \end{aligned} \quad (18)$$

The local martingale L is denoted by $\mathcal{E}(\mu \star M)$ and named the *Doléans-Dade exponential* (alternatively, the *stochastic exponential*) of the process $\mu \star M$. If $\mu > -1$, the process L is nonnegative and is a martingale if $\forall t, \mathbb{E}(L_t) = 1$ (this is the case if μ satisfies $-1 + \delta < \mu_s < C$ where C and $\delta > 0$ are two constants).

If μ is not greater than -1 , then the process L defined in equation (16) may take negative values.

Change of Probability Measure

Let μ be a predictable process such that $\mu > -1$, and $\int_0^t \lambda_s |\mu_s| ds < \infty$, and let L be the positive exponential local martingale solution of

$$dL_t = L_t - \mu_t dM_t, \quad L_0 = 1 \quad (19)$$

Assume that L is a martingale, and let $\mathbb{Q}$ be the probability measure equivalent to $\mathbb{P}$ defined on $\mathcal{F}_t$ by $\mathbb{Q}|_{\mathcal{F}_t} = L_t \mathbb{P}|_{\mathcal{F}_t}$. Under $\mathbb{Q}$, the process

$$\begin{aligned} M_t^\mu &:= M_t - \int_0^t \mu_s \lambda_s ds \\ &= N_t - \int_0^t (\mu_s + 1) \lambda_s ds \quad t \geq 0 \end{aligned} \quad (20)$$

is a local martingale, hence N is a $\mathbb{Q}$ -inhomogeneous Poisson process, with intensity $\lambda(1 + \mu)$.

Compound Poisson Processes

Definition and Properties

Let λ be a positive number, and $F(dy)$ be a probability law on $\mathbb{R}$. A (λ, F) -compound Poisson process is a process $X = (X_t, t \geq 0)$ of the form

$$X_t = \sum_{n=1}^{N_t} Y_n = \sum_{n>0, T_n \leq t} Y_n \quad (21)$$

where N is a standard Poisson process with intensity $\lambda > 0$, and the $(Y_n, n \geq 1)$ are i.i.d. square-integrable random variables with law $F(dy) = \mathbb{P}(Y_1 \in dy)$, independent of N .

Proposition 4 A compound Poisson process has stationary and independent increments; for fixed t , the

cumulative distribution function of X_t is

$$\mathbb{P}(X_t \leq x) = e^{-\lambda t} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} F^{*n}(x) \quad (22)$$

where the star indicates a convolution.

If $\mathbb{E}(|Y_1|) < \infty$, the process $(Z_t = X_t - t\lambda\mathbb{E}(Y_1), t \geq 0)$ is a martingale and $\mathbb{E}(X_t) = \lambda t\mathbb{E}(Y_1)$.

If $\mathbb{E}(Y_1^2) < \infty$, the process $(Z_t^2 - t\lambda\mathbb{E}(Y_1^2), t \geq 0)$ is a martingale and $\text{Var}(X_t) = \lambda t\mathbb{E}(Y_1^2)$.

Introducing the random measure $\mu = \sum_{n=1}^{\infty} \delta_{T_n, Y_n}$ on $\mathbb{R}^+ \times \mathbb{R}$, that is,

$$\mu(\omega,]0, t], A) = \sum_{n>0, T_n(\omega) \leq t} \mathbb{1}_{Y_n(\omega) \in A} \quad (23)$$

and denoting by $(f * \mu)_t$, the integral

$$\begin{aligned} \int_0^t \int_{\mathbb{R}} f(x) \mu(\omega, ds, dx) &= \sum_{n>0, T_n \leq t} f(Y_n(\omega)) \\ &= \sum_{n=1}^{N_t} f(Y_n(\omega)) \end{aligned} \quad (24)$$

we obtain that

$$\begin{aligned} M_t^f &= (f * \mu)_t - t\lambda\mathbb{E}(f(Y_1)) \\ &= \int_0^t \int_{\mathbb{R}} f(x) (\mu(\omega, ds, dx) - \lambda F(dx) ds) \end{aligned} \quad (25)$$

is a martingale.

Martingales

Proposition 5 *If X is a (λ, F) -compound Poisson process, for any α such that $\int_{-\infty}^{\infty} e^{\alpha x} F(dx) < \infty$, the process*

$$Z_t = \exp\left(\alpha X_t - \lambda t \int_{-\infty}^{\infty} (e^{\alpha x} - 1) F(dx)\right) \quad (26)$$

is a martingale and

$$\begin{aligned} \mathbb{E}(e^{\alpha X_t}) &= \exp\left(\lambda t \int_{-\infty}^{\infty} (e^{\alpha x} - 1) F(dx)\right) \\ &= \exp(\lambda t (\mathbb{E}(e^{\alpha Y_1}) - 1)) \end{aligned} \quad (27)$$

In other words, for any α such that $\mathbb{E}(e^{\alpha X_t}) < \infty$ (or equivalently $\mathbb{E}(e^{\alpha Y_1}) < \infty$), the process $(e^{\alpha X_t} / \mathbb{E}(e^{\alpha X_t}), t \geq 0)$ is a martingale. More generally, let f be a bounded Borel function. Then, the process

$$\exp\left(\sum_{n=1}^{N_t} f(Y_n) - \lambda t \int_{-\infty}^{\infty} (e^{f(x)} - 1) F(dx)\right) \quad (28)$$

is a martingale. In particular,

$$\begin{aligned} \mathbb{E}\left(\exp\left(\sum_{n=1}^{N_t} f(Y_n)\right)\right) \\ = \exp\left(\lambda t \int_{-\infty}^{\infty} (e^{f(x)} - 1) F(dx)\right) \end{aligned} \quad (29)$$

Change of Measure

Let X be a (λ, F) -compound Poisson process, $\tilde{\lambda} > 0$, and $\tilde{F}$ a probability measure on $\mathbb{R}$, absolutely continuous with respect to F , with Radon–Nikodym density φ , that is, $\tilde{F}(dx) = \varphi(x)F(dx)$. The process

$$L_t = \exp\left(t(\lambda - \tilde{\lambda}) + \sum_{s \leq t} \ln\left(\frac{\tilde{\lambda}}{\lambda} \varphi(\Delta X_s)\right)\right) \quad (30)$$

is a positive martingale (take $f(x) = \ln((\tilde{\lambda}/\lambda) \varphi(x))$ in equation (28)) with expectation 1. Set $d\mathbb{Q}|_{\mathcal{F}_t} = L_t d\mathbb{P}|_{\mathcal{F}_t}$.

Proposition 6 *Under $\mathbb{Q}$, the process X is a $(\tilde{\lambda}, \tilde{F})$ -compound Poisson process.*

Let α be such that $\mathbb{E}(e^{\alpha Y_1}) < \infty$. The particular case with $\varphi(x) = (e^{\alpha x} / \mathbb{E}(e^{\alpha Y_1}))$ and $\tilde{\lambda} = \lambda \mathbb{E}(e^{\alpha Y_1})$ corresponds to the Esscher transform for which

$$d\mathbb{Q}|_{\mathcal{F}_t} = \frac{e^{\alpha X_t}}{\mathbb{E}(e^{\alpha X_t})} d\mathbb{P}|_{\mathcal{F}_t} \quad (31)$$

We emphasize that there exist changes of probability that do not preserve the compound Poisson process property. For the predictable representation theorem, see **Point Processes**.

An Example: Double Exponential Model

The compound Poisson process is said to be a double exponential process if the law of the random variable Y_1 is

$$F(dx) = \left(p\theta_1 e^{-\theta_1 x} \mathbb{1}_{\{x>0\}} + (1-p)\theta_2 e^{\theta_2 x} \mathbb{1}_{\{x<0\}} \right) dx \quad (32)$$

where $p \in]0, 1[$ and $\theta_i, i = 1, 2$ are positive numbers. Under an Esscher transform, this model is still a double exponential model. This particular dynamic allows one to compute the Laplace transform of the first hitting times of a given level.

End Notes

^aWe recall that adapted continuous-on-left processes are predictable. The process N is not predictable.

References

- [1] Brémaud, P. (1981). *Point Processes and Queues: Martingale Dynamics*, Springer-Verlag, Berlin.
- [2] Çinlar, E. (1975). *Introduction to Stochastic Processes*, Prentice Hall.
- [3] Cont, R. & Tankov, P. (2004). *Financial Modeling with Jump Processes*, Chapman & Hall/CRC.
- [4] Jeanblanc, M., Yor, M. & Chesney, M. (2009). *Mathematical Models for Financial Markets*, Springer, Berlin.
- [5] Karlin, S. & Taylor, H. (1975). *A First Course in Stochastic Processes*, Academic Press, San Diego.
- [6] Protter, P.E. (2005). *Stochastic Integration and Differential Equations*, 2nd Edition, Springer, Berlin.

Related Articles

Lévy Processes; Martingales; Martingale Representation Theorem.

MONIQUE JEANBLANC

Point Processes

This article gives a brief overview of general point processes. We refer to the books [1–5], for proofs and advanced results.

Marked Point Processes

Definition

An increasing sequence of random times is called a *univariate point process*. A simple example is the Poisson process.

Given a univariate point process, we associate to every time T_n a mark Z_n . More precisely, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(Z_n, n \geq 1)$ a sequence of random variables taking values in a measurable space $(E, \mathcal{E})$, and $(T_n, n \geq 1)$ an increasing sequence of nonnegative random variables. We assume that $\lim T_n = \infty$, so that there is only a finite number of n such that, for a given t , one has $T_n \leq t$. We define the process $\mathbf{N}$ as follows. For each set, $A \in \mathcal{E}$, $N_t(A) = \sum_n \mathbb{1}_{\{T_n \leq t\}} \mathbb{1}_{\{Z_n \in A\}}$ is the number of “marks” in the set A before time t . The natural filtration of $\mathbf{N}$ is

$$\mathcal{F}_t^{\mathbf{N}} = \sigma(N_s(A), s \leq t, A \in \mathcal{E}) \quad (1)$$

The predictable σ -algebra $\mathcal{P}$ is the σ -algebra defined on $\Omega \times \mathbb{R}^+$ that is generated by the sets

$$A \times \{0\}, A \in \mathcal{F}_0^{\mathbf{N}}; \quad A \times]s, t], A \in \mathcal{F}_s^{\mathbf{N}}, s \leq t \quad (2)$$

The associated random counting measure $\mu(\omega, ds, dz)$ is defined as follows: let Φ be a map

$$(t, \omega, z) \in (\mathbb{R}^+, \Omega, E) \rightarrow \Phi(t, \omega, z) \in \mathbb{R} \quad (3)$$

We set

$$\begin{aligned} \int_{[0,t]} \int_E \Phi(s, z) \mu(ds, dz) &= \sum_{n=1}^{\infty} \Phi(T_n, Z_n) \mathbb{1}_{\{T_n \leq t\}} \\ &= \sum_{n=1}^{N_t(E)} \Phi(T_n, Z_n) \end{aligned} \quad (4)$$

The process $\mathbf{N}$ is called a *marked point process*. This is a generalization of the compound Poisson process: we have introduced, in particular, a spatial dimension for the size of jumps, which are no more i.i.d. random variables.

A map Φ is predictable if it is $\mathcal{P} \otimes \mathcal{E}$ measurable. The *compensator* of the marked point process $\mathbf{N}$ is the unique predictable random measure ν on $(\mathbb{R}^+ \times E, \mathcal{G} \otimes \mathcal{E})$ such that, for every bounded predictable process Φ

$$\begin{aligned} \mathbb{E} \left(\int_0^t \int_E \Phi(s, z; \omega) \mu(\omega; ds, dz) \right) \\ = \mathbb{E} \left(\int_0^t \int_E \Phi(s, z; \omega) \nu(\omega; ds, dz) \right) \end{aligned} \quad (5)$$

In the case of a marked point process on $\mathbb{R} \times \mathbb{R}^d$, the compensator admits an explicit representation: let $G_n(dt, dz)$ be a regular version of the conditional distribution of (T_{n+1}, Z_{n+1}) with respect to $\mathcal{F}_{T_n}^{\mathbf{N}} = \sigma\{(T_1, Z_1), \dots, (T_n, Z_n)\}$. Then,

$$\nu(dt, dz) = \sum_n \mathbb{1}_{\{T_n < t \leq T_{n+1}\}} \frac{G_n(dt, dz)}{G_n([t, \infty[\times \mathbb{R}^d)} \quad (6)$$

Intensity Process

In what follows, we assume that, for any $A \in \mathcal{E}$, the process $(N_t(A), t \geq 0)$ admits the $\mathbf{F}$ -predictable intensity $(\lambda_t(A), t \geq 0)$, that is, there exists a non-negative process $(\lambda_t(A), t \geq 0)$ such that

$$N_t(A) - \int_0^t \lambda_s(A) ds \quad (7)$$

is an $\mathbf{F}$ -martingale. Then, if $X_t = \sum_{n=1}^{N_t(E)} \Phi(T_n, Z_n)$ where Φ is an $\mathbf{F}$ -predictable process that satisfies

$$\mathbb{E} \left(\int_{[0,t]} \int_E |\Phi(s, z)| \lambda_s(dz) ds \right) < \infty \quad (8)$$

the process

$$\begin{aligned} X_t - \int_0^t \int_E \Phi(s, z) \lambda_s(dz) ds \\ = \int_{[0,t]} \int_E \Phi(s, z) [\mu(ds, dz) - \lambda_s(dz) ds] \end{aligned} \quad (9)$$

is a martingale and, in particular,

$$\begin{aligned} \mathbb{E} \left(\int_{[0,t]} \int_E \Phi(s, z) \mu(ds, dz) \right) \\ = \mathbb{E} \left(\int_{[0,t]} \int_E \Phi(s, z) \lambda_s(dz) ds \right) \end{aligned} \quad (10)$$

The random measure $\mu(ds, dz) - \lambda_s(dz)ds$ is the compensated measure of μ .

Example Compound Poisson Process. Let $X_t = \sum_{n=1}^{N_t} Y_n$ be a (λ, F) -compound Poisson process. We can consider the Y_n s as marks and introduce the marked point process $N_t(A) = \sum_{n=1}^{N_t} \mathbb{1}_{Y_n \in A}$. For any A , the process $(N_t(A), t \geq 0)$ is a compound Poisson process, and $(N_t(A) - \lambda t P(Y_1 \in A), t \geq 0)$ is a martingale. The intensity of the marked point process $\mathbf{N}$ is $\lambda_t(dz) = \lambda F(dz)$. Moreover, if A_i are disjoint sets, the processes $N(A_i)$ are independent. The counting random measure μ satisfies

$$\int_0^t \int_{\mathbb{R}} f(x) \mu(\omega; ds, dx) = \sum_{k=1}^{N_t} f(Y_k) \quad (11)$$

and we obtain, in particular, that, as in the article on Poisson processes (see **Poisson Process**)

$$M_t^f = \int_0^t \int_{\mathbb{R}} f(x) (\mu(\omega; ds, dx) - ds \lambda F(dx)) \quad (12)$$

is a martingale.

Predictable Representation Property

Let $\mathbf{F}^{\mathbf{N}}$ be the filtration generated by the marked point process with intensity $\lambda_s(dz)$. Then, any $(\mathbb{P}, \mathbf{F}^{\mathbf{N}})$ -martingale M admits the representation

$$M_t = M_0 + \int_0^t \int_E \Phi(s, x) (\mu(ds, dx) - \lambda_s(dx)ds) \quad (13)$$

where Φ is a $\mathbf{F}^{\mathbf{N}}$ -predictable process such that

$$\mathbb{E} \left(\int_0^t \int_E |\Phi(s, x)| \lambda_s(dx) ds \right) < \infty \quad (14)$$

Change of Probability Measure

Let μ be the random measure of a marked point process with intensity $\lambda_t(A) = \alpha_t m_t(A)$, where m is a probability measure. We shall say that the marked point process admits $(\alpha_t, m_t(dz))$ as P -local characteristics. Let $(\psi_t, h_t(z))$ be two predictable positive processes such that

$$\int_0^t \psi_s \alpha_s ds < \infty, \quad \int_E h_s(z) m_t(dz) = 1 \quad (15)$$

Let L be the solution of

$$\begin{aligned} dL_t = L_{t-} \int_E (\psi_t h_t(z) - 1) (\mu(dt, dz) \\ - \alpha_t m_t(dz) dt), \quad L_0 = 1 \end{aligned} \quad (16)$$

If $\mathbb{E}(L_t) = 1$ (so that L is a martingale), setting $\mathbb{Q}|_{\mathcal{F}_t} = L_t \mathbb{P}|_{\mathcal{F}_t}$, the marked point process has the $\mathbb{Q}$ -local characteristics $(\psi_t \alpha_t, h_t(z) m_t(dz))$.

Example Compound Poisson Process. The change of measure for compound Poisson processes can be written in terms of random measures. Let

$$\begin{aligned} L_t = \exp \left(\int_{\mathbb{R}} f(x) N_t(dx) \right. \\ \left. - t \lambda \int_{-\infty}^{\infty} (e^{f(x)} - 1) F(dx) \right) \\ = \exp \left(\int_0^t \int_{\mathbb{R}} f(x) \mu(ds, dx) \right. \\ \left. - t \int_{-\infty}^{\infty} (e^{f(x)} - 1) \lambda F(dx) \right) \end{aligned} \quad (17)$$

be a martingale. Define $d\mathbb{Q}|_{\mathcal{F}_t} = L_t d\mathbb{P}|_{\mathcal{F}_t}$. Then,

$$\int_0^t \int_{\mathbb{R}} (\mu(ds, dx) - ds e^{f(x)} \lambda F(dx)) \quad (18)$$

is a $\mathbb{Q}$ -martingale as obtained in the article on Poisson processes (see **Poisson Process**).

Poisson Point Processes

Poisson Measures

Let $(E, \mathcal{E})$ be a measurable space. A random measure μ on $(E, \mathcal{E})$ is a Poisson measure with intensity ν , where ν is a σ -finite measure on $(E, \mathcal{E})$, if

1. for every set $B \in \mathcal{E}$ with $\nu(B) < \infty$, $\mu(B)$ follows a Poisson distribution with parameter $\nu(B)$;
2. for disjoint sets $B_i, i \leq n$, the variables $\mu(B_i), i \leq n$ are independent.

Point Processes

Let $(E, \mathcal{E})$ be a measurable space and δ an additional point. We set $E_\delta = E \cup \delta$, $\mathcal{E}_\delta = \sigma(\mathcal{E}, \{\delta\})$.

Definition 1 Let $\mathbf{e}$ be a stochastic process defined on a probability space $(\Omega, \mathcal{F}, P)$, taking values in $(E_\delta, \mathcal{E}_\delta)$. The process $\mathbf{e}$ is a point process if

1. the map $(t, \omega) \rightarrow \mathbf{e}_t(\omega)$ is $\mathcal{B}([0, \infty[) \otimes \mathcal{F}$ -measurable;
2. the set $D_\omega = \{t : \mathbf{e}_t(\omega) \neq \delta\}$ is a.s. countable.

For every measurable set B of $]0, \infty[\times E$, we set

$$N^B(\omega) := \sum_{s \geq 0} \mathbb{1}_B(s, \mathbf{e}_s(\omega)) \quad (19)$$

In particular, if $B =]0, t] \times \Gamma$, we write

$$N_t^\Gamma = N^B = \text{Card}\{s \leq t : \mathbf{e}(s) \in \Gamma\} \quad (20)$$

Poisson Point Processes

Definition 2 An $\mathbf{F}$ -Poisson point process $\mathbf{e}$ is a point process such that

1. $N_t^E < \infty$ a.s. for every t
2. for any $\Gamma \in \mathcal{E}$, the process N^Γ is $\mathbf{F}$ -adapted
3. for any s and t and any $\Gamma \in \mathcal{E}$, $N_{s+t}^\Gamma - N_t^\Gamma$ is independent from $\mathcal{F}_t$ and distributed as N_s^Γ .

In particular, for any disjoint family $(\Gamma_i, i = 1, \dots, d)$, the d -dimensional process $(N_t^{\Gamma_i}, i = 1, \dots, d)$ is a Poisson process.

Definition 3 The σ -finite measure on $\mathcal{E}$ defined by

$$\mathbf{n}(\Gamma) = \frac{1}{t} \mathbb{E}(N_t^\Gamma) \quad (21)$$

is called the characteristic measure of $\mathbf{e}$.

If $\mathbf{n}(\Gamma) < \infty$, the process $N_t^\Gamma - t\mathbf{n}(\Gamma)$ is an $\mathbf{F}$ -martingale.

Proposition 1 (Compensation Formula).

Let H be a measurable positive process vanishing at δ . Then

$$\begin{aligned} & \mathbb{E} \left[\sum_{s \geq 0} H(s, \omega, \mathbf{e}_s(\omega)) \right] \\ &= \mathbb{E} \left[\int_0^\infty ds \int H(s, \omega, u) \mathbf{n}(du) \right] \end{aligned} \quad (22)$$

If, for any t , $\mathbb{E} \left[\int_0^t ds \int H(s, \omega, u) \mathbf{n}(du) \right] < \infty$, the process

$$\sum_{s \leq t} H(s, \omega, \mathbf{e}_s(\omega)) - \int_0^t ds \int H(s, \omega, u) \mathbf{n}(du) \quad (23)$$

is a martingale.

Proposition 2 (Exponential Formula).

If f is a measurable function such that $\int_0^t ds \int |f(s, u)| \mathbf{n}(du) < \infty$ for every t , then,

$$\begin{aligned} & \mathbb{E} \left[\exp \left(i \sum_{0 < s \leq t} f(s, \mathbf{e}_s) \right) \right] \\ &= \exp \left(\int_0^t ds \int (e^{if(s, u)} - 1) \mathbf{n}(du) \right) \end{aligned} \quad (24)$$

Moreover, if $f \geq 0$,

$$\begin{aligned} & \mathbb{E} \left[\exp \left(- \sum_{0 < s \leq t} f(s, \mathbf{e}_s) \right) \right] \\ &= \exp \left(- \int_0^t ds \int (1 - e^{-f(s, u)}) \mathbf{n}(du) \right) \end{aligned} \quad (25)$$

References

- [1] Cont, R. & Tankov, P. (2004). *Financial Modeling with Jump Processes*, Chapman & Hall/CRC.
- [2] Dellacherie, C. & Meyer, P.-A. (1980). *Probabilités et Potentiel, chapitres*, Hermann, Paris, Chapter V-VIII. English translation (1982), *Probabilities and Potential Chapters V-VIII*, North-Holland.

- [3] Jacod, J. & Shiryaev, A.N. (2003). *Limit Theorems for Stochastic Processes*, 2nd Edition, Springer Verlag.
- [4] Last, G. & Brandt, A. (1995). *Marked Point Processes on the Real Line. The Dynamic Approach*, Springer, Berlin.
- [5] Protter, P.E. (2005). *Stochastic Integration and Differential Equations*, 2nd Edition, Springer, Berlin.

Related Articles

Lévy Processes; Martingales; Martingale Representation Theorem.

MONIQUE JEANBLANC

Compensators

In probability theory, the *compensator* of a stochastic process designates a quantity that, once subtracted from a stochastic process, yields a martingale.

Compensator of a Random Time

Let $(\Omega, \mathbf{G}, \mathbb{P})$ be a filtered probability space and τ a $\mathbf{G}$ -stopping time. The process $H_t = \mathbb{1}_{\tau \leq t}$ is a $\mathbf{G}$ -adapted increasing process, hence a $\mathbf{G}$ -submartingale and admits a Doob–Meyer decomposition as

$$H_t = M_t + \Lambda_t \quad (1)$$

where M is a $\mathbf{G}$ -local martingale and Λ a $\mathbf{G}$ -predictable increasing process. The process Λ , called the *$\mathbf{G}$ -compensator* of H , is constant after τ , that is, $\Lambda_t = \Lambda_{t \wedge \tau}$. The process Λ “compensates” H with the meaning that $H - \Lambda$ is a martingale. If τ is $\mathbf{G}$ -predictable, then $\Lambda_t = H_t$. The continuity of Λ is equivalent to the fact that τ is a $\mathbf{G}$ -totally inaccessible stopping time. If Λ is absolutely continuous with respect to the Lebesgue measure, that is, if $\Lambda_t = \int_0^t \lambda_s^{\mathbf{G}} ds$, the nonnegative $\mathbf{G}$ -adapted process $\lambda^{\mathbf{G}}$ is called the *intensity rate* of τ . Note that $\lambda_t^{\mathbf{G}}$ is null on the set $\tau \leq t$.

For any integrable random variable $X \in \mathcal{G}_T$, one has

$$\mathbb{E}(X \mathbb{1}_{T < \tau} | \mathcal{G}_t) = \mathbb{1}_{\{t < \tau\}} (V_t - \mathbb{E}(\Delta V_\tau \mathbb{1}_{\tau \leq T} | \mathcal{G}_t)) \quad (2)$$

with $V_t = e^{\Lambda_t} \mathbb{E}(X e^{-\Lambda_\tau} | \mathcal{G}_t)$.

In the following examples, τ is a given random time, that is, a nonnegative random variable, and $\mathbf{H}$ the natural filtration of H (i.e., the smallest filtration satisfying the usual conditions such that the process H is adapted). The random time τ is a $\mathbf{H}$ -stopping time.

Elementary Case

Let τ be an exponential random variable with constant parameter λ . Then, the $\mathbf{H}$ -compensator of H is $\lambda(t \wedge \tau)$. More generally, if τ is a nonnegative random variable with cumulative distribution function F , taken continuous on the right ($F(t) = \mathbb{P}(\tau \leq t)$) such

that $\forall t, F(t) < 1$, the $\mathbf{H}$ -compensator of τ is $\Lambda_t = \int_0^{t \wedge \tau} \frac{dF(s)}{1-F(s-)}$. If F is continuous, the $\mathbf{H}$ -compensator is $\Lambda_t = -\ln(1 - F(t \wedge \tau))$.

Cox Processes

Let $\mathbf{F}$ be a given filtration, λ a given $\mathbf{F}$ -adapted nonnegative process, $\Lambda_t^{\mathbf{F}} = \int_0^t \lambda_s ds$, and Θ a random variable with exponential law, independent of $\mathbf{F}$. Let us define the random time τ as

$$\tau = \inf \{t : \Lambda_t^{\mathbf{F}} \geq \Theta\} \quad (3)$$

Then, the process

$$\mathbb{1}_{\tau \leq t} - \int_0^{t \wedge \tau} \lambda_s ds = \mathbb{1}_{\tau \leq t} - \Lambda_{t \wedge \tau}^{\mathbf{F}} \quad (4)$$

is a martingale in the filtration $\mathbf{G} = \mathbf{F} \vee \mathbf{H}$, the smallest filtration that contains $\mathbf{F}$, making τ a stopping time (in fact a totally inaccessible stopping time). The $\mathbf{G}$ -compensator of H is $\Lambda_t = \Lambda_{t \wedge \tau}^{\mathbf{F}}$, and the $\mathbf{G}$ -intensity rate is $\lambda_t^{\mathbf{G}} = \mathbb{1}_{t < \tau} \lambda_t$. In that case, for an integrable random variable $X \in \mathcal{F}_T$, one has

$$\mathbb{E}(X \mathbb{1}_{T < \tau} | \mathcal{G}_t) = \mathbb{1}_{t < \tau} e^{\Lambda_t^{\mathbf{F}}} \mathbb{E}(X e^{-\Lambda_t^{\mathbf{F}}} | \mathcal{F}_t) \quad (5)$$

and, for H , an $\mathbf{F}$ -predictable (bounded) process

$$\begin{aligned} \mathbb{E}(H_\tau \mathbb{1}_{\tau \leq T} | \mathcal{G}_t) &= H_\tau \mathbb{1}_{\tau \leq t} \\ &+ \mathbb{1}_{t < \tau} e^{\Lambda_t^{\mathbf{F}}} \mathbb{E} \left(\int_t^T H_s e^{-\Lambda_s^{\mathbf{F}}} \lambda_s ds | \mathcal{F}_t \right) \end{aligned} \quad (6)$$

Conditional Survival Probability

Assume now that τ is a nonnegative random variable on the filtered probability space $(\Omega, \mathbf{F}, \mathbb{P})$ with conditional survival probability $G_t := \mathbb{P}(\tau > t | \mathcal{F}_t)$, taken continuous on the right and let $\mathbf{G} = \mathbf{F} \vee \mathbf{H}$. The random time τ is a $\mathbf{G}$ -stopping time.

If τ is an $\mathbf{F}$ -predictable stopping time (hence a $\mathbf{G}$ -predictable stopping time), then $G_t = \mathbb{1}_{\tau > t}$ and $\Lambda = H$.

In what follows, we assume that $G_t > 0$ and we introduce the Doob–Meyer decomposition of the $\mathbf{F}$ -supermartingale G , that is, $G_t = Z_t - A_t$, where Z is an $\mathbf{F}$ -martingale and A is an increasing $\mathbf{F}$ -predictable process. Then, the $\mathbf{G}$ -compensator of τ is $\Lambda_t = \int_0^{t \wedge \tau} (G_{s-})^{-1} dA_s$. If $dA_t = a_t dt$, the $\mathbf{G}$ -intensity rate is $\lambda_t^{\mathbf{G}} = \mathbb{1}_{t < \tau} (G_{t-})^{-1} a_t$. Moreover, if G

is continuous, then for an integrable random variable $X \in \mathcal{F}_T$, one has

$$\mathbb{E}(X \mathbb{1}_{T < \tau} | \mathcal{G}_t) = \mathbb{1}_{t < \tau} (G_t)^{-1} \mathbb{E}(X G_T | \mathcal{F}_t) \quad (7)$$

It is often convenient to introduce the $\mathbf{F}$ -adapted process $\tilde{\lambda}_t = (G_t)^{-1} a_t$, equal to $\lambda_t^{\mathbf{G}}$ on the set $t < \tau$. We shall call this process the *predefault-intensity* process.

A particular case occurs when the process G is nonincreasing and absolutely continuous with respect to the Lebesgue measure, that is, $G_t = \int_t^\infty g_s ds$, where $g \geq 0$. In that case, the $\mathbf{G}$ -adapted intensity rate is $\lambda_t^{\mathbf{G}} = (G_t)^{-1} g_t \mathbb{1}_{t < \tau}$, the predefault intensity is $\tilde{\lambda}_t = (G_t)^{-1} g_t$ and, for an integrable random variable $X \in \mathcal{F}_T$,

$$\mathbb{E}(X \mathbb{1}_{T < \tau} | \mathcal{G}_t) = \mathbb{1}_{t < \tau} e^{\Lambda_t^{\mathbf{F}}} \mathbb{E}(X e^{-\Lambda_t^{\mathbf{F}}} | \mathcal{F}_t) \quad (8)$$

where $\Lambda^{\mathbf{F}}$ is the $\mathbf{F}$ -adapted process defined as

$$\Lambda_t^{\mathbf{F}} = \int_0^t \tilde{\lambda}_s ds = \int_0^t (G_s)^{-1} g_s ds \quad (9)$$

Aven's Lemma

The Aven lemma has the following form: let $(\Omega, \mathcal{G}_t, \mathbb{P})$ be a filtered probability space and N be a counting process. Assume that $E(N_t) < \infty$ for any t . Let $(h_n, n \geq 1)$ be a sequence of real numbers converging to 0, and

$$Y_t^{(n)} = \frac{1}{h_n} E(N_{t+h_n} - N_t | \mathcal{G}_t) \quad (10)$$

Assume that there exists λ_t and y_t nonnegative $\mathbb{G}$ -adapted processes such that

1. For any t , $\lim Y_t^{(n)} = \lambda_t$ (11)
2. For any t , there exists for almost all ω an $n_0 = n_0(t, \omega)$ such that

$$|Y_s^{(n)} - \lambda_s(\omega)| \leq y_s(\omega), \quad s \leq t, \quad n \geq n_0(t, \omega) \quad (12)$$

3. $\int_0^t y_s ds < \infty, \forall t$ (13)

Then, $N_t - \int_0^t \lambda_s ds$ is a $\mathbb{G}$ -martingale.

For the particular case of a random time, we obtain the following: assume that $\lim_{h \rightarrow 0} \frac{1}{h} \mathbb{P}(t < \tau \leq t +$

$h | \mathcal{G}_t) = \lambda_t^{\mathbf{G}}$, and that, there exists a Lebesgue integrable process y such that $|\frac{1}{h} \mathbb{P}(t < \tau \leq t + h | \mathcal{G}_t) - \lambda_t^{\mathbf{G}}| \leq y_t$ for any h small enough. Then $\lambda^{\mathbf{G}}$ is the $\mathbf{G}$ -intensity of τ .

In the case of conditional survival probability model, the predefault intensity $\tilde{\lambda}^{\mathbf{G}}$ is

$$\tilde{\lambda}_t^{\mathbf{G}} = \lim_{h \rightarrow 0} \frac{1}{h \mathbb{P}(t < \tau | \mathcal{F}_t)} \mathbb{P}(t < \tau \leq t + h | \mathcal{F}_t) \quad (14)$$

See [2] for an extensive study.

Shrinking

Assume that $\mathbf{G}^*$ is a subfiltration of $\mathbf{G}$ such that τ is a $\mathbf{G}^*$ (and a $\mathbf{G}$) stopping time. Assume that τ admits a $\mathbf{G}$ -intensity rate equal to $\lambda^{\mathbf{G}}$. Then, the $\mathbf{G}^*$ -intensity of τ is $\lambda_t^* = \mathbb{E}(\lambda_t^{\mathbf{G}} | \mathcal{F}_t^*)$ (see [1]).

As we have seen above, in the survival probability approach, the value of the intensity can be given in terms of the conditional survival probability. Assume that $G_t = \mathbb{P}(\tau > t | \mathcal{F}_t) = Z_t - \int_0^t a_s ds$, where Z is an $\mathbf{F}$ -martingale and that $\mathbf{G}^* = \mathbf{F}^* \vee \mathbf{H}$ where, $\mathbf{F}^* \subset \mathbf{F}$. Then, the $\mathbf{F}^*$ -conditional survival probability of τ is

$$G_t^* = \mathbb{P}(\tau > t | \mathcal{F}_t^*) = \mathbb{E}(G_t | \mathcal{F}_t^*) = X_t^* - \int_0^t a_s^* ds \quad (15)$$

where X^* is an $\mathbf{F}^*$ -martingale and $a_s^* = \mathbb{E}(a_s | \mathcal{F}_s^*)$. It follows that the $\mathbf{G}^*$ -intensity rate of τ writes as (we assume, for simplicity, that G and G^* are continuous)

$$\lambda_t^* = \mathbb{1}_{t < \tau} \frac{a_t^*}{G_t^*} = \mathbb{1}_{t < \tau} \frac{\mathbb{E}(\lambda_t^{\mathbf{G}} G_t | \mathcal{F}_t^*)}{\mathbb{E}(G_t | \mathcal{F}_t^*)} \quad (16)$$

It is useful to note that one can start with a model in which τ is an $\mathbf{F}$ -predictable stopping time (hence $\mathbf{G} = \mathbf{F}$, and a $\mathbf{G}$ -intensity rate does not exist) and consider a smaller filtration (e.g., the trivial filtration) for which there exists an intensity rate, computed by means of the conditional survival probability.

Compensator of an Increasing Process

The notion of interest in this section is that of *dual predictable projection*, which we define as follows:

Proposition 1 *Let A be an integrable increasing process (not necessarily $\mathbf{F}$ -adapted). There*

exists a unique $\mathbf{F}$ -predictable increasing process $(A_t^{(p)}, t \geq 0)$, called the $\mathbf{F}$ -dual predictable projection of A such that

$$\mathbb{E} \left(\int_0^\infty H_s dA_s \right) = \mathbb{E} \left(\int_0^\infty H_s dA_s^{(p)} \right) \quad (17)$$

for any positive $\mathbf{F}$ -predictable process H .

The definition of compensator of a random time can be interpreted in terms of dual predictable projection: if τ is a random time, the F -predictable compensator associated with τ is the dual predictable projection A^τ of the increasing process $\mathbb{1}_{\{\tau \leq t\}}$. It satisfies

$$\mathbb{E}(k_\tau) = \mathbb{E} \left(\int_0^\infty k_s dA_s^\tau \right) \quad (18)$$

for any positive, $\mathbf{F}$ -predictable process k .

Examples

Covariation Processes. Let M be a martingale and $[M]$ its quadratic variation process. If $[M]$ is integrable, its compensator is $\langle M \rangle$.

Standard Poisson Process. If N is a Poisson process, $(M_t = N_t - \lambda t, t \geq 0)$ is a martingale, and λt is the compensator of N ; the martingale M is called the *compensated martingale*.

Compensated Poisson Integrals. Let N be a time inhomogeneous Poisson process with deterministic intensity λ and $\mathbf{F}^N$ its natural filtration. The process

$$\left(M_t = N_t - \int_0^t \lambda(s) ds, t \geq 0 \right) \quad (19)$$

is an $\mathbf{F}^N$ -martingale. The increasing function $\Lambda(t) := \int_0^t \lambda(s) ds$ is called the (deterministic) *compensator* of N .

Random Measures

Definitions

The *compensator* of a random measure μ is the unique random measure ν such that

1. for every predictable process H , the process $(H \star \nu)$ is predictable (the measure ν is said to be predictable) and
2. for every predictable process H such that the process $|H| \star \mu$ is increasing and locally integrable, the process $(H \star \mu - H \star \nu)$ is a local martingale.

Examples

If N is a Lévy process with Lévy measure ν

$$\begin{aligned} & \int_{\Lambda} f(x) N_t(\cdot, dx) - t \int_{\Lambda} f(x) \nu(dx) \\ &= \sum_{0 < s \leq t} (f(\Delta X_s) \mathbb{1}_{\Lambda}(\Delta X_s)) - t \int_{\Lambda} f(x) \nu(dx) \end{aligned} \quad (20)$$

is a martingale, the compensator of $\int_{\Lambda} f(x) N_t(\cdot, dx)$ is $t \int_{\Lambda} f(x) \nu(dx)$.

For other examples see the article on point processes (see **Point Processes**).

References

- [1] Brémaud, P. & Yor, M. (1978). Changes of filtration and of probability measures, *Zeit Wahr and Verw Gebiete* **45**, 269–295.
- [2] Zeng, Y. (2006). *Compensators of Stopping Times*, PhD thesis, Cornell University.

Further Reading

- Brémaud, P. (1981). *Point Processes and Queues. Martingale Dynamics*, Springer-Verlag, Berlin.
- Çınlar, E. (1975). *Introduction to Stochastic Processes*, Prentice Hall.
- Cont, R. & Tankov, P. (2004). *Financial Modeling with Jump Processes*, Chapman & Hall/CRC.
- Jeanblanc, M., Yor, M. & Chesney, M. (2009). *Mathematical Models for financial Markets*, Springer, Berlin.
- Karlin, S. & Taylor, H. (1975). *A First Course in Stochastic Processes*, Academic Press, San Diego.

Related Articles

Doob–Meyer Decomposition; Filtrations; Intensity-based Credit Risk Models; Point Processes.

MONIQUE JEANBLANC

Heavy Tails

The three most cited stylized properties attributed to log-returns of financial assets or stocks are (i) a kurtosis much larger than 3, the kurtosis of a normal distribution; (ii) serial dependence without correlation; and (iii) volatility clustering. Any realistic and useful model for log-returns must account for all three of these characteristics. In this article, the focus is on the large kurtosis property, which is indicative of heavy tails in the returns. Although this stylized fact may not draw the same level of attention as the other two, it can have a serious impact on modeling and inference questions related to financial time series. One such application is the estimation of the Value at Risk, which is an important entity in the finance industry. For example, financial institutions would like to estimate large quantiles of the absolute returns, that is, the level at which the probability that an absolute return exceeds this value is small such as 0.01 or less. The estimation of these large quantities is extremely sensitive to the shape assumed for the tail of the marginal distribution. A light-tailed assumption for the tails can severely underestimate the actual quantiles of the marginal distribution. In addition to Value at Risk, heavy tails can impact the estimation of key measures of dependencies in financial time series. This includes the sample autocorrelation of the time series and of functions of the time series such as absolute values and squares. Standard central limit theory for mixing sequences generally directly applies to the sample autocorrelation functions (ACFs) of a financial time series and its squares, provided the fourth and eight moments, respectively, are finite. If these moments are infinite, as well may be the case for financial time series, then the asymptotic behavior of the sample ACFs is often nonstandard. As it turns out, GARCH processes and stochastic volatility (SV) processes, which are the primary modeling engines for financial returns, exhibit heavy tails in the marginal distribution. We focus on heavy tails and how the concept of regular variation plays a vital role in both these processes.

It is often a misconception to associate heavy-tailed distributions with a very large variance. Rather, the term is used to describe data that exhibit bursts of outlying observations. These outlying observations could be orders of magnitude larger than the median

of the observations. In the early 1960s, Mandelbrot (see **Mandelbrot, Benoit**) [31], Mandelbrot and Taylor [32], and Fama [21] realized that the marginal distribution of returns appeared to be heavy tailed. To cope with heavy tails, they considered non-Gaussian stable distributions for the marginals. Since this class of distributions has infinite variance, it was a slightly controversial approach. On the other hand, for many financial time series, there is evidence that the marginal distribution may have a finite variance but an infinite fourth moment. Figure 1 contains two financial time series that exhibit heavy tails. Figure 1(a) consists of the daily pound/US dollar exchange rate from October 1, 1981 to June 28, 1985, while Figure 1(b) displays the log-returns of the daily closing price of Merck stock from January 2, 2003 through April 28, 2006. One can certainly detect the occasional bursts of outlying observations in both series that are representative of heavy tails. As described in the second section (see Figure 3c and d), there is statistical evidence that the tail behavior of the marginal distribution is heavy with possibly infinite fourth moments.

Regular variation is a natural and often used concept to describe and model heavy-tailed phenomena. Many processes that are designed to model financial time series, such as the GARCH and heavy-tailed SV processes, have the property that all finite-dimensional distributions are regularly varying. For such processes, one can apply standard results from extreme value theory for establishing limiting behavior of the extremes of the process, the sample ACF of the process and its squares, and a host of other statistics. The regular variation condition and its properties are described in the second section. In the third section, some of the main results on regular variation for GARCH and SV processes, respectively, are described. The fourth section describes some of the applications of the regular variation conditions mentioned in the third section, with emphasis on extreme values, point processes, and sample autocorrelations.

Regular Variation

Multivariate regular variation plays an indispensable role in extreme value theory and often serves as the starting point for modeling multivariate extremes. In some respect, one can regard a random vector that is regularly varying as the heavy-tailed analog

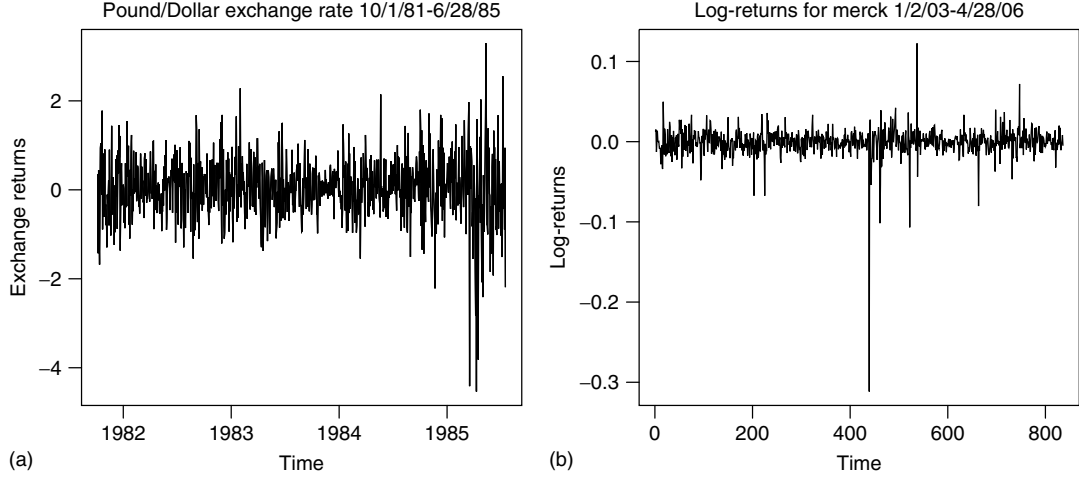


Figure 1 Log-returns for US/pound exchange rate, October 1, 1981 to June 28, 1985 (a) and log-returns for closing price of Merck stock, January 2, 2003 to April 28, 2006 (b)

of a Gaussian random vector. Unlike a Gaussian random vector, which is characterized by the mean vector and all pairwise covariances, a regular varying random vector in d dimensions is characterized by two components, an index $\alpha > 0$ and a random vector Θ with values in $\mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1}$ denotes the unit sphere in $\mathbb{R}^d$ with respect to the norm $|\cdot|$. The random vector $\mathbf{X}$ is said to be *regularly varying* with index $-\alpha$ if for all $t > 0$,

$$\frac{P(|\mathbf{X}| > tu, \mathbf{X}/|\mathbf{X}| \in \cdot)}{P(|\mathbf{X}| > u)} \xrightarrow{v} t^{-\alpha} P(\Theta \in \cdot) \quad (1)$$

as $u \rightarrow \infty$

The symbol $\xrightarrow{v}$ stands for vague convergence on $\mathbb{S}^{d-1}$; vague convergence of measures is treated in detail in [27]. See [24, 36, 37] for background on multivariate regular variation. In this context, the convergence in equation (1) holds for all continuity sets $A \in \mathcal{B}(\mathbb{S}^{d-1})$ of Θ . In particular, equation (1) implies that the modulus of the random vector $|\mathbf{X}|$ is regularly varying, that is,

$$\lim_{u \rightarrow \infty} \frac{P(|\mathbf{X}| > tu)}{P(|\mathbf{X}| > u)} = t^{-\alpha} \quad (2)$$

Hence, roughly speaking, from the defining equation (1), the modulus and angular parts of the random vector, $|\mathbf{X}|$ and $\mathbf{X}/|\mathbf{X}|$, are *independent* in the limit,

that is,

$$P(\mathbf{X}/|\mathbf{X}| \in A | |\mathbf{X}| > u) \rightarrow P(\Theta \in A) \quad (3)$$

as $u \rightarrow \infty$

The distribution of Θ is often called the *spectral measure* of the regularly varying random vector. The modulus has power-law-like tails in the sense that

$$P(|\mathbf{X}| > x) = L(x)x^{-\alpha} \quad (4)$$

where $L(x)$ is a slowly varying function, that is, for any $t > 0$, $L(tx)/L(x) \rightarrow 1$ as $x \rightarrow \infty$. This property implies that the r th moments of $|\mathbf{X}|$ are infinite for $r > \alpha$ and finite for $r < \alpha$.

There is a second characterization of regular variation that is often useful in applications. Replacing u in equation (1) by the sequence $a_n > 0$ satisfying, $nP(|\mathbf{X}| > a_n) \rightarrow 1$ (i.e., we may take a_n to be the $1 - n^{-1}$ quantile of $|\mathbf{X}|$), we obtain

$$nP(|\mathbf{X}| > ta_n, \mathbf{X}/|\mathbf{X}| \in \cdot) \xrightarrow{v} t^{-\alpha} P(\Theta \in \cdot) \quad (5)$$

as $n \rightarrow \infty$

As expected, the multivariate regular variation condition collapses to the standard condition in the one-dimensional case $d = 1$. In this case, $\mathbb{S}^0 = \{-1, 1\}$, so that the random variable X is regular

varying if and only if $|X|$ is regularly varying

$$\lim_{u \rightarrow \infty} \frac{P(|X| > tu)}{P(|X| > u)} = t^{-\alpha} \quad (6)$$

and the tail balancing condition,

$$\begin{aligned} \lim_{u \rightarrow \infty} \frac{P(X > u)}{P(|X| > u)} &= p \quad \text{and} \\ \lim_{u \rightarrow \infty} \frac{P(X < -u)}{P(|X| > u)} &= q \end{aligned} \quad (7)$$

holds, where p and q are nonnegative constants with $p + q = 1$. The Pareto distribution, t -distribution, and nonnormal stable distributions are all examples of one-dimensional distributions that are regularly varying.

Example 1 (Independent components). Suppose that $\mathbf{X} = (X_1, X_2)'$ consists of two independent and identically distributed (i.i.d.) components, where X_1 is regularly varying random variable. The scatter plot of 10 000 replicates of these pairs, where X_1 has a t -distribution with 3 degrees of freedom, is displayed in Figure 2(a). The t -distribution is regularly varying, with index α being equal to the degrees of freedom. In this case, the spectral measure is a discrete distribution, which places equal mass at the intersection of

the unit circle and the coordinate axes. That is,

$$P\left(\Theta = \frac{\pi k}{2}\right) = \frac{1}{4} \quad \text{for } k = -1, 0, 1, 2 \quad (8)$$

The scatter plot in Figure 2 reflects the form of the spectral distribution. The points that are far from the origin occur only near the coordinate axes. The interpretation is that the probability that both components of the random vector are large at the same time is quite small.

Example 2 (Totally Dependent Components). In contrast to the independent case of Example 1, suppose that both components of the vector are identical, that is, $\mathbf{X} = (X, X)$, with X regularly varying in one dimension. Independent replicates of this random vector would just produce points lying on a 45° line through the origin. Here, it is easy to see that the vector is regularly varying with spectral measure given by

$$P\left(\Theta = \frac{\pi}{4}\right) = p \quad \text{and} \quad P\left(\Theta = \frac{-\pi}{4}\right) = q \quad (9)$$

Example 3 (AR(1) Process). Let $\{X_t\}$ be the AR(1) process defined by the recursion:

$$X_t = 0.9X_{t-1} + Z_t \quad (10)$$

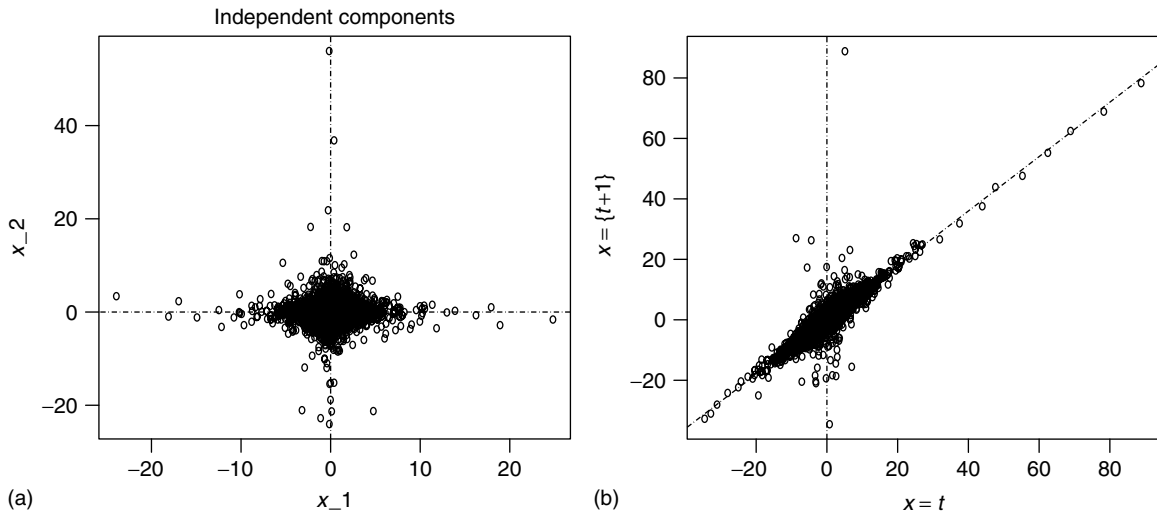


Figure 2 Scatter plot of 10 000 pairs of observations with i.i.d. components having a t -distribution with 3 degrees of freedom (a) and 10 000 observations of (X_t, X_{t+1}) from an AR(1) process (b)

where $\{Z_t\}$ is an i.i.d. sequence of random variables that have a symmetric stable distribution with exponent 1.8. This stable distribution is regularly varying with index $\alpha = 1.8$. Since $X_t = \sum_{j=0}^{\infty} 0.9^j Z_{t-j}$ is a linear process, it follows [14, 15] that X_t is also symmetric and regularly varying with index 1.8. In fact, X_t has a symmetric stable distribution with exponent 1.8 and scale parameter $(1 - 0.9^{1.8})^{-1/1.8}$. The scatter plot of consecutive observations (X_t, X_{t+1}) based on 10 000 observations generated from an AR(1) process is displayed in Figure 2(b). It can be shown that all finite-dimensional distributions of this time series are regularly varying. The spectral distribution of the vector consisting of two consecutive observations $\mathbf{X} = (X_t, X_{t+1})$ is given by

$$\begin{aligned} P(\Theta = \pm \arctan(0.9)) &= 0.9898 \quad \text{and} \\ P(\Theta = \pm \pi/2) &= 0.0102 \end{aligned} \quad (11)$$

As seen in Figure 2, one can see that most of the points in the scatter plot, especially those far from the origin, cluster tightly around the line through the origin with slope 0.9. This corresponds to the large mass at $\arctan(0.9)$ of the distribution of Θ . One can also detect a smattering of extreme points clustered around the vertical axis.

Estimation of α

A great deal of attention in the extreme value theory community has been devoted to the estimation of α in the regular variation condition (1). The generic Hill estimate is often a good starting point for this task. There are more sophisticated versions of Hill estimates, see [23] for a nice treatment of Hill estimators, but for illustration we stick with the standard version. For observations $X_1, \dots, X_n$ from a nonnegative-valued time series, let $X_{n:1} > \dots > X_{n:n}$ be the corresponding descending order statistics. If the data were in fact i.i.d. from a Pareto distribution, then the maximum likelihood estimator of α^{-1} based on the largest $m + 1$ order statistics is

$$\hat{\alpha}^{-1} = \frac{1}{m} \sum_{j=1}^m (\ln X_{n:j} - \ln X_{n:m+1}) \quad (12)$$

Different values of m produce an array of α estimates. The typical operating procedure is to plot the estimate of α versus m and choose a value

of m where the plot appears horizontal for an extended segment. See [7, 37] for other procedures for selecting m . There is the typical bias versus variance trade-off, with larger m producing smaller variance but larger bias. Figure 3 contains graphs of the Hill estimate of α as a function of m for the two simulated series in Figure 2 and the exchange rate and log-return data of Figure 1. In all cases, one can see a range of m for which the graph of $\hat{\alpha}$ is relatively flat. Using this segment as an estimate of α , we would estimate the index as approximately 3 for the two simulated series, approximately 3 for the exchange rate data, and around 3.5 for the stock price data. (The value of α for the two simulated series is indeed 3.) Also displayed on the plots are 95% confidence intervals for α , assuming the data are i.i.d. As suggested by these plots, the return data appear to have quite heavy tails.

Estimation of the Spectral Distribution

Using property (3), a naive estimate of the distribution of Θ is based on the angular components $\mathbf{X}_t/|\mathbf{X}_t|$ in the sample. One simply uses the empirical distribution of these angular pieces for which the modulus $|\mathbf{X}_t|$ exceeds some large threshold. More details can be found in [37]. For the scatter plots in Figure 2, we produced in Figure 4 kernel density estimates of the spectral density function for the random variable Θ on $(-\pi, \pi]$. One can see in the graph of the i.i.d. data, the large spikes at values of $\theta = -\pi, -\pi/2, 0, \pi/2, \pi$ corresponding to the coordinate axes (the values at $-\pi$ and π should be grouped together). On the other hand for the AR(1) process, the density estimate puts large mass at $\theta = \arctan(0.9)$ and $\theta = \arctan(0.9) - \pi$ corresponding to the line with slope 0.9 in the first and third quadrants, respectively. Since there are only a few points on the vertical axis, the density estimate does not register much mass at 0 and π .

Regular Variation for GARCH and SV Processes

GARCH Processes

The autoregressive conditional heteroscedastic (ARCH) process developed by Engle [19] and its generalized version, GARCH, developed by Engle

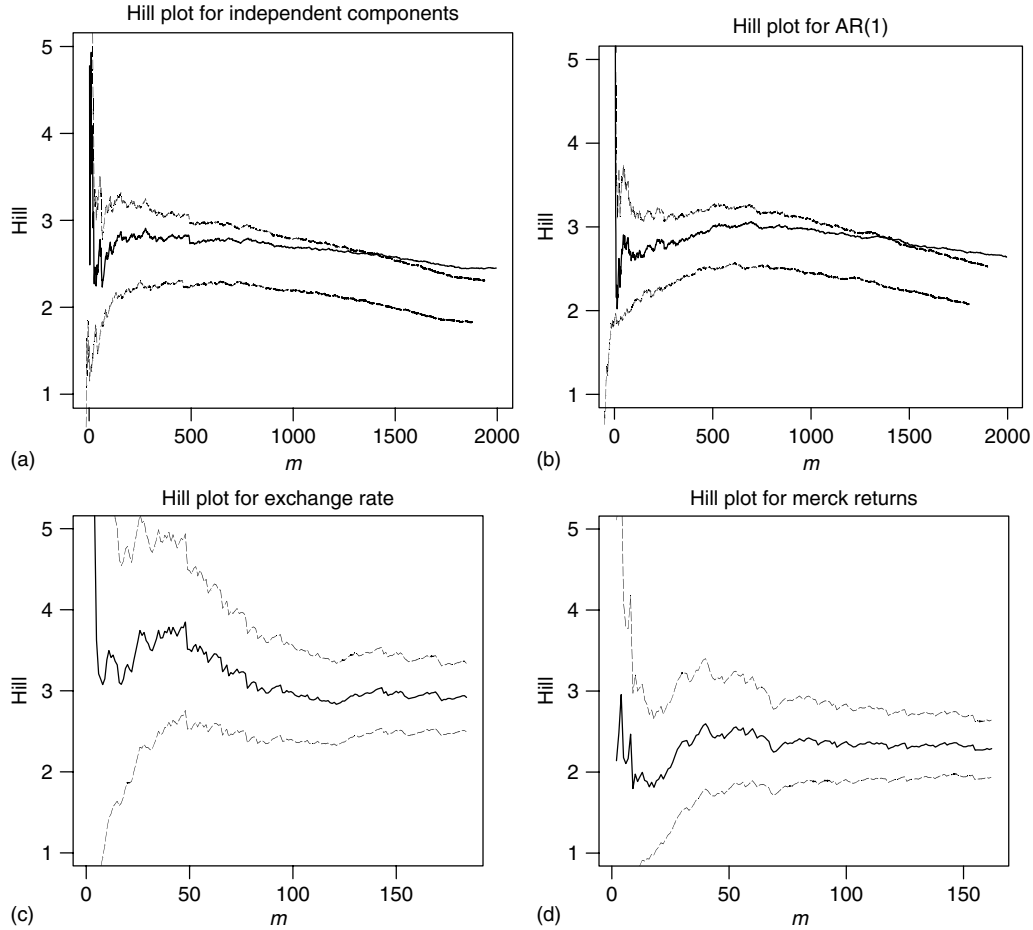


Figure 3 Hill plots for tail index: (a) i.i.d. data in Figure 2; (b) AR(1) process in Figure 2; (c) log-returns for US/pound exchange rate; and (d) log-returns for Merck stock, January 2, 2003 to April 28, 2006

and Bollerslev [20] are perhaps the most popular models for financial time series (see **GARCH Models**). Although there are many variations of the GARCH process, we focus on the traditional version. We say that $\{X_t\}$ is a GARCH(p, q) process if it is a strictly stationary solution of the equations:

$$\begin{aligned} X_t &= \sigma_t Z_t \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^p \alpha_i X_{t-i}^2 \\ &\quad + \sum_{j=1}^q \beta_j \sigma_{t-j}^2, \quad t \in \mathbb{Z} \end{aligned} \quad (13)$$

where the *noise* or *innovations* sequence $(Z_t)_{t \in \mathbb{Z}}$ is an i.i.d. sequence with mean zero and unit variance. It is usually assumed that all coefficients α_i and β_j are nonnegative, with $\alpha_0 > 0$. For identification purposes, the variance of the noise is assumed to be 1 since otherwise its standard deviation can be absorbed into σ_t . (σ_t) is referred to as the *volatility* sequence of the GARCH process.

The parameters are typically chosen to ensure that a causal and strictly stationary solution to the equations (13) exists. This means that X_t has a representation as a measurable function of the past and present noise values $Z_s, s \leq t$. The necessary and sufficient conditions for the existence and uniqueness of a stationary ergodic solution to equation (13) are

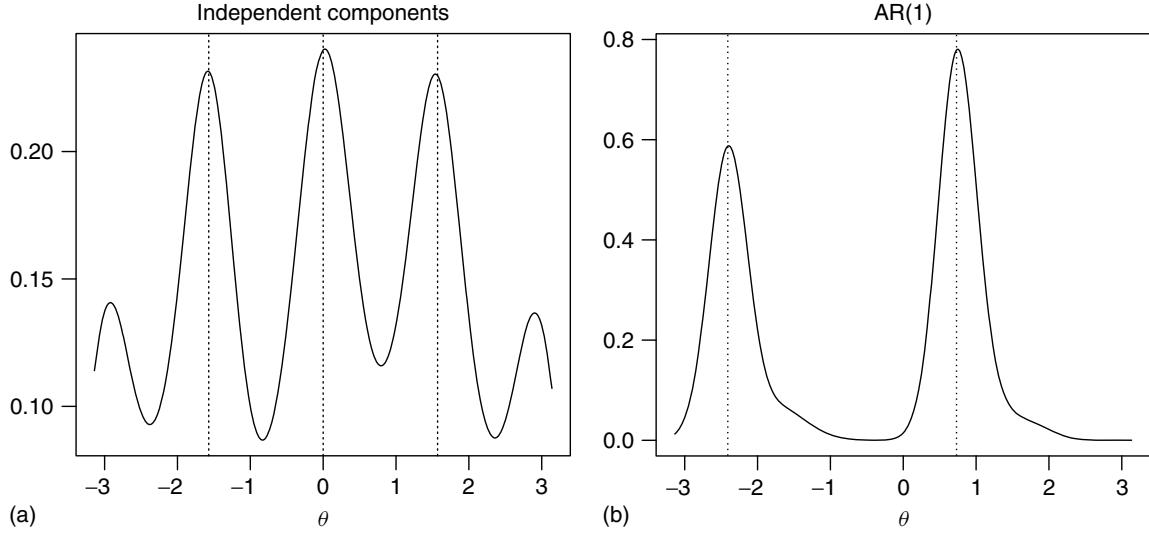


Figure 4 The estimation of the spectral density function for i.i.d. components (a) and for the AR(1) process (b) from Figure 2

given in [35] for the GARCH(1, 1) case and for the general GARCH(p, q) case in [4]; see [30] for a summary of the key properties of a GARCH process. In some cases, one only assumes weak stationarity, in which case the conditions on the parameters reduce substantially. A GARCH process is weakly stationary if and only if

$$\alpha_0 > 0 \quad \text{and} \quad \sum_{j=1}^p \alpha_j + \sum_{j=1}^q \beta_j < 1 \quad (14)$$

To derive properties of the tail of the finite-dimensional distributions of a GARCH process, including the marginal distribution, it is convenient to embed the squares X_t^2 and σ_t^2 in a stochastic recurrence equation (SRE). This embedding can be used to derive other key properties of the process beyond the finite-dimensional distributions. For example, conditions for stationarity and β -mixing can be established from the properties of SREs and general theory of Markov chains. Here, we focus on the tail behavior.

One builds an SRE by including the volatility process in the state vector. An SRE takes the form

$$\mathbf{Y}_t = \mathbf{A}_t \mathbf{Y}_{t-1} + \mathbf{B}_t \quad (15)$$

where $\mathbf{Y}_t$ is an m -dimensional random vector, $\mathbf{A}_t$ is an $m \times m$ random matrix, $\mathbf{B}_t$ is a random vector, and $\{(\mathbf{A}_t, \mathbf{B}_t)\}$ is an i.i.d. sequence. Under suitable conditions on the coefficient matrices and error matrices, one can derive various properties about the Markov chain $\mathbf{Y}_t$. For example, iteration of equation (15) yields a unique stationary and causal solution:

$$\mathbf{Y}_t = \mathbf{B}_t + \sum_{i=1}^{\infty} \mathbf{A}_t \cdots \mathbf{A}_{t-i+1} \mathbf{B}_{t-i}, \quad t \in \mathbb{Z} \quad (16)$$

To ensure almost surely (a.s.) convergence of the infinite series in equation (16), and hence the existence of a unique strictly stationary solution to equation (15), it is assumed that the *top Lyapunov exponent* given by

$$\gamma = \inf_{n \geq 1} n^{-1} E \log \|\mathbf{A}_n \cdots \mathbf{A}_1\| \quad (17)$$

is negative, where $\|\cdot\|$ is the operator norm corresponding to a given norm in $\mathbb{R}^m$.

Now, the GARCH process, at least its squares, can be embedded into an SRE by choosing

$$\mathbf{Y}_t = \begin{pmatrix} \sigma_{t+1}^2 \\ \vdots \\ \sigma_{t-q+2}^2 \\ X_t^2 \\ \vdots \\ X_{t-p+1}^2 \end{pmatrix}, \quad \mathbf{A}_t = \begin{pmatrix} \alpha_1 Z_t^2 + \beta_1 & \beta_2 & \cdots & \beta_{q-1} & \beta_q & \alpha_2 & \alpha_3 & \cdots & \alpha_p \\ 1 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & 0 & \cdots & 0 \\ Z_t^2 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

$$\mathbf{B}_t = (\alpha_0, 0, \dots, 0)' \quad (18)$$

where, as required, $\{(\mathbf{A}_t, \mathbf{B}_t)\}$ is an i.i.d. sequence. The top row in the SRE for the GARCH specification follows directly from the definition of the squared volatility process σ_{t+1}^2 and the property that $X_t = \sigma_t Z_t$.

In general, the top Lyapunov coefficient γ for the GARCH SRE cannot be calculated explicitly. However, a sufficient condition for $\gamma < 0$ is given as

$$\sum_{i=1}^p \alpha_i + \sum_{j=1}^q \beta_j < 1 \quad (19)$$

see p. 122 [4]. It turns out that this condition is also necessary and sufficient for the existence of a weakly stationary solution to the GARCH recursions. The solution will also be strictly stationary in this case.

It has been noted that for many financial time series, the GARCH(1,1) often provides an adequate model or is at least a good starter model. This is one of the few models where the Lyapunov coefficient can be computed explicitly. In this case, the SRE equation essentially collapses to the one-dimensional SRE given as

$$\sigma_{t+1}^2 = \alpha_0 + (\alpha_1 Z_t^2 + \beta_1) \sigma_t^2 = A_t \sigma_t^2 + \alpha_0 \quad (20)$$

where $A_t = \alpha_1 Z_t^2 + \beta_1$. The elements in the second row in the vector and matrix components of equation (18) play no role in this case. Hence,

$$\begin{aligned} \gamma &= n^{-1} E \log (A_n \cdots A_1) = E \log A_1 \\ &= E \log (\alpha_1 Z^2 + \beta_1) \end{aligned} \quad (21)$$

The conditions [35], $E \log(\alpha_1 Z^2 + \beta_1) < 0$ and $\alpha_0 > 0$, are necessary and sufficient for the existence of a stationary causal nondegenerate solution to the GARCH(1,1) equations.

Once the squares and volatility sequence, X_t^2 and σ_t^2 , respectively, are embedded in an SRE, then one can apply classical theory for SREs as developed by Kesten [28], (see also [22]), and extended by Basrak *et al.* [2], to establish regular variation of the tails of X_t^2 and σ_t^2 . The following result by Basrak *et al.* [1] summarizes the key results applied to a GARCH process.

Theorem 1 *Consider the process $(\mathbf{Y}_t)$ in equation (18) obtained from embedding a stationary GARCH process into the SRE (18). Assume that Z has a positive density on $\mathbb{R}$ such that $E(|Z|^h) < \infty$ for $h < h_0$ and $E(|Z|^{h_0}) = \infty$ for some $h_0 \in (0, \infty]$. Then with $\mathbf{Y} = \mathbf{Y}_1$, there exist $\alpha > 0$, a constant $c > 0$, and a random vector Θ on the unit sphere $\mathbb{S}^{p+q-2}$ such that*

$$x^{\alpha/2} P(|\mathbf{Y}| > x) \rightarrow c \quad \text{as } x \rightarrow \infty \quad (22)$$

and for every $t > 0$

$$\frac{P(|\mathbf{Y}| > tx, \mathbf{Y}/|\mathbf{Y}| \in \cdot)}{P(|\mathbf{Y}| > x)} \xrightarrow{w} t^{-\alpha/2} P(\Theta \in \cdot) \quad \text{as } x \rightarrow \infty \quad (23)$$

where $\xrightarrow{w}$ denotes weak convergence on the Borel σ -field of $\mathbb{S}^{p+q-2}$.^a

It follows that the components of the vector of $\mathbf{Y}$ are also regularly varying so that

$$\begin{aligned} P(|X_1| > x) &\sim c_1 x^{-\alpha} \quad \text{and} \\ P(\sigma_1 > x) &\sim c_2 x^{-\alpha} \end{aligned} \quad (24)$$

for some positive constants c_1 and c_2 . A straightforward application of Breiman's lemma [6], (cf. [13], Section 4), allows us to remove the absolute values in X_1 to obtain

$$\begin{aligned} P(X_1 > x) &= P(\sigma_1 Z_1^+ > x) \\ &\sim E((Z_1^+)^{\alpha}) P(\sigma_1 > x) \end{aligned} \quad (25)$$

$$\begin{aligned} P(X_1 \leq -x) &= P(-\sigma_1 Z_1^- \leq -x) \\ &\sim E((Z_1^-)^{\alpha}) P(\sigma_1 > x) \end{aligned} \quad (26)$$

where $Z_1^{\pm}$ are the respective positive and negative parts of Z_1 . With the exception of simple models such as the GARCH(1,1), there is no explicit formula for the index α of regular variation of the marginal distribution. In principle, α could be estimated from the data using a Hill style estimator, but an enormous sample size would be required in order to obtain a precise estimate of the index.

In the GARCH(1,1) case, α is found by solving the following equation:

$$E[(\alpha_1 Z^2 + \beta_1)^{\alpha/2}] = 1 \quad (27)$$

This equation can be solved for α by numerical and/or simulation methods for fixed values of α_1 and β_1 from the stationarity region of a GARCH(1,1) process and assuming a concrete density for Z . (See [12] for a table of values of α for various choices of α_1 and β_1 .) Note that in the case of an integrated GARCH (IGARCH) process where $\alpha_1 + \beta_1 = 1$, then we have $\alpha = 2$. This holds regardless of the distribution of Z_1 , provided it has a finite variance. Since the marginal distribution of an IGARCH process has Pareto-like tails with index 2, the variance is infinite.

While equations (25) and (26) describe only the regular variation of the marginal distribution, it is also true that the finite-dimensional distributions are regularly varying. To see this in the GARCH(1,1) case, we note that the volatility process is given as

$$\sigma_{t+1}^2 = (\alpha_1 Z_t^2 + \beta_1) \sigma_t^2 + \beta_0 \quad (28)$$

so that

$$\begin{aligned} (\sigma_1^2, \dots, \sigma_m^2) &= (1, \alpha_1 Z_1^2 + \beta_1, (\alpha_1 Z_2^2 + \beta_1) \\ &\quad \times (\alpha_1 Z_1^2 + \beta_1), \dots, \alpha_1 Z_{m-1}^2 + \beta_1) \cdots \\ &\quad \times (\alpha_1 Z_1^2 + \beta_1) \sigma_1^2 + \mathbf{R}_m \\ &= \mathbf{D}_m \sigma_1^2 + \mathbf{R}_m \end{aligned} \quad (29)$$

where $\mathbf{R}_m$ has tails that are lighter than those for σ_1^2 . Now since $\mathbf{D}_m = (D_1, \dots, D_m)$ is independent of σ_1^2 and has a $\alpha/2 + \delta$ moment for some $\delta > 0$, it follows by a generalization of Breiman's lemma [1] that

$$\mathbf{U}_m := (X_1^2, \dots, X_m^2) = \mathbf{F}_m \sigma_1^2 + \mathbf{R}_m \quad (30)$$

where $\mathbf{F}_m = (Z_1^2 D_1, \dots, Z_m^2 D_m)$ is regularly varying with

$$\begin{aligned} \lim_{x \rightarrow \infty} \frac{P(|\mathbf{U}_m| > x, \mathbf{U}_m/|\mathbf{U}_m| \in A)}{P(|\mathbf{U}_m| > x)} \\ &= \lim_{x \rightarrow \infty} \frac{P(|\mathbf{F}_m| \sigma_1^2 > x, \mathbf{F}_m/|\mathbf{F}_m| \in A)}{P(|\mathbf{F}_m| \sigma_1^2 > x)} \\ &= \frac{E(|\mathbf{F}_m|^{\alpha/2} I_A(\mathbf{F}_m/|\mathbf{F}_m|))}{E|\mathbf{F}_m|^{\alpha/2}} \end{aligned} \quad (31)$$

It follows that the finite-dimensional distributions of a GARCH process are regularly varying.

Stochastic Volatility Processes

The SV process also starts with the multiplicative model (13)

$$X_t = \sigma_t Z_t \quad (32)$$

with (Z_t) being an i.i.d. sequence of random variables. If $\text{var}(Z_t) < \infty$, then it is conventional to assume that Z_t has mean 0 and variance 1. Unlike the GARCH process, the volatility process (σ_t) for SV processes is assumed to be independent of the sequence (Z_t) . Often, one assumes that $\log \sigma_t^2$ is a linear Gaussian process given by

$$\log \sigma_t^2 = Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \eta_{t-j} \quad (33)$$

where (ψ_j) is a sequence of square summable coefficients and (η_t) is a sequence of i.i.d. $N(0, \sigma^2)$ random variables independent of (Z_t) . If $\text{var}(Z_t)$ is

finite and equal to 1, then the SV process $X_t = \sigma_t Z_t = \exp^{Y_t/2} Z_t$ is white noise with mean 0 and variance $\exp\{\mu + \sigma^2 \sum_{j=0}^{\infty} \psi_j^2/2\}$. One advantage of such processes is that one can explicitly compute the autocovariance function (ACVF) of any power of X_t and its absolute values. For example, the ACVF of the squares of (X_t) is, for $h > 0$, given as

$$\begin{aligned} \gamma_{|X|^2}(h) &= E(\exp\{Y_0 + Y_h\}) - (E \exp\{Y_0\})^2 \\ &= \exp\left\{2\mu + \sigma^2 \sum_{i=0}^{\infty} \psi_i^2\right\} \\ &\quad \times \left[\exp\left\{\sigma^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+h}\right\} - 1\right] \\ &= e^{2\mu} e^{\gamma_Y(0)} \left[e^{\gamma_Y(h)} - 1\right] \end{aligned} \quad (34)$$

Note that as $h \rightarrow \infty$,

$$\gamma_{|X|^2}(h) \sim e^{2\mu} e^{\gamma_Y(0)} \left[e^{\gamma_Y(h)} - 1\right] \sim e^{2\mu} e^{\gamma_Y(0)} \gamma_Y(h) \quad (35)$$

so that the ACVF of the SV for the squares converges to zero at the same rate as the log-volatility process.

If Z_t has a Gaussian distribution, then the tail of X_t remains light although a bit heavier than a Gaussian [3]. This is in contrast to the GARCH case where an i.i.d. Gaussian input leads to heavy-tailed marginals of the process. On the other hand, for SV processes, if the Z_t have heavy tails, for example, if Z_t has a t -distribution, then Davis and Mikosch [10] show that X_t is regularly varying. Furthermore, in this case, any finite collection of X_t 's has the same limiting joint tail behavior as an i.i.d. sequence with regularly varying marginals. Specifically, the two random vectors, $(X_1, \dots, X_k)'$ and $(E|\sigma_1|^\alpha)^{1/\alpha} (Z_1, \dots, Z_k)'$ have the same joint tail behavior.

Limit Theory GARCH and SV Processes

Convergence of Maxima

If (X_t) is a stationary sequence of random variables with common distribution function F , then often one can directly relate the limiting distribution of the maxima, $M_n = \max\{X_1, \dots, X_n\}$ to F . Assuming

that X_1 is regularly varying with index $-\alpha$ and choosing the sequence (a_n) such that $n(1 - F(a_n)) \rightarrow 1$, then

$$F^n(a_n x) \rightarrow G(x) = \begin{cases} 0, & x \leq 0 \\ e^{-x^{-\alpha}}, & x > 0 \end{cases} \quad (36)$$

This relation is equivalent to convergence in distribution of the maxima of the associated independent sequence $(\hat{X}_t)$ (i.e., the sequence $(\hat{X}_t)$ is i.i.d. with common distribution function F) normalized by a_n to the Fréchet distribution G . Specifically, if $\hat{M}_n = \max\{\hat{X}_1, \dots, \hat{X}_n\}$, then

$$P(a_n^{-1} M_n \leq x) \rightarrow G(x) \quad (37)$$

Under mild mixing conditions on the sequence (X_t) [29], we have

$$P(a_n^{-1} M_n \leq x) \rightarrow H(x) \quad (38)$$

with H a nondegenerate distribution function if and only if

$$H(x) = G^\theta(x) \quad (39)$$

for some $\theta \in (0, 1]$. The parameter θ is called the *extremal index* and can be viewed as a sample size adjustment for the maxima of the dependent sequence due to clustering of the extremes. The case $\theta = 1$ corresponds to no clustering, in which case the limiting behavior of M_n and $\hat{M}_n$ are identical. In case $\theta < 1$, M_n behaves asymptotically like the maximum of $n\theta$ independent observations. The reciprocal of the extremal index $1/\theta$ of a stationary sequence (X_t) also has the interpretation as the expected size of clusters of high-level exceedances in the sequence.

There are various sufficient conditions for ensuring that $\theta = 1$. Perhaps the most common anticlustering condition is D' [28], which has the following form:

$$\limsup_{n \rightarrow \infty} n \sum_{t=2}^{[n/k]} P(X_1 > a_n x, X_t > a_n x) = O(1/k) \quad (40)$$

as $k \rightarrow \infty$. Hence, if the stationary process (X_t) satisfies a mixing condition and D' , then

$$P(a_n^{-1} M_n \leq x) \rightarrow G(x) \quad (41)$$

Returning to the GARCH setting, we assume that the conditions of Theorem 1 are satisfied. Then we know that $P(|X| > x) \sim c_1 x^{-\alpha}$ for some $\alpha, c_1 > 0$, and we can even specify the value of α in the GARCH(1, 1) case by solving equation (27). Now choosing $a_n = n^{1/\alpha} c_1^{1/\alpha}$, we have $nP(|X_1| > a_n) \rightarrow 1$ and defining $M_n = \max\{|X_1|, \dots, |X_n|\}$, we obtain

$$P(a_n^{-1} M_n \leq x) \rightarrow \exp\{-\theta_1 x^{-\alpha}\} \quad (42)$$

where the extremal index θ_1 is strictly less than 1. Explicit formulae for the extremal index of a general GARCH process are hard to come by. In some special cases, such as the ARCH(1) and the GARCH(1,1), there are more explicit expressions. For example, in the GARCH(1,1) case, the extremal index θ_1 for the maxima of the absolute values of the GARCH process is given by Mikosch and Stărică [34]

$$\theta_1 = \frac{\lim_{k \rightarrow \infty} E \left(|Z_1|^\alpha - \max_{j=2, \dots, k+1} \left| Z_j^2 \prod_{i=2}^j A_i \right|^{\alpha/2} \right)}{E|Z_1|^\alpha} \quad (43)$$

The above expression can be evaluated by Monte-Carlo simulation, see, for example, [25] for the ARCH(1) case with standard normal noise Z_t ; see [18], Section 8.1, where one can also find some advice as to how the extremal index of a stationary sequence can be estimated from data.

The situation is markedly different for SV processes. For the SV process with either light- or heavy-tailed noise, one can show that D' is satisfied and hence the extremal index is always 1 (see [3] for the light-tailed case and [10] for the heavy-tailed case). Hence, although both GARCH and SV models exhibit stochastic clustering, only the GARCH process displays extremal clustering.

Convergence of Point Processes

The theory of point processes plays a central role in extreme value theory and in combination with regular variation can be a powerful tool for establishing limiting behavior of other statistics beyond extreme order statistics. As in the previous section, suppose that $(\hat{X}_t)$ is an i.i.d. sequence of nonnegative random variables with common distribution F that has

regularly varying tails with index $-\alpha$. Choosing the sequence a_n satisfying $n(1 - F(a_n)) \rightarrow 1$, we have

$$nP(\hat{X}_1 > a_n x) \rightarrow x^{-\alpha} \quad (44)$$

as $n \rightarrow \infty$. Now equation (44) can be strengthened to the statement

$$nP(a_n^{-1} \hat{X}_1 \in B) \rightarrow \nu(B) \quad (45)$$

for all suitably chosen Borel sets B , where the measure ν is defined by its value on intervals of the form $(a, b]$ with $a > 0$ as

$$\nu(a, b] = a^{-\alpha} - b^{-\alpha} \quad (46)$$

The convergence in equation (46) can be connected with the convergence in the distribution of a sequence of point processes. For a bounded Borel set B in $E = [0, \infty] \setminus \{0\}$, define the sequence of point processes $(\hat{N}_n)$ by

$$\hat{N}_n(B) = \# \{a_n^{-1} \hat{X}_j \in B, j = 1, 2, \dots, n\} \quad (47)$$

If B is the interval $(a, b]$ with $0 < a < b \leq \infty$, then since the $\hat{X}_j$ are i.i.d., $\hat{N}_n(B)$ has a binomial distribution with number of trials n and probability of success

$$p_n = P(a_n^{-1} \hat{X}_1 \in (a, b]) \quad (48)$$

It then follows from equation (46) that $\hat{N}_n(B)$ converges in distribution to a Poisson random variable $N(B)$ with mean $\nu(B)$. In fact, we have the stronger point process convergence:

$$\hat{N}_n \xrightarrow{d} N \quad (49)$$

where N is a Poisson process on E with mean measure $\nu(dx)$ and $\xrightarrow{d}$ denotes convergence in distribution of point processes. For our purposes, $\xrightarrow{d}$ for point processes means that for any collection of *bounded*^b Borel sets $B_1, \dots, B_k$ for which $P(N(\partial B_j) > 0) = 0$, $j = 1, \dots, k$, we have

$$(\hat{N}_n(B_1), \dots, \hat{N}_n(B_k)) \xrightarrow{d} (N(B_1), \dots, N(B_k)) \quad (50)$$

on $\mathbb{R}^k$ [18, 29, 36].

As an application of equation (49), define $\hat{M}_{n,k}$ to be the k th largest among $\hat{X}_1, \dots, \hat{X}_n$. For $y \leq x$, the event $\{a_n^{-1}\hat{M}_n \leq x, a_n^{-1}\hat{M}_{n,k} \leq y\} = \{\hat{N}_n(x, \infty) = 0, \hat{N}_n(y, x] \leq k-1\}$ and hence

$$\begin{aligned} P(a_n^{-1}\hat{M}_n \leq x, a_n^{-1}\hat{M}_{n,k} \leq y) \\ &= P(\hat{N}_n(x, \infty) = 0, \hat{N}_n(y, x] \leq k-1) \\ &\rightarrow P(N(x, \infty) = 0, N(y, x] \leq k-1) \\ &= e^{-x^{-\alpha}} \sum_{j=0}^{k-1} (y^{-\alpha} - x^{-\alpha})^j / j! \end{aligned} \quad (51)$$

As a second application of the limiting Poisson convergence in equation (49), the limiting Poisson process $\hat{N}$ has points located at $\Gamma_k^{-1/\alpha}$, where $\Gamma_k = E_1 + \dots + E_k$ is the sum of k i.i.d. unit exponentially distributed random variables. Then if $\alpha < 1$, the result is more complicated; if $\alpha \geq 1$, we obtain the convergence of partial sums:

$$a_n^{-1} \sum_{t=1}^n \hat{X}_t \xrightarrow{d} \sum_{j=0}^{\infty} \Gamma_j^{-1/\alpha} \quad (52)$$

In other words, the sum of the points of the point process N_n converges in distribution to the sum of points in the limiting Poisson process.

For a stationary time series (X_t) with heavy tails that satisfy a suitable mixing condition, such as strong mixing, and the anticlustering condition D' , then the convergence in equation (49) remains valid, as well as the limit in equation (52), at least for positive random variables. For example, this is the case for SV processes. If the condition D' is replaced by the assumption that all finite-dimensional random variables are regularly varying, then there is a point convergence result for N_n corresponding to (X_t) . However, the limit point process in this case is more difficult to describe. Essentially, the point process has anchors located at the Poisson points $\Gamma_j^{-1/\alpha}$. At each of these anchor locations, there is an independent cluster of points that can be described by the distribution of the angular measures in the regular variation condition [8, 9]. These conditions can then be applied to functions of the data, such as lagged products, to establish the convergence in distribution of the sample autocovariance function. This is the subject of the following section.

The Behavior of the Sample Autocovariance and Autocorrelation Functions

The ACF is one of the principal tools used in classical time series modeling. For a stationary Gaussian process, the dependence structure of the process is completely determined by the ACF. The ACF also conveys important dependence information for linear process. To some extent, the dependence governed by a linear filter can be fully recovered from the ACF. For the time series consisting of financial returns, the data are uncorrelated, so the value of the ACF is substantially diminished. Nevertheless, the ACF of other functions of the process such as the squares and absolute values can still convey useful information about the nature of the nonlinearity in the time series. For example, slow decay of the ACF of the squares is consistent with the volatility clustering present in the data. For a stationary time series (X_t) , the ACVF and ACF are defined as

$$\begin{aligned} \gamma_X(h) &= \text{cov}(X_0, X_h) \quad \text{and} \\ \rho_X(h) &= \text{corr}(X_0, X_h) = \frac{\gamma_X(h)}{\gamma_X(0)}, \quad h \geq 0 \end{aligned} \quad (53)$$

respectively. Now for observations $X_1, \dots, X_n$ from the stationary time series, the ACVF and ACF are estimated by their sample counterparts, namely, by

$$\hat{\gamma}_X(h) = \frac{1}{n} \sum_{t=1}^{n-h} (X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n) \quad (54)$$

and

$$\hat{\rho}_X(h) = \frac{\hat{\gamma}_X(h)}{\hat{\gamma}_X(0)} = \frac{\sum_{t=1}^{n-h} (X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n)}{\sum_{t=1}^n (X_t - \bar{X}_n)^2} \quad (55)$$

where $\bar{X}_n = n^{-1} \sum_{t=1}^n X_t$ is the sample mean.

Even though the sample ACVF is an average of random variables, its asymptotic behavior is determined by the extremes values, at least in the case of heavy-tailed data. Regular variation and point process theory are the two ingredients that play a key role in deriving limit theory for the sample ACVF and ACF. In particular, one applies the point process techniques alluded to in the previous section to the

stationary process consisting of products $(X_t X_{t+h})$. The first such results were established by Davis and Resnick [14–16] in a linear process setting. Extensions by Davis and Hsing [8] and Davis and Mikosch [9] allowed one to consider more general time series models beyond those linear. The main idea is to consider a point process N_n based on products of the form $X_t X_{t+h}/a_n^2$. After establishing convergence of this point process, in many cases one can apply the continuous mapping theorem to show that the sum of the points that comprise N_n converges in distribution to the sum of the points that make up the limiting point process. Although the basic idea for establishing these results is rather straightforward, the details are slightly complex. These ideas have been applied to the case of GARCH processes in [1] and to SV processes in [10], which are summarized below.

The GARCH Case

The scaling in the limiting distribution for the sample ACF depends on the index of regular variation α specified in Theorem 1. We summarize the results for the various cases of α .

1. If $\alpha \in (0, 2)$, then $\hat{\rho}_X(h)$ and $\hat{\rho}_{|X|}(h)$ have nondegenerate limit distributions. The same statement holds for $\hat{\rho}_{X^2}(h)$ when $\alpha \in (0, 4)$.
2. If $\alpha \in (2, 4)$, then both $\hat{\rho}_X(h)$, $\hat{\rho}_{|X|}(h)$ converge in probability to their deterministic counterparts $\rho_X(h)$, $\rho_{|X|}(h)$, respectively, at the rate $n^{1-2/\alpha}$ and the limit distribution is a complex function of non-Gaussian stable random variables.
3. If $\alpha \in (4, 8)$, then

$$n^{1-4/(2\alpha)}(\hat{\rho}_{X^2}(h) - \rho_{X^2}(h)) \xrightarrow{d} S_{\alpha/2}(h) \quad (56)$$

where the random variable $S_{\alpha/2}(h)$ is a function of infinite variance stable random variables.

4. If $\alpha > 4$, then the one can apply standard central limit theorems for stationary mixing sequences to establish a limiting normal distribution [17, 26]. In particular, $(\hat{\rho}_X(h))$ and $(\hat{\rho}_{|X|}(h))$ have Gaussian limits at $\sqrt{n}$ -rates. The corresponding result holds for (X_t^2) when $\alpha > 8$.

These results show that the limit theory for the sample ACF of a GARCH process is rather complicated when the tails are heavy. In fact, there is considerable empirical evidence based on extreme

value statistics as described in the second section, indicating that log-return series might not have a finite fourth or fifth moment^c and then the limit results above would show that the usual confidence bands for the sample ACF based on the central limit theorem and the corresponding $\sqrt{n}$ -rates are far too optimistic in this case.

The Stochastic Volatility Case

For a more direct comparison with the GARCH process, we choose a distribution for the noise process that matches the power law tail of the GARCH with index α . Then

$$\left(\frac{n}{\ln n}\right)^{1/\alpha} \hat{\rho}_X(h) \quad \text{and} \quad \left(\frac{n}{\ln n}\right)^{1/(2\alpha)} \hat{\rho}_{X^2}(h) \quad (57)$$

converge in distribution for $\alpha \in (0, 2)$ and $\alpha \in (0, 4)$, respectively. This illustrates the excellent large sample behavior of the sample ACF for SV models even if ρ_X and ρ_{X^2} are not defined [11, 13]. Thus, even if $\text{var}(Z_t) = \infty$ or $E Z_t^4 = \infty$, the estimates $\hat{\rho}_X(h)$ and $\hat{\rho}_{X^2}(h)$, respectively, converge to zero at a rapid rate. This is in marked contrast with the situation for GARCH processes, where under similar conditions on the marginal distribution, the respective sample ACFs converge in distribution to random variables without any scaling.

End Notes

^aBasrak *et al.* [1] proved this result under the condition that $\alpha/2$ is not an even integer. Boman and Lindskog [5] removed this condition.

^bHere bounded means bounded away from zero.

^cSee, for example, [18], Chapter 6, and [33].

References

- [1] Basrak, B., Davis, R.A. & Mikosch, T. (2002). Regular variation of GARCH processes, *Stochastic Processes and Their Applications* **99**, 95–116.
- [2] Basrak, B., Davis, R.A. & Mikosch, T. (2002). A characterization of multivariate regular variation, *The Annals of Applied Probability* **12**, 908–920.
- [3] Breidt, F.J. & Davis, R.A. (1998). Extremes of stochastic volatility models, *The Annals of Applied Probability* **8**, 664–675.
- [4] Bougerol, P. & Picard, N. (1992). Stationarity of GARCH processes and of some nonnegative time series, *Journal of Econometrics* **52**, 115–127.

- [5] Boman, J. & Lindsog, F. (2007). *Support Theorems for the Radon Transform and Cramér-Wold Theorems*. Technical report, KTH, Stockholm.
- [6] Breiman, L. (1965). On some limit theorems similar to the arc-sin law, *Theory of Probability and Its Applications* **10**, 323–331.
- [7] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer, London.
- [8] Davis, R.A. & Hsing, T. (1995). Point process and partial sum convergence for weakly dependent random variables with infinite variance, *Annals of Probability* **23**, 879–917.
- [9] Davis, R.A. & Mikosch, T. (1998). The sample autocorrelations of heavy-tailed processes with applications to ARCH, *Annals of Statistics* **26**, 2049–2080.
- [10] Davis, R.A. & Mikosch, T. (2001). Point process convergence of stochastic volatility processes with application to sample autocorrelation, *Journal of Applied Probability* **38A**, 93–104.
- [11] Davis, R.A. & Mikosch, T. (2001). The sample autocorrelations of financial time series models, in W.J. Fitzgerald, R.L. Smith, A.T. Walden & P.C. Young, (eds), *Nonlinear and Nonstationary Signal Processing*, Cambridge University Press, Cambridge, pp. 247–274.
- [12] Davis, R.A. & Mikosch, T. (2009). Extreme value theory for GARCH processes, in *Handbook of Financial Time Series*, T. Andersen, R.A. Davis, J.-P. Kreiss & T. Mikosch, eds, Springer, New York, pp. 187–200.
- [13] Davis, R.A. & Mikosch, T. (2009). Probabilistic properties of stochastic volatility models, in T. Andersen, R.A. Davis, J.-P. Kreiss & T. Mikosch, (eds), *Handbook of Financial Time Series*, Springer, New York, pp. 255–267.
- [14] Davis, R.A. & Resnick, S.I. (1985). Limit theory for moving averages of random variables with regularly varying tail probabilities, *Annals of Probability* **13**, 179–195.
- [15] Davis, R.A. & Resnick, S.I. (1985). More limit theory for the sample correlation function of moving averages, *Stochastic Processes and Their Applications* **20**, 257–279.
- [16] Davis, R.A. & Resnick, S.I. (1986). Limit theory for the sample covariance and correlation functions of moving averages, *Annals of Statistics* **14**, 533–558.
- [17] Doukhan, P. (1994). *Mixing Properties and Examples*, Lecture Notes in Statistics, Springer Verlag, New York. Vol. 85.
- [18] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.
- [19] Engle, R.F. (1982). Autoregressive conditional heteroscedastic models with estimates of the variance of United Kingdom inflation, *Econometrica* **50**, 987–1007.
- [20] Engle, R.F. & Bollerslev, T. (1986). Modelling the persistence of conditional variances. With comments and a reply by the authors, *Econometric Reviews* **5**, 1–87.
- [21] Fama, E.F. (1965). The behaviour of stock market prices, *Journal of Business* **38**, 34–105.
- [22] Goldie, C.M. (1991). Implicit renewal theory and tails of solutions of random equations, *Annals of Applied Probability* **1**, 126–1–1.
- [23] Haan, L. & Ferreira, A. (2006). *Extreme Value Theory: An Introduction*, Springer, New York.
- [24] Haan, L. & Resnick, S.I. (1977). Limit theory for multivariate sample extremes, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **40**, 317–337.
- [25] Haan, Lde., Resnick, S.I., Rootzén, H. & Vries, C. Gde. (1989). Extremal behaviour of solutions to a stochastic difference equation with applications to ARCH processes, *Stochastic Processes and Their Applications* **32**, 213–224.
- [26] Ibragimov, I.A. & Linnik, Yu.V. (1971). *Independent and Stationary Sequences of Random Variables*, Wolters-Noordhoff, Groningen.
- [27] Kallenberg, O. (1983). *Random Measures*, 3rd edition, Akademie-Verlag, Berlin.
- [28] Kesten, H. (1973). Random difference equations and renewal theory for products of random matrices, *Acta Mathematica* **131**, 207–248.
- [29] Leadbetter, M.R., Lindgren, G. & Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer, New York.
- [30] Linder, A. (2009). Stationarity, mixing, distributional properties and moments of GARCH(p,q) processes, in T. Andersen, R.A. Davis, J.-P. Kreiss, and T. Mikosch, (eds), *Handbook of Financial Time Series*, Springer, New York.
- [31] Mandelbrot, B. (1963). The variation of certain speculative prices, *Journal of Business* **36**, 394–419.
- [32] Mandelbrot, B. & Taylor, H. (1967). On the distribution of stock price differences, *Operations Research* **15**, 1057–1062.
- [33] Mikosch, T. (2003). Modelling dependence and tails of financial time series, in B. Finkenstadt & H. Rootzen, (eds), *Extreme Values in Finance, Telecommunications and the Environment*, Chapman & Hall, pp. 185–286.
- [34] Mikosch, T. & Stărică, C. (2000). Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process, *Annals of Statistics* **28**, 1427–1451.
- [35] Nelson, D.B. (1990). Stationarity and persistence in the GARCH(1,1) model, *Econometric Theory* **6**, 318–334.
- [36] Resnick, S.I. (1987). *Extreme Values, Regular Variation, and Point Processes*, Springer, New York.
- [37] Resnick, S.I. (2007). *Heavy Tail Phenomena; Probabilistic and Statistical Modeling*, Springer, New York.

Further Reading

- Resnick, S.I. (1986). Point processes, regular variation and weak convergence, *Advances in Applied Probability* **18**, 66–138.
- Taylor, S.J. (1986). *Modelling Financial Time Series*, Wiley, Chichester.

Related Articles

Risk Measures: Statistical Estimation; Stochastic Volatility Models; Volatility.

Extreme Value Theory; GARCH Models; Mandelbrot, Benoit; Mixture of Distribution Hypothesis;

RICHARD A. DAVIS

Filtering

The Filtering Problem

Consider a randomly evolving system, the state of which is denoted by x_t and this state may not be directly observable. Denote by y_t the observation at time $t \in [0, T]$ (x_t and y_t may be vector-valued): y_t is supposed to be probabilistically related to x_t . For instance, y_t may represent a noisy measurement of x_t .

The process x_t is generally supposed to evolve in a Markovian way according to a given (*a priori*) distribution $p(x_t | x_s)$, $s \leq t$. The dynamics of y_t are given in terms of the process x_t ; a general assumption is that, given x_t , the process y_t is independent of its past and so one may consider as given the distribution $p(y_t | x_t)$. The information on x_t at a given $t \in [0, T]$ is thus represented by the past and present observations of y_t , that is, by $y_0^t := \{y_s; s \leq t\}$ or, equivalently, by the filtration $\mathcal{F}_t^y := \sigma\{y_s; s \leq t\}$. This information, combined with the *a priori* dynamics of x given by $p(x_t | x_s)$ can, via a Bayes-type formula, be synthesized in the conditional or posterior distribution $p(x_t | y_0^t)$ of x_t , given y_0^t , and this distribution is called the *filter distribution*.

The filtering problem consists now in determining, possibly in a recursive way, the filter distribution at each $t \leq T$. It can also be seen as a dynamic extension of Bayesian statistics: for $x_t \equiv x$ an unknown parameter, the dynamic model for x given by $p(x_t | x_s)$ reduces to a prior distribution for x and the filter $p(x | y_0^t)$ is then simply the posterior distribution of x , given the observations y_s , $s \leq t$.

In many applications, it suffices to determine a synthetic value of the filter distribution $p(x_t | y_0^t)$. In particular, given an (integrable) function $f(\cdot)$, one may want to compute

$$\begin{aligned} E\{f(x_t) | y_0^t\} &= E\{f(x_t) | \mathcal{F}_t^y\} \\ &= \int f(x) \, dp(x | y_0^t) \end{aligned} \quad (1)$$

The quantity in equation (1) may be seen as the best estimate of $f(x_t)$, given y_0^t , with respect to the mean square error criterion in the sense that $E\{(E\{f(x_t) | y_0^t\}) - f(x_t)\}^2 \leq E\{(g(y_0^t) - f(x_t))^2\}$ for all measurable (and integrable) functions $g(y_0^t)$ of the available information. In this sense, one may also consider

$E\{f(x_t) | \mathcal{F}_t^y\}$ as the *optimal filter* for $f(x_t)$. Notice that determining $E\{f(x_t) | \mathcal{F}_t^y\}$ is no more restrictive than determining the entire filter distribution $p(x_t | y_0^t)$; in fact, by taking $f(x) = e^{i\lambda x}$ for a generic λ , the $E\{f(x_t) | \mathcal{F}_t^y\}$ in equation (1) leads to the conditional characteristic function of x_t given y_0^t .

Related to the filtering problem, are the *prediction problem*, that is, that of determining $p(x_t | y_0^s)$ for $s < t$, and the *interpolation or smoothing problem* concerning $p(x_t | y_0^s)$ for $t < s$. Given the Bayesian nature of the filtering problem, one can also consider the so-called *combined filtering and parameter estimation problem*: if the dynamics $p(x_t | x_s)$ for x include an unknown parameter θ , one may consider the problem of determining the joint conditional distribution $p(x_t, \theta | \mathcal{F}_t^y)$.

Models for the Filtering Problem

To solve a given filtering problem, one has to specify the two basic inputs, namely, $p(x_t | x_s)$ and $p(y_t | x_t)$. A classical model in discrete time is

$$\begin{cases} x_{t+1} = a(t, x_t) + b(t, x_t) w_t \\ y_t = c(t, x_t) + v_t \end{cases} \quad (2)$$

where w_t and v_t are (independent) sequences of independent random variables and the distribution of x_0 is given. Notice that in equation (2) the process x_t is Markov and y_t represents the indirect observations of x_t , affected by additive noise.

The continuous time counterpart is

$$\begin{cases} dx_t = a(t, x_t) dt + b(t, x_t) dw_t \\ dy_t = c(t, x_t) dt + dv_t \end{cases} \quad (3)$$

and notice that, here, y_t represents the cumulative observations up to t . These basic models allow for various extensions: x_t may, for example, be a jump-diffusion process or a Markov process with a finite number of states, characterized by its transition intensities. Also the observations may more generally be a jump-diffusion such as

$$dy_t = c(t, x_t) dt + dv_t + dN_t \quad (4)$$

where N_t is a doubly stochastic Poisson process, the intensity $\lambda_t = \lambda(x_t)$ of which depends on x_t . Further generalizations are, of course, possible.

Analytic Solutions of the Filtering Problem

Discrete Time. By the Markov property of the process x_t and the fact that, given x_t , the process y_t is independent of its past, with the use of Bayes' formula one easily obtains the following two-step recursions

$$\begin{cases} p(x_t | y_0^{t-1}) = \int p(x_t | x_{t-1}) dp(x_{t-1} | y_0^{t-1}) \\ p(x_t | y_0^t) \propto p(y_t | x_t) p(x_t | y_0^{t-1}) \end{cases} \quad (5)$$

where $\propto$ denotes “proportional to” and the first step corresponds to the *prediction step* while the second one is the *updating step*. The recursions start with $p(x_0 | y_0^0) = p(x_0)$. Although equation (5) represents a fully recursive relation, its actual computation is made difficult not only by the presence of the integral in x_{t-1} , but also by the fact that this integral is parameterized by x_t that, in general, takes infinitely many values. Depending on the model, one can however obtain explicit solutions as will be shown below. The most general of such situations arises when one can find a finitely parameterized class of distributions of x_t that is closed under the operator implicit in equation (5), that is, such that, whenever $p(x_{t-1} | y_0^{t-1})$ belongs to this class, then $p(x_t | y_0^t)$ also belongs to it. A classical case is the linear conditionally Gaussian case that corresponds to a model of the form

$$\begin{cases} x_{t+1} = A_t(y_0^t)x_t + B_t(y_0^t)w_t \\ y_t = C_t(y_0^t)x_t + R_t(y_0^t)v_t \end{cases} \quad (6)$$

where the coefficients may depend on the entire past of the observations y_t , and w_t, v_t are independent i.i.d. sequences of standard Gaussian random variables. For such a model, $p(x_t | y_0^t)$ is Gaussian at each t and therefore characterized by mean and (co)variance that can be recursively computed by the well-known *Kalman–Bucy filter*. Denoting

$$\begin{aligned} \hat{x}_{t|t-1} &:= E\{x_t | y_0^{t-1}\}; \hat{x}_{t|t} := E\{x_t | y_0^t\} \\ P_{t|t-1} &:= E\{(x_t - \hat{x}_{t|t-1})(x_t - \hat{x}_{t|t-1})' | y_0^{t-1}\} \\ P_{t|t} &:= E\{(x_t - \hat{x}_{t|t})(x_t - \hat{x}_{t|t})' | y_0^t\} \end{aligned} \quad (7)$$

the Kalman–Bucy filter is given by (dropping for simplicity the dependence on y_0^t),

$$\begin{cases} \hat{x}_{t|t-1} = A_{t-1} \hat{x}_{t-1|t-1} \\ P_{t|t-1} = A_{t-1} P_{t-1|t-1} A_{t-1}' + B_{t-1} B_{t-1}' \end{cases} \quad (8)$$

which represents the *prediction step*, and

$$\begin{aligned} \hat{x}_{t|t} &= \hat{x}_{t|t-1} + L_t[y_t - C_t \hat{x}_{t|t-1}] \\ P_{t|t} &= P_{t|t-1} - L_t C_t P_{t|t-1} \end{aligned} \quad (9)$$

which represents the *updating step* with $\hat{x}_{0|-1}$ the mean of x_0 and $P_{0|-1}$ its variance. Furthermore,

$$L_t := P_{t|t-1} C_t' [C_t P_{t|t-1} C_t' + R_t R_t']^{-1} \quad (10)$$

Notice that, in the prediction step, the estimate of x_t is propagated one step further on the basis of the given *a priori* dynamics of x_t , while in the updating step one takes into account the additional information coming from the current observation. A crucial role in the updating step given by equation (9) is played by

$$\begin{aligned} y_t - C_t \hat{x}_{t|t-1} &= y_t - C_t A_{t-1} \hat{x}_{t-1|t-1} \\ &= y_t - C_t E\{x_t | y_0^{t-1}\} \\ &= y_t - E\{y_t | y_0^{t-1}\} \end{aligned} \quad (11)$$

which represents the new information given by y_t with respect to its best estimate $E\{y_t | y_0^{t-1}\}$ and is therefore called *innovation*.

The Kalman–Bucy filter has been extremely successful and has also been applied to Gaussian models that are nonlinear by simply linearizing the nonlinear coefficient functions around the current best estimate of x_t . In this way, one obtains an approximate filter, called the *extended Kalman filter*.

Exact solutions for the discrete time filtering problem can also be obtained for the case when x_t is a finite-state Markov chain with, say, N states defined by its transition probability matrix. In this case, the filter is characterized by its conditional state probability vector that we denote by $\pi_t = (\pi_t^1, \dots, \pi_t^N)$ with $\pi_t^i := P\{x_t = i | \mathcal{F}_t^y\}$.

Continuous Time. For the solution of a general continuous time problem, we have two main approaches, namely, the *innovations approach* that extends the innovation representation of the Kalman

filter where, combining equations (8) and (9), this latter representation is given by

$$\hat{x}_{t|t} = A_{t-1} \hat{x}_{t-1|t-1} + L_t [y_t - C_t A_{t-1} \hat{x}_{t-1|t-1}] \quad (12)$$

and the so-called *reference probability approach*. For the sake of brevity, we discuss here only the innovations approach (*Kushner–Stratonovich equation*) and we do it for the case of the model in equation (3) mentioning briefly possible extensions to other cases. For the reference probability approach (Zakai equation), we refer to the literature (for instance, [8, 19]).

We denote by $\mathcal{L}$ the generator of the Markov diffusion x_t in equation (3), that is, assuming $x \in \mathbb{R}^n$, for a function $\phi(t, x) \in \mathbb{C}^{1,2}$, we have

$$\begin{aligned} \mathcal{L}\phi(t, x) &= a(t, x)\phi_x(t, x) \\ &+ \frac{1}{2} \sum_{i,j=1}^n \sigma_{ij}(t, x)\phi_{x_i x_j}(t, x) \end{aligned} \quad (13)$$

with $\sigma(t, x) := b(t, x)b'(t, x)$. Furthermore, for a generic (integrable) $f(\cdot)$, we let $\hat{f}_t := E\{f(x_t) \mid \mathcal{F}_t^y\}$. The innovations approach now leads, in case of model given by equation (3), to the following dynamics, also called the *Kushner–Stratonovich equation* (see e.g., [19, 8]):

$$\begin{aligned} d\hat{f}_t &= \widehat{\mathcal{L}f(x_t)} dt + [c(t, \widehat{x_t})\widehat{f(x_t)} \\ &- c(t, \widehat{x_t})\hat{f}_t]' [dy_t - c(t, \widehat{x_t}) dt] \end{aligned} \quad (14)$$

which (see equation (3)) is based on the innovations $dy_t - c(t, \widehat{x_t}) dt = dy_t - E\{dy_t \mid \mathcal{F}_t^y\}$. In addition to the stochastic integral, the main difficulty with equation (14) is that, to compute $\hat{f}$, one needs $\widehat{cf}$, which, in turn, requires $\widehat{c^2 f}$, and so on. In other words, equation (14) is not a closed system of stochastic differential equations. Again, for particular models, equation (14) leads to a closed system as it happens with the linear-Gaussian version of equation (3) that leads to the continuous time Kalman–Bucy filter, which is analogous to its discrete time counterpart. A further case arises when x_t is finite-state Markov with transition intensity matrix $Q = \{q_{ij}\}$, $i, j = 1, \dots, N$. Putting $\pi_t(i) := P\{x_t = i \mid \mathcal{F}_t^y\}$ and taking $f(\cdot)$ as the indicator function of the various values of x_t ,

equation (14) becomes (on replacing $\mathcal{L}$ by Q)

$$\begin{aligned} d\pi_t(j) &= \sum_{i=1}^N \pi_t(i) q_{ij} dt \\ &+ \pi_t(j) \left[c(t, j) - \sum_{i=1}^N \pi_t(i) c(t, i) \right] \\ &\times \left[dy_t - \sum_{i=1}^N \pi_t(i) c(t, i) dt \right] \end{aligned} \quad (15)$$

For more results when x_t is finite-state Markov, we refer to [10], and, in particular, see [11].

We just mention that one can write the dynamics of $\hat{f}_t$ also in the case of jump-diffusion observations as in equation (4) (see [17]) and one can, furthermore, obtain an evolution equation, a stochastic partial differential equation (PDE), for the conditional density $p(x_t) = p(x_t \mid y_0^t)$, whenever it exists, that involves the formal adjoint $\mathcal{L}^*$ of the $\mathcal{L}$ in equation (13) (see [19]).

Numerical Solutions of the Filtering Problem

As we have seen, an explicit analytic solution to the filtering problem can be obtained only for special models so that, remaining within analytic solutions, in general, one has to use an approximation approach. As already mentioned, one such approximation consists in linearizing the nonlinear model, both in discrete and continuous time, and this leads to the extended Kalman filter. Another approach consists in approximating the original model by one where x_t is finite-state Markov. The latter approach goes back mainly to Kushner and coworkers; see, for example, [18] (for a financial application, see also [13]). A more direct numerical approach is simulation-based and given by the so-called *particle approach to filtering* that has been successfully introduced more recently and that is summarized next.

Simulation-based Solution (Particle Filters).

Being simulation-based, this solution method as such is applicable only to discrete time models; continuous time models have to be first discretized in time. There are various variants of particle filters but, analogous to the analytical approaches, they all proceed along two steps, a prediction step and an updating step, and

at each step the relevant distribution (predictive and filter distribution, respectively) is approximated by a discrete probability measure supported by a finite number of points. These approaches vary mainly in the updating step.

A simple version of a particle filter is as follows (see [3]): in the generic period $t - 1$ approximate $p(x_{t-1} | y_0^{t-1})$ by a discrete distribution $((x_{t-1}^1, p_{t-1}^1), \dots, (x_{t-1}^L, p_{t-1}^L))$ where p_{t-1}^i is the probability that $x_{t-1} = x_{t-1}^i$. Consider each location x_{t-1}^i as the position of a “particle”.

1. Prediction step

Propagate each of the particles $x_{t-1}^i \rightarrow \hat{x}_t^i$ over one time period, using the given (discrete time) evolution dynamics of x_t : referring to the model in equation (2) just simulate independent trajectories of x_t starting from the various x_{t-1}^i . This leads to an approximation of $p(x_t | y_0^{t-1})$ by the discrete distribution $((\hat{x}_t^1, \hat{p}_t^1), \dots, (\hat{x}_t^L, \hat{p}_t^L))$ where one puts $\hat{p}_t^i = p_{t-1}^i$.

2. Updating step

Update the weights using the new observation y_t by putting $p_t^i = c p_{t-1}^i p(y_t | \hat{x}_t^i)$ where c is the normalization constant (see the second relation in equation (5) for an analogy).

Notice that $p(y_t | \hat{x}_t^i)$ may be viewed as the likelihood of particle $\hat{x}_t^i$, given the observation y_t , so that in the updating step one weighs each particle according to its likelihood. There exist various improvements of this basic setup. There are also variants, where in the updating step each particle is made to branch into a random number of offsprings, where the mean number of offsprings is taken to be proportional to the likelihood of that position. In this latter variant, the number of particles increases and one can show that, under certain assumptions, the empirical distribution of the particles converges to the true filter distribution. There is a vast literature on particle filters, of which we mention [5] and, in particular, [1].

Filtering in Finance

There are various situations in finance where filtering problems may arise, but one typical situation is given by factor models. These models have proven to be useful for capturing the complicated nonlinear dynamics of real asset prices, while at the same time being parsimonious and numerically tractable. In

addition, with Markovian factor processes, Markov-process techniques can be fruitfully applied. In many financial applications of factor models, the investors have only incomplete information about the actual state of the factors and this may induce model risk. In fact, even if the factors are associated with economic quantities, some of them are difficult to observe precisely. Furthermore, abstract factors without economic interpretation are often included in the specification of a model to increase its flexibility. Under incomplete information of the factors, their values have to be inferred from observable quantities and this is where filtering comes in as an appropriate tool.

Most financial problems concern pricing as well as portfolio management, in particular, hedging and portfolio optimization. While portfolio management is performed under the physical measure, for pricing, one has to use a martingale measure. Filtering problems in finance may therefore be considered under the physical or the martingale measures, or under both (see [22]). In what follows, we shall discuss filtering for pricing problems, with examples from term structure and credit risk, as well as for portfolio management. More general aspects can be found, for example, in the recent papers [6, 7], and [23].

Filtering in Pricing Problems

This section is to a large extent based on [14]. In Markovian factor models, the price of an asset at a generic time t can, under full observation of the factors, be expressed as an instantaneous function $\Psi(t, x_t)$ of time and the value of the factors. Let $\mathcal{G}_t$ denote the full filtration that measures all the processes of interest, and let $\mathcal{F}_t \subset \mathcal{G}_t$ be a subfiltration representing the information of an investor. What is an arbitrage-free price in the filtration $\mathcal{F}_t$? Assume the asset to be priced is a European derivative with maturity T and claim $H \in \mathcal{F}_T$. Let N be a numeraire, adapted to the investor filtration $\mathcal{F}_t$, and let Q^N be the corresponding martingale measure. One can easily prove the following:

Lemma 1 *Let $\Psi(t, x_t) = N_t E^{Q^N} \left\{ \frac{H}{N_T} \mid \mathcal{G}_t \right\}$ be the arbitrage-free price of the claim H under the full*

information $\mathcal{G}_t$ and $\hat{\Psi}(t) = N_t E^{Q^N} \left\{ \frac{H}{N_t} \mid \mathcal{F}_t \right\}$ the corresponding arbitrage-free price in the investor filtration. It then follows that

$$\hat{\Psi}(t) = E^{Q^N} \{ \Psi(t, x_t) \mid \mathcal{F}_t \} \quad (16)$$

Furthermore, if the savings account $B_t = \exp\{\int_0^t r_s ds\}$ with corresponding martingale measure Q is $\mathcal{F}_t$ -adapted, then

$$\hat{\Psi}(t) = E^Q \{ \Psi(t, x_t) \mid \mathcal{F}_t \} \quad (17)$$

We thus see that, to compute the right-hand sides in equation (16) or equation (17), namely, the price of a derivative under restricted information given its price under full information, one has to solve the filtering problem for x_t given $\mathcal{F}_t$ under a martingale measure. We present now two examples.

Example 1 (*Term structure of interests*). The example is a simplified version adapted from [15]. Consider a factor model for the term structure where the unobserved (multivariate) factor process x_t satisfies the linear-Gaussian model

$$dx_t = F x_t dt + D dw_t \quad (18)$$

In this case, the term structure is exponentially affine in x_t and one has

$$p(t, T; x_t) = \exp[A(t, T) - B(t, T) x_t] \quad (19)$$

with $A(t, T)$, $B(t, T)$ satisfying well-known first-order ordinary differential equations to exclude arbitrage. Passing to log-prices for the bonds, one gets the linear relationship $y_t^T := \log p(t, T; x_t) = A(t, T) - B(t, T) x_t$. Assume now that investors cannot observe x_t , but they can observe the short rate and the log-prices of a finite number n of zero-coupon bonds, perturbed by additive noise. This leads to a system of the form

$$\begin{cases} dx_t = F x_t dt + D dw_t \\ dr_t = (\alpha_t^0 + \beta_t^0 x_t) dt + \sigma_t^0 dw_t + dv_t^0 \\ dy_t^i = (\alpha_t^i + \beta_t^i x_t) dt + \sigma_t^i dw_t + (T_i - t) dv_t^i \\ \quad ; i = 1, \dots, n \end{cases} \quad (20)$$

where v_t^i , $i = 0, \dots, n$ are independent Wiener processes and the coefficients are related to those in equations (18) and (19). The time-dependent volatility in the perturbations of the log-prices reflects the fact that it tends to zero as time approaches maturity.

From the filtering point of view, the system (20) is a linear-Gaussian model with x_t unobserved and the observations given by (r_t, y_t^i) . We shall thus put $\mathcal{F}_t = \sigma\{r_s, y_s^i; s \leq t, i = 1, \dots, n\}$. The filter distribution is Gaussian and, via the Kalman filter, one can obtain its conditional mean m_t and (co)variance Σ_t . Applying Lemma 1 and using the moment-generating function of a Gaussian random variable, we obtain the arbitrage-free price, in the investor filtration, of an illiquid bond with maturity T as follows:

$$\begin{aligned} \hat{p}(t, T) &= E\{p(t, T; x_t) \mid \mathcal{F}_t\} \\ &= \exp[A(t, T)] E\{\exp[-B(t, T) x_t] \mid \mathcal{F}_t\} \\ &= \exp[A(t, T) - B(t, T) m_t] \\ &\quad + \frac{1}{2} B(t, T) \Sigma_t B'(t, T) \end{aligned} \quad (21)$$

For the given setup, the expectation is under the martingale measure Q with the money market account B_t as numeraire. To apply Lemma 1, we need the numeraire to be observable and this contrasts with the assumption that r_t is observable only in noise. This difficulty can be overcome (see [14]), but by suitably changing the drifts in equation (20) (corresponding to a translation of w_t), one may however consider the model in equation (20) also under a martingale measure for which the numeraire is different from B_t and observable.

A further filter application to the term structure of interest rates can be found in [2].

Example 2 (*Credit risk*). One of the main issues in credit risk is the modeling of the dynamic evolution of the default state of a given portfolio. To formalize the problem, given a portfolio of m obligors, let $y_t := (y_{t,1}, \dots, y_{t,m})$ be the default indicator process where $y_{t,i} := \mathbf{1}_{\{\tau_i \leq t\}}$ with τ_i the random default time of obligor i , $i = 1, \dots, m$. In line with the factor modeling philosophy, it is natural to assume that default intensities depend on an unobservable latent process x_t . In particular, if $\lambda_i(t)$ is the default intensity of obligor i , $i = 1, \dots, m$, assume $\lambda_i(t) = \lambda_i(x_t)$. Note that this generates *information-driven contagion*: it is, in fact, well known that the intensities with respect to $\mathcal{F}_t$ are given by $\hat{\lambda}_i(t) = E\{\lambda_i(x_t) \mid \mathcal{F}_t\}$. Hence the news that an obligor has defaulted leads, via filtering, to an update of the distribution

of x_t and thus to a jump in the default intensities of the still surviving obligors. In this context, we shall consider the pricing of illiquid credit derivatives on the basis of the investor filtration supposed to be given by the default history and noisily observed prices of liquid credit derivatives.

We assume that, conditionally on x_t , the defaults are independent with intensities $\lambda_i(x_t)$ and that (x_t, y_t) is jointly Markov. A credit derivative has the payoff linked to default events in a given reference portfolio and so one can think of it as a random variable $H \in \mathcal{F}_T^y$ with T being the maturity. Its full information price at the generic $t \leq T$, that is, in the filtration $\mathcal{G}_t$ that measures also x_t , is given by $\tilde{H}_t = E\{e^{-r(T-t)} H \mid \mathcal{G}_t\}$ where r is the short rate and the expectation is under a given martingale measure Q . By the Markov property of (x_t, y_t) , one gets a representation of the form

$$\tilde{H}_t = E\{e^{-r(T-t)} H \mid \mathcal{G}_t\} := a(t, x_t, y_t) \quad (22)$$

for a suitable $a(\cdot)$. In addition to the default history, we assume that the investor filtration also includes noisy observations of liquid credit derivatives. In view of equation (22), it is reasonable to model such observations as

$$dz_t = \gamma(t, x_t, y_t) dt + d\beta_t \quad (23)$$

where the various quantities may also be column vectors, β_t is an independent Wiener process and $\gamma(\cdot)$ is a function of the type of $a(\cdot)$ in equation (22). The investor filtration is then $\mathcal{F}_t = \mathcal{F}_t^y \vee \mathcal{F}_t^z$. The price at $t < T$ of the credit derivative in the investor filtration is now $H_t = E\{e^{-r(T-t)} H \mid \mathcal{F}_t\}$ and by Lemma 1 we have

$$H_t = E\{e^{-r(T-t)} H \mid \mathcal{F}_t\} = E\{a(t, x_t, y_t) \mid \mathcal{F}_t\} \quad (24)$$

Again, if one knows the price $a(t, x_t, y_t)$ in $\mathcal{G}_t$, one can thus obtain the price in $\mathcal{F}_t$ by computing the right-hand side in equation (24) and for this we need the filter distribution of x_t given $\mathcal{F}_t$.

To define the corresponding filtering problem, we need a more precise model for (x_t, y_t) (the process z_t is already given by equation (23)). Since y_t is a jump process, the model cannot be one of those for which we had described an explicit analytic solution. Without entering into details, we refer to [13] (see also [14]), where a jump-diffusion model is considered that allows for common jumps between

x_t and y_t . In [13] it is shown that an arbitrarily good approximation to the filter solution can be obtained both analytically and by particle filtering.

We conclude this section with a couple of additional remarks:

1. Traditional credit risk models are either structural models or reduced-form (intensity-based) models. Example 2 belongs to the latter class. In *structural models*, the default of the generic obligor/firm i is defined as the first passage time of the asset value $V_i(t)$ of the firm at a given (possibly stochastic) barrier $K_i(t)$, that is,

$$\tau_i = \inf\{t \geq 0 \mid V_i(t) \leq K_i(t)\} \quad (25)$$

In such a context, filtering problems may arise when either $V_i(t)$ or $K_i(t)$ or both are not exactly known/observable (see e.g., [9]).

2. *Can a structural model also be seen as a reduced-form model?* At first sight, this is not clear since τ_i in equation (25) is predictable, while in intensity-based models it is totally inaccessible. However, it turns out (see e.g., [16]) that, while τ_i in equation (25) is predictable with respect to the full filtration (measuring also $V_i(t)$ and $K_i(t)$), it becomes totally inaccessible in the smaller investor filtration that, say, does not measure $V_i(t)$ and, furthermore, it admits an intensity.

Filtering in Portfolio Management Problems

Rather than presenting a general treatment (for this, we refer to [21] and the references therein), we discuss here two specific examples in models with unobserved factors, one in discrete time and one in continuous time. Contrary to the previous section on pricing, here we shall work under the physical measure P .

A Discrete Time Case. To motivate the model, start from the classical continuous time asset price model $dS_t = S_t[a dt + x_t dw_t]$ where w_t is Wiener and x_t is the nondirectly observable volatility process (factor). For $y_t := \log S_t$, one then has

$$dy_t = \left(a - \frac{1}{2}x_t^2\right) dt + x_t dw_t \quad (26)$$

Passing to discrete time with step δ , let for $t = 0, \dots, T$ the process x_t be a Markov chain with m

states $x^1, \dots, x^m$ (may result from a time discretization of a continuous time x_t) and

$$y_t = y_{t-1} + \left(a - \frac{1}{2}x_{t-1}^2\right)\delta + x_{t-1}\sqrt{\delta}\varepsilon_t \quad (27)$$

with ε_t i.i.d. standard Gaussian as it results from equation (26) by applying the Euler–Maruyama scheme. Notice that (x_t, y_t) is Markov. Having for simplicity only one stock to invest in, denote by ϕ_t the number of shares of stock held in the portfolio in period t with the rest invested in a riskless bond B_t (for simplicity assume $r = 0$). The corresponding self-financed wealth process then evolves according to

$$V_{t+1}^\phi = V_t^\phi + \phi_t (e^{y_{t+1}} - e^{y_t}) := F(V_t^\phi, \phi_t, y_t, y_{t+1}) \quad (28)$$

and ϕ_t is supposed to be adapted to $\mathcal{F}_t^y$; denote by $\mathcal{A}$ the class of such strategies. Given a horizon T , consider the following investment criterion

$$\begin{aligned} J_{\text{opt}}(V_0) &= \sup_{\phi \in \mathcal{A}} J(V_0, \phi) \\ &= \sup_{\phi \in \mathcal{A}} E \left\{ \sum_{t=0}^{T-1} r_t(x_t, y_t, V_t^\phi, \phi_t) \right. \\ &\quad \left. + f(x_T, y_T, V_T^\phi) \right\} \end{aligned} \quad (29)$$

which, besides portfolio optimization, includes also hedging problems. The problem in equations (27), (28), and (29) is now a stochastic control problem under partial/incomplete information given that x_t is an unobservable factor process.

A standard approach to dynamic optimization problems under partial information is to transform them into corresponding complete information ones whereby x_t is replaced by its filter distribution given $\mathcal{F}_t^y$. Letting $\pi_t^i := P\{x_t = x^i \mid \mathcal{F}_t^y\}$, $i = 1, \dots, m$ we first adapt the filter dynamics in equation (5) to our situation to derive a recursive relation for $\pi_t = (\pi_t^1, \dots, \pi_t^m)$. Being x_t finite-state Markov, $p(x_{t+1} \mid x_t)$ is given by the transition probability matrix and the integral in equation (5) reduces to a sum. On the other hand, $p(y_t \mid x_t)$ in equation (5) corresponds to the model in equation (2) that does not include our model in equation (27) for y_t . One can however easily see that equation (27) leads to a

distribution of the form $p(y_t \mid x_{t-1}, y_{t-1})$, and equation (5) can be adapted to become here

$$\begin{cases} \pi_0 = \mu & \text{(initial distribution for } x_t) \\ \pi_t^i \propto \sum_{j=1}^m p(y_t \mid x_{t-1} = j, y_{t-1}) \\ \quad p(x_t = i \mid x_{t-1} = j) \pi_{t-1}^j \end{cases} \quad (30)$$

In addition, we may consider the law of y_t conditional on $(\pi_{t-1}, y_{t-1}) = (\pi, y)$ that is given by

$$\begin{aligned} Q_t(\pi, y, dy') &= \sum_{i,j=1}^m p(y' \mid x_{t-1} = j, y) \\ &\quad p(x_t = i \mid x_{t-1} = j) \pi^j \end{aligned} \quad (31)$$

From equations (30) and (31), it follows easily that (π_t, y_t) is a sufficient statistic and an $\mathcal{F}_t^y$ -Markov process.

To transform the original partial information problem with criterion (29) into a corresponding complete observation problem, put $\hat{r}_t(\pi, y, v, \phi) = \sum_{i=1}^m r_t(x^i, y, v, \phi) \pi^i$ and $\hat{f}(\pi, y, v) = \sum_{i=1}^m f(x^i, y, v) \pi^i$ so that, by double conditioning, one obtains

$$\begin{aligned} J(V_0, \phi) &= E \left\{ \sum_{t=0}^{T-1} E \left\{ r_t(x_t, y_t, V_t^\phi, \phi_t) \mid \mathcal{F}_t^y \right\} \right. \\ &\quad \left. + E \left\{ f(x_T, y_T, V_T^\phi) \mid \mathcal{F}_T^y \right\} \right\} \\ &= E \left\{ \sum_{t=0}^{T-1} \hat{r}_t(\pi_t, y_t, V_t^\phi, \phi_t) + \hat{f}(\pi_T, y_T, V_T^\phi) \right\} \end{aligned} \quad (32)$$

Owing to the Markov property of (π_t, y_t) , one can write the following (backward) dynamic programming recursions:

$$\begin{cases} u_T(\pi, y, v) = \hat{f}(\pi, y, v) \\ u_t(\pi, y, v) = \sup_{\phi \in \mathcal{A}} [\hat{r}_t(\pi, y, v, \phi) \\ \quad + E \{ u_{t+1}(\pi_{t+1}, y_{t+1}, \\ \quad F(v, \phi, y, y_{t+1})) \mid (\pi_t, y_t) = (\pi, y) \}] \end{cases} \quad (33)$$

where the function $F(\cdot)$ was defined in equation (28), and ϕ here refers to the generic choice of $\phi = \phi_t$ in period t . It leads to the optimal investment strategy ϕ^* and the optimal value $J_{\text{opt}}(V_0) = u_0(\mu, y_0, V_0)$. It can, in fact, be shown that the strategy and value thus

obtained are optimal also for the original incomplete information problem when ϕ there is required to be $\mathcal{F}_t^y$ -adapted.

To actually compute the recursions in equation (33), one needs the conditional law of (π_{t+1}, y_{t+1}) given (π_t, y_t) , which can be deduced from equations (30) and (31). In this context, notice that, even if x is m -valued, π_t takes values in the m -dimensional simplex that is ∞ -valued. To actually perform the calculation, one needs an approximation leading to a finite-valued process (π_t, y_t) and to this effect various approaches have appeared in the literature (for an approach with numerical results see [4]).

A Continuous Time Case. Consider the following market model where x_t is an unobserved factor process and S_t is the price of a single risky asset:

$$\begin{cases} dx_t = F_t(x_t) dt + R_t(x_t) dM_t \\ dS_t = S_t [a_t(S_t, x_t) dt + \sigma_t(S_t) dw_t] \end{cases} \quad (34)$$

with w_t a Wiener process and M_t a not necessarily continuous martingale, independent of w_t . Since, in continuous time, $\int_0^t \sigma_s^2 ds$ can be estimated by the empirical quadratic variation of S_t , in order not to have degeneracy in the filter to be derived below for x_t , we do not let $\sigma(\cdot)$ depend also on x_t . For the riskless asset, we assume for simplicity that its price is $B_t \equiv \text{const}$ (short rate $r = 0$). In what follows, it is convenient to consider log-prices $y_t = \log S_t$, for which

$$\begin{aligned} dy_t &= [a_t(S_t, x_t) - \frac{1}{2}\sigma_t^2(S_t)] dt + \sigma_t(S_t) dw_t \\ &:= A_t(y_t, x_t) dt + B_t(y_t) dw_t \end{aligned} \quad (35)$$

Investing in this market in a self-financing way and denoting by ρ_t the fraction of wealth invested in the risky asset, we have from $\frac{dV_t}{V_t} = \rho_t \frac{dS_t}{S_t} = \rho_t \frac{d}{e^{y_t}} e^{y_t}$ that

$$\begin{aligned} dV_t &= V_t \left[\rho_t \left(A_t(y_t, x_t) + \frac{1}{2} B_t^2(y_t) \right) dt \right. \\ &\quad \left. + \rho_t B_t(y_t) dw_t \right] \end{aligned} \quad (36)$$

We want to consider the problem of maximization of expected utility from terminal wealth, without

consumption, and with a power utility function. Combining equations (34), (35), and (36) we obtain the following portfolio optimization problem under incomplete information where the factor process x_t is not observed and where we shall require that ρ_t is $\mathcal{F}_t^y$ -adapted:

$$\begin{cases} dx_t = F_t(x_t) dt + R_t(x_t) dM_t & (\text{unobserved}) \\ dy_t = A_t(y_t, x_t) dt + B_t(y_t) dw_t & (\text{observed}) \\ dV_t = V_t \left[\rho_t \left(A_t(y_t, x_t) + \frac{1}{2} B_t^2(y_t) \right) dt \right. \\ \quad \left. + \rho_t B_t(y_t) dw_t \right] \\ \sup_{\rho} E \{ (V_T)^\mu \}, \quad \mu \in (0, 1) \end{cases} \quad (37)$$

As in the previous discrete time case, we shall now transform this problem into a corresponding one under complete information, thereby replacing the unobserved state variable x_t by its filter distribution, given $\mathcal{F}_t^y$, that is, $\pi_t(x) := p(x_t | \mathcal{F}_t^y)_{x_t=x}$. Even if x_t is finite-dimensional, $\pi_t(\cdot)$ is ∞ -dimensional. We have seen above cases where the filter distribution is finitely parameterized, namely, the linear-Gaussian case (*Kalman filter*) and when x_t is finite-state Markov. The parameters characterizing the filter were seen to evolve over time driven by the innovations process (see equations (8), (10) and (14)). In what follows, we then assume that the filter is parameterized by a vector process $\xi_t \in \mathbb{R}^p$, that is, $\pi_t(x) := p(x_t | \mathcal{F}_t^y)_{x_t=x} = \pi(x; \xi_t)$ and that ξ_t satisfies

$$d\xi_t = \beta_t(y_t, \xi_t) dt + \eta_t(y_t, \xi_t) d\bar{w}_t \quad (38)$$

where $\bar{w}_t$ is Wiener and given by the innovations process. We now specify this innovations process $\bar{w}_t$ for our general model in equation (37). To this effect, putting $A_t(y_t, \xi_t) := \int A_t(y_t, x) d\pi_t(x; \xi_t)$, let

$$d\bar{w}_t := B_t^{-1}(y_t) [dy_t - A_t(y_t, \xi_t) dt] \quad (39)$$

and notice that, replacing dy_t from equation (35), this definition implies a translation of the original $(P, \mathcal{F}_t)$ -Wiener w_t , that is,

$$d\bar{w}_t = dw_t + B_t^{-1}(y_t) [A_t(y_t, x_t) - A_t(y_t, \xi_t)] dt \quad (40)$$

and thus the implicit change of measure $P \rightarrow \bar{P}$ with

$$\begin{aligned} \frac{d\bar{P}}{dP}|_{\mathcal{F}_T} = \exp \left\{ \int_0^T [A_t(y_t, \xi_t) - A_t(y_t, x_t)] \right. \\ \left. \times B_t^{-1}(y_t) dw_t - \frac{1}{2} \int_0^T [A_t(y_t, \xi_t) \right. \\ \left. - A_t(y_t, x_t)]^2 B_t^{-2}(y_t) dt \right\} \quad (41) \end{aligned}$$

We obtain thus as the complete information problem corresponding to equation (37), the following, which is defined on the space $(\Omega, \mathcal{F}, \mathcal{F}_t, \bar{P})$ with Wiener $\bar{w}_t$:

$$\begin{cases} d\xi_t = \beta_t(y_t, \xi_t) dt + \eta_t(y_t, \xi_t) d\bar{w}_t \\ dy_t = A_t(y_t, \xi_t) dt + B_t(y_t) d\bar{w}_t \\ dV_t = V_t \left[\rho_t \left(A_t(y_t, \xi_t) + \frac{1}{2} B_t^2(y_t) \right) dt \right. \\ \left. + \rho_t B_t(y_t) d\bar{w}_t \right] \\ \sup_{\rho} \bar{E} \{ (V_T)^\mu \}, \quad \mu \in (0, 1) \end{cases} \quad (42)$$

One can now use methods for complete information problems to solve equation (42), and it can also be shown that the solution to equation (42) gives a solution of the original problem for which ρ_t was assumed $\mathcal{F}_t^y$ -adapted.

We remark that other reformulations of the incomplete information problem as a complete information one are also possible (see e.g., [20]).

A final comment concerns hedging under incomplete information (incomplete market). When using the quadratic hedging criterion, that is, $\min_{\rho} E_{S_0, V_0} \{(H_T - V_T^\rho)^2\}$, its quadratic nature implies that if $\phi_t^*(x_t, y_t)$ is the optimal strategy (number of units invested in the risky asset) under complete information also of x_t , then, under the partial information $\mathcal{F}_t^y$, the optimal strategy is simply the projection $E\{\phi_t^*(x_t, y_t) | \mathcal{F}_t^y\}$ that can be computed on the basis of the filter of x_t given $\mathcal{F}_t^y$ (see [12]).

References

- [1] Bain, A. & Crisan, D. (2009). Fundamentals of stochastic filtering, in *Series: Stochastic Modelling and Applied Probability*, Vol. 60, Springer Science+Business Media, New York.
- [2] Bhar, R. Chiarella, C. Hung, H. & Runggaldier, W. (2005). The volatility of the instantaneous spot interest rate implied by arbitrage pricing—a dynamic Bayesian approach. *Automatica* **42**, 1381–1393.
- [3] Budhiraja, A., Chen, L. & Lee, C. (2007). A survey of nonlinear methods for nonlinear filtering problems, *Physica D* **230**, 27–36.
- [4] Corsi, M., Pham, H. & Runggaldier, W.J. (2008). Numerical approximation by quantization of control problems in finance under partial observations, to appear in *Mathematical Modeling and Numerical Methods in Finance. Handbook of Numerical Analysis*, A. Bensoussan & Q. Zhang, eds, Elsevier, Vol. 15.
- [5] Crisan, D., Del Moral, P. & Lyons, T. (1999). Interacting particle systems approximations of the Kushner–Stratonovich equation, *Advances in Applied Probability* **31**, 819–838.
- [6] Cvitanic, J., Liptser, R. & Rozovski, B. (2006). A filtering approach to tracking volatility from prices observed at random times, *The Annals of Applied Probability* **16**, 1633–1652.
- [7] Cvitanic, J., Rozovski, B. & Zaliapin, I. (2006). Numerical estimation of volatility values from discretely observed diffusion data, *Journal of Computational Finance* **9**, 1–36.
- [8] Davis, M.H.A. & Marcus, S.I. (1981). An Introduction to nonlinear filtering, in *Stochastic Systems: The Mathematics of Filtering and Identification and Applications* M. Hazewinkel & J.C. Willems, eds, D.Reidel, Dordrecht, pp. 53–75.
- [9] Duffie, D. & Lando, D. (2001). Term structure of credit risk with incomplete accounting observations, *Econometrica* **69**, 633–664.
- [10] Elliott, R.J. (1993). New finite-dimensional filters and smoothers for noisily observed Markov chains, *IEEE Transactions on Information Theory*, **IT-39**, 265–271.
- [11] Elliott, R.J., Aggoun, L. & Moore, J.B. (1994). Hidden Markov models: estimation and control, in *Applications of Mathematics*, Springer-Verlag, Berlin-Heidelberg-New York, Vol. 29.
- [12] Frey, R. & Runggaldier, W. (1999). Risk-minimizing hedging strategies under restricted information: the case of stochastic volatility models observed only at discrete random times, *Mathematical Methods of Operations Research* **50**(3), 339–350.
- [13] Frey, R. & Runggaldier, W. (2008). *Credit risk and incomplete information: a nonlinear filtering approach*, preprint, Universitat Leipzig, Available from www.math.uni-leipzig.de/~7Efrey/publications-frey.html.
- [14] Frey, R. & Runggaldier, W.R. Nonlinear filtering in models for interest-rate and credit risk, to appear in *Handbook of Nonlinear Filtering*, D. Crisan & B. Rozovski, eds, Oxford University Press (to be published in 2009).
- [15] Gombani, A., Jaschke, S. & Runggaldier, W. (2005). A filtered no arbitrage model for term structures with noisy data, *Stochastic Processes and Applications* **115**, 381–400.

-
- [16] Jarrow, R. & Protter, P. (2004). Structural versus reduced-form models: a new information based perspective, *Journal of Investment Management*, **2**, 1–10.
- [17] Kliemann, W., Koch, G. & Marchetti, F. (1990). On the unnormalized solution of the filtering problem with counting process observations, *IEEE IT-36*, 1415–1425.
- [18] Kushner, H.J. & Dupuis, P. (1992). Numerical methods for stochastic control Problems in continuous time, in *Applications of Mathematics*, Springer, New York, Vol. 24.
- [19] Liptser, R.S. & Shiryaev, A.N. (2001). *Statistics of random processes, Series: Applications of Mathematics; Stochastic Modelling and Applied Probability*, Springer-Verlag, Berlin, Vols. I, II.
- [20] Nagai, H. & Runggaldier, W.J. (2008). PDE approach to utility maximization for market models with hidden Markov factors, in *Seminar on Stochastic Analysis, Random Fields and Applications V*, R.C. Dalang, M. Dozzi, & F. Russo, eds, *Progress in Probability*, Birkhäuser Verlag, Vol. 59, pp. 493–506.
- [21] Pham, H. Portfolio optimization under partial observation: theoretical and numerical aspects, to appear in *Handbook of Nonlinear Filtering*, D. Crisan & B. Rozovski, eds, Oxford University Press (to be published in 2009).
- [22] Runggaldier, W.J. (2004). Estimation via stochastic filtering in financial market models, in *Mathematics of Finance. Contemporary Mathematics*, G. Yin & Q. Zhang, eds, AMS, Vol. 351, pp. 309–318.
- [23] Zeng, Y. (2003). A partially observed model for micro-movement of asset prices with Bayes estimation via filtering, *Mathematical Finance*, **13**, 411–444.

WOLFGANG RUNGALDIER

Filtrations

The notion of *filtration*, introduced by Doob, has become a fundamental feature of the theory of stochastic processes. Most basic objects, such as martingales, semimartingales, stopping times, or Markov processes, involve the notion of filtration.

Definition 1 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A filtration $\mathbb{F}$, on $(\Omega, \mathcal{F}, \mathbb{P})$, is an increasing family $(\mathcal{F}_t)_{t \geq 0}$ of sub- σ -algebras of $\mathcal{F}$. In other words, for each t , $\mathcal{F}_t$ is a σ -algebra included in $\mathcal{F}$ and if $s \leq t$, $\mathcal{F}_s \subset \mathcal{F}_t$. A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ endowed with a filtration $\mathbb{F}$ is called a *filtered probability space*.

We now give a definition that is very closely related to that of a filtration.

Definition 2 A stochastic process $(X_t)_{t \geq 0}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is adapted to the filtration $(\mathcal{F}_t)$ if, for each $t \geq 0$, X_t is $\mathcal{F}_t$ -measurable.

A stochastic process X is always adapted to its *natural filtration* $\mathbb{F}^X$, where for each $t \geq 0$, $\mathcal{F}_t^X = \sigma(X_s, s \leq t)$ (the last notation means that $\mathcal{F}_t$ is the smallest σ -algebra with respect to which all the variables $(X_s, s \leq t)$ are measurable). $\mathbb{F}^X$ is, hence, the smallest filtration to which X is adapted.

The parameter t is often thought of as time, and the σ -algebra $\mathcal{F}_t$ represents the set of information available at time t , that is, events that have occurred up to time t . Thus, the filtration $\mathbb{F}$ represents the evolution of the information or knowledge of the world with time. If X is an adapted process, then X_t , its value at time t , depends only on the evolution of the universe prior to t .

Definition 3 Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space.

1. The filtration $\mathbb{F}$ is said to be *complete* if $(\Omega, \mathcal{F}, \mathbb{P})$ is complete and if $\mathcal{F}_0$ contains all the $\mathbb{P}$ -null sets.
2. The filtration $\mathbb{F}$ is said to satisfy the *usual hypotheses* if it is complete and right continuous, that is, for all $t \geq 0$, $\mathcal{F}_t = \mathcal{F}_{t+}$, where

$$\mathcal{F}_{t+} = \bigcap_{u > t} \mathcal{F}_u \quad (1)$$

Some fundamental theorems, such as the D  but theorem, require the usual hypotheses. Hence naturally, very often in the literature on the theory of stochastic processes and mathematical finance, the underlying filtered probability spaces are assumed to satisfy the usual hypotheses. This assumption is not very restrictive for the following reasons:

1. Any filtration can easily be made complete and right continuous; indeed, given a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, we first complete the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and then we add all the $\mathbb{P}$ -null sets to every $\mathcal{F}_{t+}$, $t \geq 0$. The new filtration thus obtained satisfies the usual hypotheses and is called the *usual augmentation* of $\mathbb{F}$;
2. Moreover, in most classical and encountered cases, the filtration $\mathbb{F}$ is right continuous. Indeed, this is the case when, for instance, $\mathbb{F}$ is the natural filtration of a Brownian motion, a L  vy process, a Feller process, or a Hunt process [8, 9].

Enlargements of Filtrations

For more precise and detailed references, the reader can consult the books [4–6, 8] or the survey article [7].

Generalities

Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space satisfying the usual hypotheses. Let $\mathbb{G}$ be another filtration satisfying the usual hypotheses and such that $\mathcal{F}_t \subset \mathcal{G}_t$ for every $t \geq 0$. One natural question is, how are the $\mathbb{F}$ -semimartingales modified when considered as stochastic processes in the larger filtration $\mathbb{G}$? Given the importance of semimartingales and martingales (in particular, in mathematical finance where they are used to model prices), it seems natural to characterize situations where the semimartingale or martingale properties are preserved.

Definition 4 We shall say that the pair of filtrations $(\mathbb{F}, \mathbb{G})$ satisfies the *(H') hypothesis* if every $\mathbb{F}$ -semimartingale is a $\mathbb{G}$ -semimartingale.

Remark 1 In fact, using a classical decomposition of semimartingales due to Jacod and M  min, it is enough to check that every $\mathbb{F}$ -bounded martingale is a $\mathbb{G}$ -semimartingale.

Definition 5 We shall say that the pair of filtrations $(\mathbb{F}, \mathbb{G})$ satisfies the (H) hypothesis if every $\mathbb{F}$ -local martingale is a $\mathbb{G}$ -local martingale.

The theory of enlargements of filtrations, developed in the late 1970s, provides answers to questions such as those mentioned earlier. Currently, this theory has been widely used in mathematical finance, especially in insider trading models and in models of default risk. The insider trading models are usually based on the so-called *initial enlargements of filtrations*, whereas the models of default risk fit well in the framework of the *progressive enlargements of filtrations*. More precisely, given a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, there are essentially two ways of enlarging filtrations:

- *initial enlargements*, for which $\mathcal{G}_t = \mathcal{F}_t \vee \mathcal{H}$ for every $t \geq 0$, that is, the new information $\mathcal{H}$ is brought in at the origin of time and
- *progressive enlargements*, for which $\mathcal{G}_t = \mathcal{F}_t \vee \mathcal{H}_t$ for every $t \geq 0$, that is, the new information is brought in progressively as the time t increases.

Before presenting the basic theorems on enlargements of filtrations, we state a useful theorem due to Stricker.

Theorem 1 (Stricker [10]). Let $\mathbb{F}$ and $\mathbb{G}$ be two filtrations as above, such that for all $t \geq 0$, $\mathcal{F}_t \subset \mathcal{G}_t$. If (X_t) is a $\mathbb{G}$ -semimartingale that is $\mathbb{F}$ -adapted, then it is also an $\mathbb{F}$ -semimartingale.

Initial Enlargements of Filtrations

The most important theorem on initial enlargements of filtrations is due to Jacod and deals with the special case where the initial information brought in at the origin of time consists of the σ -algebra generated by a random variable. More precisely, let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space satisfying the usual assumptions. Let Z be an $\mathcal{F}$ measurable random variable. Define

$$\mathcal{G}_t = \bigcap_{\varepsilon > 0} \left(\mathcal{F}_{t+\varepsilon} \vee \sigma\{Z\} \right), \quad t \geq 0 \quad (2)$$

In financial models, the filtration $\mathbb{F}$ represents the public information in a financial market and the random variable Z stands for the additional (anticipating) information of an insider.

The conditional laws of Z given $\mathcal{F}_t$, for $t \geq 0$, play a crucial role in initial enlargements.

Theorem 2 (Jacod's criterion). Let Z be an $\mathcal{F}$ measurable random variable and let $Q_t(\omega, dx)$ denote the regular conditional distribution of Z given $\mathcal{F}_t$, $t \geq 0$. Suppose that for each $t \geq 0$, there exists a positive σ -finite measure $\eta_t(dx)$ (on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$) such that

$$Q_t(\omega, dx) \ll \eta_t(dx) \text{ almost surely} \quad (3)$$

Then every $\mathbb{F}$ -semimartingale is a $\mathbb{G}$ -semimartingale.

Remark 2 In fact, this theorem still holds for random variables with values in a standard Borel space. Moreover, the existence of the σ -finite measure $\eta_t(dx)$ is equivalent to the existence of one positive σ -finite measure $\eta(dx)$ such that $Q_t(\omega, dx) \ll \eta(dx)$ and in this case η can be taken to be the distribution of Z .

Now we give classical corollaries of Jacod's theorem.

Corollary 1 Let Z be independent of $\mathcal{F}_\infty$. Then, every $\mathbb{F}$ -semimartingale is a $\mathbb{G}$ -semimartingale.

Corollary 2 Let Z be a random variable taking on only a countable number of values. Then every $\mathbb{F}$ -semimartingale is a $\mathbb{G}$ -semimartingale.

In some cases, it is possible to obtain an explicit decomposition of an $\mathbb{F}$ -local martingale as a $\mathbb{G}$ -semimartingale [4–8]. For example, if $Z = B_{t_0}$, for some fixed time $t_0 > 0$ and a Brownian Motion B , it can be shown that Jacod's criterion holds for $t < t_0$ and that every $\mathbb{F}$ -local martingale is a semimartingale for $0 \leq t < t_0$, but not necessarily including t_0 . Indeed in this case, there are $\mathbb{F}$ -local martingales that are not $\mathbb{G}$ -semimartingales. Moreover, B is a $\mathbb{G}$ -semimartingale, which decomposes as

$$B_t = B_0 + \tilde{B}_t + \int_0^{t \wedge t_0} ds \frac{B_{t_0} - B_s}{t_0 - s} \quad (4)$$

where $(\tilde{B}_t)$ is a $\mathbb{G}$ Brownian Motion.

Remark 3 There are cases where Jacod's criterion does not hold but where other methods apply [4, 6, 7].

Progressive Enlargements of Filtrations

Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space satisfying the usual hypotheses, and $\rho : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ be a random time. We enlarge the initial filtration $\mathbb{F}$ with the process $(\rho \wedge t)_{t \geq 0}$, so that the new enlarged filtration $\mathbb{F}^\rho$ is the smallest filtration (satisfying the usual assumptions) containing $\mathbb{F}$ and making ρ a stopping time (i.e., for all $t \geq 0$, $\mathcal{F}_t^\rho = \mathcal{K}_{t+}^\rho$, where $\mathcal{K}_t^\rho = \mathcal{F}_t \vee \sigma(\rho \wedge t)$). One may interpret ρ as the instant of default of an issuer; the given filtration $\mathbb{F}$ can be thought of as the filtration of default-free prices, for which ρ is not a stopping time. Then, the filtration $\mathbb{F}^\rho$ is the defaultable market filtration used for the pricing of defaultable assets.

A few processes play a crucial role in our discussion:

- the $\mathbb{F}$ -supermartingale

$$Z_t^\rho = \mathbb{P}[\rho > t \mid \mathcal{F}_t] \quad (5)$$

chosen to be càdlàg, associated to ρ by Azéma [1];

- the $\mathbb{F}$ -dual optional projection of the process $\mathbf{1}_{\{\rho \leq t\}}$, denoted by A_t^ρ (see [7, 8] for a definition of dual optional projections); and
- the càdlàg martingale

$$\mu_t^\rho = \mathbb{E}[A_\infty^\rho \mid \mathcal{F}_t] = A_t^\rho + Z_t^\rho \quad (6)$$

Theorem 3 Every $\mathbb{F}$ -local martingale (M_t) , stopped at ρ , is an $\mathbb{F}^\rho$ -semimartingale, with canonical decomposition:

$$M_{t \wedge \rho} = \tilde{M}_t + \int_0^{t \wedge \rho} \frac{d\langle M, \mu^\rho \rangle_s}{Z_{s-}^\rho} \quad (7)$$

where $(\tilde{M}_t)$ is an $\mathbb{F}^\rho$ -local martingale.

The most interesting case in the theory of progressive enlargements of filtrations is when ρ is an *honest time* or equivalently the end of an $\mathbb{F}$ optional set Γ , that is,

$$\rho = \sup\{t : (t, \omega) \in \Gamma\} \quad (8)$$

Indeed, in this case, the pair of filtrations $(\mathbb{F}, \mathbb{F}^\rho)$ satisfies the (H') hypothesis: every $\mathbb{F}$ -local martingale (M_t) is an $\mathbb{F}^\rho$ -semimartingale, with canonical decomposition:

$$M_t = \tilde{M}_t + \int_0^{t \wedge \rho} \frac{d\langle M, \mu^\rho \rangle_s}{Z_{s-}^\rho} - \int_\rho^t \mathbf{1}_{\{\rho \leq t\}} \frac{d\langle M, \mu^\rho \rangle_s}{1 - Z_{s-}^\rho} \quad (9)$$

The next decomposition formulas are used for pricing in default models:

Proposition 1

1. Let $\xi \in L^1$. Then a càdlàg version of the martingale $\xi_t = \mathbb{E}[\xi \mid \mathcal{F}_t^\rho]$, on the set $\{t < \rho\}$, is given by:

$$\xi_t \mathbf{1}_{t < \rho} = \frac{1}{Z_t^\rho} \mathbf{1}_{t < \rho} \mathbb{E}[\xi \mathbf{1}_{t < \rho} \mid \mathcal{F}_t] \quad (10)$$

2. Let $\xi \in L^1$ and let ρ be an honest time. Then a càdlàg version of the martingale $\xi_t = \mathbb{E}[\xi \mid \mathcal{F}_t^\rho]$ is given as

$$\begin{aligned} \xi_t &= \frac{1}{Z_t^\rho} \mathbb{E}[\xi \mathbf{1}_{t < \rho} \mid \mathcal{F}_t] \mathbf{1}_{t < \rho} \\ &\quad + \frac{1}{1 - Z_t^\rho} \mathbb{E}[\xi \mathbf{1}_{t \geq \rho} \mid \mathcal{F}_t] \mathbf{1}_{t \geq \rho} \end{aligned} \quad (11)$$

The (H) Hypothesis

The (H) hypothesis, in contrast to the (H') hypothesis, is sometimes presented as a no-arbitrage condition in default models. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space satisfying the usual assumptions. Let $\mathbb{F}$ and $\mathbb{G}$ be two subfiltrations of $\mathcal{F}$, with

$$\mathcal{F}_t \subset \mathcal{G}_t \quad (12)$$

Brémaud and Yor [2] have proven the following characterization of the (H) hypothesis:

Theorem 4 The following are equivalent:

1. Every $\mathbb{F}$ -martingale is a $\mathbb{G}$ -martingale.
2. For all $t \geq 0$, the sigma fields $\mathcal{G}_t$ and $\mathcal{F}_\infty$ are independent conditionally on $\mathcal{F}_t$.

Remark 4 We also say that $\mathbb{F}$ is *immersed* in $\mathbb{G}$.

In the framework of the progressive enlargement of some filtration $\mathbb{F}$ with a random time ρ , the (H) hypothesis is equivalent to one of the following hypothesis [3]:

1. $\forall t$, the σ -algebras $\mathcal{F}_\infty$ and $\mathcal{F}_t^\rho$ are conditionally independent given $\mathcal{F}_t$.
2. For all bounded $\mathcal{F}_\infty$ measurable random variables $\mathbf{F}$ and all bounded $\mathcal{F}_t^\rho$ measurable random variables $\mathbf{G}_t$, we have

$$\mathbb{E}[\mathbf{F} \mathbf{G}_t \mid \mathcal{F}_t] = \mathbb{E}[\mathbf{F} \mid \mathcal{F}_t] \mathbb{E}[\mathbf{G}_t \mid \mathcal{F}_t] \quad (13)$$

3. For all bounded $\mathcal{F}_t^o$ measurable random variables $\mathbf{G}_t$:

$$\mathbb{E}[\mathbf{G}_t \mid \mathcal{F}_\infty] = \mathbb{E}[\mathbf{G}_t \mid \mathcal{F}_t] \quad (14)$$

4. For all bounded $\mathcal{F}_\infty$ measurable random variables $\mathbf{F}$,

$$\mathbb{E}[\mathbf{F} \mid \mathcal{F}_t^o] = \mathbb{E}[\mathbf{F} \mid \mathcal{F}_t] \quad (15)$$

5. For all $s \leq t$,

$$\mathbb{P}[\rho \leq s \mid \mathcal{F}_t] = \mathbb{P}[\rho \leq s \mid \mathcal{F}_\infty] \quad (16)$$

In view of applications to financial mathematics, a natural question is, how is the (H) hypothesis affected when we make an equivalent change of probability measure?

Proposition 2 *Let $\mathbb{Q}$ be a probability measure that is equivalent to $\mathbb{P}$ (on $\mathcal{F}$). Then, every $(\mathbb{F}, \mathbb{Q})$ -semimartingale is a $(\mathbb{G}, \mathbb{Q})$ -semimartingale.*

Now, define

$$\frac{d\mathbb{Q}}{d\mathbb{P}} \Big|_{\mathcal{F}_t} = R_t, \quad \frac{d\mathbb{Q}}{d\mathbb{P}} \Big|_{\mathcal{G}_t} = R'_t \quad (17)$$

If $Y = \frac{d\mathbb{Q}}{d\mathbb{P}}$, then the hypothesis (H) holds under $\mathbb{Q}$ if and only if

$$\forall X \geq 0, \quad X \in \mathcal{F}_\infty, \quad \frac{\mathbb{E}_{\mathbb{P}}[XY \mid \mathcal{G}_t]}{R'_t} = \frac{\mathbb{E}_{\mathbb{P}}[XY \mid \mathcal{F}_t]}{R_t} \quad (18)$$

In particular, when $\frac{d\mathbb{Q}}{d\mathbb{P}}$ is $\mathcal{F}_\infty$ measurable, $R_t = R'_t$ and the hypothesis (H) holds under $\mathbb{Q}$.

A decomposition formula is given below.

Theorem 5 *If (X_t) is a $(\mathbb{F}, \mathbb{Q})$ -local martingale, then the stochastic process*

$$I_X(t) = X_t + \int_0^t \frac{R'_{s-}}{R'_s}$$

$$\times \left(\frac{1}{R_{s-}} d[X, R]_s - \frac{1}{R'_{s-}} d[X, R']_s \right) \quad (19)$$

is a $(\mathbb{G}, \mathbb{Q})$ -local martingale.

References

- [1] Azéma, J. (1972). Quelques applications de la théorie générale des processus I, *Inventiones Mathematicae* **18**, 293–336.
- [2] Brémaud, P. & Yor, M. (1978). Changes of filtration and of probability measures, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **45**, 269–295.
- [3] Elliott, R.J., Jeanblanc, M. & Yor, M. (2000). On models of default risk, *Mathematical Finance* **10**, 179–196.
- [4] Jeulin, T. (1980). *Semi-martingales et Grossissements d'une Filtration*, Lecture Notes in Mathematics, Springer, Vol. 833.
- [5] Jeulin, T. & Yor, M. (eds) (1985). *Grossissements de Filtrations: Exemples et Applications*, Lecture Notes in Mathematics, Springer, Vol. 1118.
- [6] Mansuy, R. & Yor, M. (2006). *Random Times and (Enlargement of Filtrations) in a Brownian Setting*, Lecture Notes in Mathematics, Springer, Vol. 1873.
- [7] Nikeghbali, A. (2006). An essay on the general theory of stochastic processes, *Probability Surveys* **3**, 345–412.
- [8] Protter, P.E. (2005). *Stochastic Integration and Differential Equations*, 2nd Edition, version 2.1, Springer.
- [9] Revuz, D. & Yor, M. (1999). *Continuous Martingales and Brownian Motion*, 3rd Edition, Springer.
- [10] Stricker, C. (1977). Quasi-martingales, martingales locales, semimartingales et filtration naturelle, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **39**, 55–63.

Further Reading

Jacod, J. (1985). Grossissement initial, hypothèse (H') , et théorème de Girsanov, in *Grossissements de Filtrations: Exemples et Applications*, T. Jeulin & M. Yor, eds, Springer, pp. 15–35.

Related Articles

Compensators; Equivalence of Probability Measures; Martingale Representation Theorem; Martingales; Poisson Process; Semimartingale.

DELIA COCULESCU & ASHKAN NIKEGHBALI

Local Times

The most obvious way to measure the time that a random process X spends at a value b on a time interval $[0, t]$ is to compute $\int_0^t 1_{\{X_s=b\}} ds$. The problem is that this expression might be equal to 0, although the process X actually visits the value b . This is realized by the real Brownian motion (for a definition of this process, see **Lévy Processes**). Indeed, if we denote this process B , then for every fixed real b the set $\{s \geq 0 : B_s = b\}$ has 0 Lebesgue measure and is infinite (and uncountable). However, one can measure the time that B spends at b by using the notion of *local time* defined by

$$L_t^b = \lim_{\epsilon \rightarrow 0} \frac{1}{2\epsilon} \int_0^t 1_{\{|B_s - b| < \epsilon\}} ds \quad (1)$$

where the limit is a pathwise limit.

For a fixed b , the process $(L_t^b, t \geq 0)$ is an increasing process that only increases at times when B takes the value b . Under the assumption that B starts at 0, the processes $(L_t^0, t \geq 0)$ and $(2 \sup_{0 \leq s \leq t} B_s, s \geq 0)$ have the same law. This identity is due to Paul Lévy.

As b varies and t is fixed, one obtains the process $(L_t^b, b \in \mathbb{R})$, which actually represents the density of occupation time of B during the time interval $[0, t]$. This fact corresponds to the following formula called the *occupation time formula*

$$\int_0^t f(B_s) ds = \int_{\mathbb{R}} f(b) L_t^b db \quad (2)$$

for every measurable bounded function f . This formula provides a definition of the local time equivalent to definition (1). For a fixed t , one does not know, special times excepted, the law of the process $(L_t^b, b \in \mathbb{R})$, but a lot of trajectorial results are established. For example, from [6], we have

$$\liminf_{t \rightarrow \infty} \sup_{x \in \mathbb{R}} L_t^x (t^{-1} \log \log t)^{1/2} = c \quad (3)$$

with $0 < c < \infty$, and

$$\limsup_{t \rightarrow \infty} \sup_{x \in \mathbb{R}} L_t^x (t \log \log t)^{-1/2} = \sqrt{2} \quad (4)$$

One of these special times is T_a , the first hitting time by B of a given value a . The law of $(L_{T_a}^b, b \in \mathbb{R})$ is described by one of the famous *Ray–Knight Theorems* (see [8, Chapter XI]).

The local time of B can also be considered as a doubly indexed process. As such it is a.s. jointly continuous in b and t (see [9]) and deterministic functions on $\mathbb{R} \times \mathbb{R}_+$ can be integrated with respect to $(L_t^b, b \in \mathbb{R}, t \geq 0)$ (see **Itô's Formula**).

Local Time of a Semimartingale

Similarly to formula (2), one can define the local time process of a semimartingale Y (for the definition of a semimartingale, see **Stochastic Integrals**) by using the following occupation time formula:

$$\int_0^t f(Y_s) d[Y]_s^c = \int_{\mathbb{R}} f(b) L_t^b(Y) db \quad (5)$$

where $([Y]_s^c, s \geq 0)$ is the continuous part of the quadratic variation of Y also denoted by $\langle Y \rangle$ (for the definition see **Stochastic Integrals**). For a fixed b $(L_t^b(Y), t \geq 0)$ is a.s. continuous.

The obtained local time process $(L_t^b(Y), b \in \mathbb{R}, t \geq 0)$ satisfies the following formula, called *Tanaka's formula*:

$$\begin{aligned} |Y_t - b| &= |Y_0 - b| + \int_0^t \text{sgn}(Y_s - b) dY_s + L_t^b \\ &+ \sum_{0 < s \leq t} \{|Y_s - b| - |Y_{s-} - b| - \text{sgn}(Y_{s-} - b) \Delta Y_s\} \end{aligned} \quad (6)$$

where the function sgn is defined by $\text{sgn}(x) = 1_{x>0} - 1_{x \leq 0}$. Tanaka's formula actually provides a definition of the local time equivalent to formula (5). Thanks to this formula, Paul Lévy's identity is extended in [5] to the continuous semimartingales starting from 0 under the form

$$(L_t^0, t \geq 0) \stackrel{(\text{law})}{=} \left(2 \sup_{0 \leq s \leq t} \int_0^s \text{sgn}(-Y_u) dY_u, t \geq 0 \right) \quad (7)$$

One can actually see Tanaka's formula as an example of extension of Itô's formula (see **Itô's Formula**).

Local time is also involved in inequalities reminiscent of the Burkholder–Davis–Gundy ones. Indeed, in [2], it is shown that there exist two universal positive and finite constants c and C such that

$$cE[\sup_t |X_t|] \leq E[\sup_a L_\infty^a] \leq CE[\sup_t |X_t|] \quad (8)$$

for any continuous local martingale X with $X_0 = 0$.

Local Time of a Markov Process

One can define the local time process of a Markov process X at a value b of its state space only if b is *regular* for X (see **Markov Processes** for the definition of Markov process). This means that starting from b , the process X then visits b at arbitrarily small times. Not every Markov process has this property. For example, a real-valued Lévy process (see **Lévy Processes** for the definition of that process) has this property at every point if its characteristic exponent ψ satisfies [3, Chapter II]

$$\int_{-\infty}^{+\infty} \mathcal{R}\left(\frac{1}{1 + \psi(x)}\right) dx < \infty \quad (9)$$

When b is regular for X , there exists a unique (up to a multiplicative constant) increasing continuous *additive functional*, that is, an adapted process $(\ell_t^b(X), t \geq 0)$ starting from 0 such that

$$\ell_{t+s}^b(X) = \ell_t^b(X) + \ell_s^b(X) \circ \theta_t \quad (10)$$

increasing only at times when X takes the value b . This process is called the *local time* at b .

When it exists, the local time process $(\ell_t^b(X), b \in E, t \geq 0)$ of a Markov process X with state space E might be jointly continuous in b and t . A necessary and sufficient condition for that property is given in [1] for Lévy processes as follows: set $h(a) = \frac{1}{\pi} \int_{-\infty}^{\infty} (1 - \cos ab) \mathcal{R}(1/\psi(b)) db$ and $m(\varepsilon) = \int_{\mathbb{R}} da 1_{\{h(a) < \varepsilon\}}$ for $\varepsilon > 0$; then the considered Lévy process has a continuous local time process if

$$\int_{0+} \sqrt{\text{Log}\left(\frac{1}{m(\varepsilon)}\right)} d\varepsilon < \infty \quad (11)$$

This result concerning Lévy processes has been extended to symmetric Markov processes in [7] and to general Markov processes in [4].

We mention that under condition (9), the local time process of a Lévy process X satisfies the same

occupation time formula as for the real Brownian motion:

$$\int_0^t f(X_s) ds = \int_E f(b) \ell_t^b(X) db \quad (12)$$

In case a random process is both a Markov process with regular points and a semimartingale, it then admits two local time processes that are different (they might coincide as in the case of the Brownian motion). As an example, consider a symmetric stable process X with index in $(1, 2)$ (for definition see **Lévy Processes**). We have $[X]^c = 0$; hence, as a semimartingale, X has a local time process that identically equals 0. However, as a Markov process, X has a local time process that satisfies formula (12) and hence differs from 0. Besides, in this case, condition (11) is satisfied.

References

- [1] Barlow, M.T. (1988). Necessary and sufficient conditions for the continuity of local times for Levy processes, *Annals of Probability* **16**, 1389–1427.
- [2] Barlow, M.T. & Yor, M. (1981). (Semi-) Martingale inequalities and local times, *Zeitschrift für Wahrscheinlichkeitstheorie verw Gebiete* **55**, 237–254.
- [3] Bertoin, J. (1996). *Lévy Processes*, Cambridge University Press.
- [4] Eisenbaum, N. & Kaspi, H. (2007). On the continuity of local times of Borel right Markov processes, *Annals of Probability* **35**, 915–934.
- [5] El Karoui, N. & Chaleyat-Maurel, M. (1978). Un problème de réflexion et ses applications au temps local et aux équations différentielles stochastiques sur $\mathbb{R}$, in *Temps locaux—Astérisque*, Société mathématiques de France, Paris, Vol. 52–53, pp. 117–144.
- [6] Kesten, H. (1965). An iterated logarithm law for local time, *Duke Mathematical Journal* **32**, 447–456.
- [7] Marcus, M. & Rosen, J. (1995). Sample path properties of the local times of strongly symmetric Markov processes via Gaussian processes, *Annals of Probability* **20**, 1603–1684.
- [8] Revuz, D. & Yor, M. (1999). *Continuous Martingale and Brownian Motion*, 3rd Edition, Springer.
- [9] Trotter, H. (1958). A property of Brownian motion paths, *Illinois Journal of Mathematics* **2**, 425–433.

NATHALIE EISENBAUM

Stochastic Integrals

If H_t represents the number of shares of a certain asset held by an investor and X_t denotes the price of the asset, the gain on $[0, t]$ from the trading strategy H is often represented as

$$\int_0^t H_t dX_t \quad (1)$$

Here, our goal is to give a precise meaning to such “stochastic integrals”, where H and X are stochastic processes verifying appropriate assumptions.

Looking at the time-series data for price evolution of, say, a stock, one realizes that placing smoothness assumptions, such as differentiability, on the paths of X would be unrealistic. Consequently, this puts us in a situation where the theory of ordinary integration is no longer sufficient for our purposes. In what follows, we construct the stochastic integral $\int H dX$ for a class of integrands and integrators that are as large as possible while satisfying certain conditions. The stochastic processes that we use are defined on a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We always assume that all the processes are jointly measurable, that is, for any process $(Y_t)_{0 \leq t < \infty}$ the map $(t, \omega) \mapsto Y_t(\omega)$ is measurable with respect to $\mathcal{B}(\mathbb{R}_+) \times \mathcal{F}$, where $\mathcal{B}(\mathbb{R}_+)$ is the Borel σ -algebra on $[0, \infty)$. In addition, we are given a *filtration* $(\mathcal{F}_t)_{0 \leq t \leq \infty}$ (see **Filtrations**), which models the accumulation of our information over time. The filtration $(\mathcal{F}_t)_{0 \leq t \leq \infty}$ is usually denoted by $\mathbb{F}$ for convenience. We say that a jointly measurable process, Y , is *adapted* (or *$\mathbb{F}$ -adapted* if we need to specify the filtration) if $Y_t \in \mathcal{F}_t$ for all $t, 0 \leq t < \infty$. We assume that the following hypotheses hold true.

Assumption 1 *The filtered complete probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ satisfies the usual hypotheses (see **Filtrations**)*

Although the above hypotheses are restrictive, they are satisfied in many situations. The natural filtration of a Lévy process, in particular a Brownian motion, satisfies the usual hypotheses once completed. The same is true for the natural filtration of any counting process or “reasonable” strong Markov process (see, e.g., [7] for a more detailed discussion of the usual hypotheses and their consequences).

Having fixed the stochastic base on which all the processes are defined, let us go back to our primary task of defining the integral $\int H dX$. If X is a process of finite variation, the theory is that of *Lebesgue–Stieltjes integration*.

Definition 1 *A stochastic process X is said to be càdlàg (for continu à droite, limites à gauche from French) if it a.s. has sample paths that are right continuous on $[0, \infty)$ with left limits on $(0, \infty)$. Similarly, a stochastic process X is said to be càglàd (for continu à gauche, limites à droite) if it a.s. has sample paths that are left continuous on $(0, \infty)$ with right limits on $[0, \infty)$. We denote the space of adapted, càdlàg (respectively, càglàd) processes by $\mathbb{D}$ (respectively, $\mathbb{L}$).*

Definition 2 *Let X be a càdlàg process. For a given $\omega \in \Omega$, the variation of the path $X(\omega)$ on the compact interval $[a, b]$ is defined as*

$$\sup_{\pi \in P} \sum_{t_i \in \pi} |X_{t_{i+1}}(\omega) - X_{t_i}(\omega)| \quad (2)$$

where P is the set of all finite partitions of $[a, b]$. X is said to be a *finite variation (FV) process* if X is càdlàg and almost all paths of X have finite variation on each compact interval of $[0, \infty)$.

If X is an FV process, for fixed ω , it induces a signed measure on $\mathbb{R}_+$ and thus we can define a jointly measurable integral $\int_0^t H_s(\omega) dX_s(\omega)$ for any bounded and jointly measurable H . In other words, the integral $\int H dX$ can be defined path by path as a Lebesgue–Stieltjes integral, if H is a jointly measurable process such that $\int_0^t H_s(\omega) dX_s(\omega)$ exists and is finite for all $t > 0$, a.s.

Unfortunately, the set of FV processes is not rich enough if one wants to give a rigorous meaning to $\int H dX$ using only Stieltjes integration. When we replace X with, say, a Brownian motion, the theory of Stieltjes integration fails to work since the Brownian motion is known to have paths of infinite variation in every compact interval of $\mathbb{R}_+$. Therefore, one needs to develop a concept of integration with respect to a class of processes that is large enough to cover processes such as the Brownian motion or the more general Lévy processes, which find frequent applications in different fields.

To find the weakest conditions on X so that $\int H dX$ is well defined, we start with the simplest

possible form for the integrand H and work gradually to extend the stochastic integral to more complex integrands by imposing conditions on X but making sure that these conditions are as minimal as possible at the same time.

The simplest integrand one can think of is of the following form:

$$\begin{aligned} H_t(\omega) &= \mathbf{1}_{[S(\omega), T(\omega)]}(t) \\ &:= \begin{cases} 1 & \text{if } S(\omega) < t \leq T(\omega) \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (3)$$

where S and T are stopping times (see **Filtrations**) with respect to $\mathbb{F}$. In financial terms, this corresponds to a buy-and-hold strategy, whereby one unit of the asset is bought at, possibly random, time S and sold at time T . If X is the stochastic process representing the price of the asset, the net profit of such a trading strategy after time T is equal to $X_T - X_S$. This leads us to define $\int H dX$ as

$$\int_0^t H_s dX_s = X_{t \wedge T} - X_{t \wedge S} \quad (4)$$

where $t \wedge T := \min\{t, T\}$ for all t , $0 \leq t < \infty$, and stopping times T . Clearly, the process H in equation (3) has paths that are left continuous and possess right limits. We could similarly have defined $\int H dX$ for H of the form, say, $\mathbf{1}_{[S, T)}$. However, there is a good reason for insisting on paths that are continuous from the left on $(0, \infty)$ as we see in Example 1. Let us denote $\int_0^t H_s dX_s$ by $(H \cdot X)_t$.

Theorem 1 *Let H be of the form (3) and M be a martingale (see **Martingales**). Then $H \cdot M$ is a martingale.*

Later, we will see that the above theorem holds for a more general class of integrands so that the stochastic integrals preserve the martingale property. The following example shows why the left continuity for H is a reasonable restriction from a financial perspective.

Example 1 Let N be a Poisson process with intensity λ and define X by $X_t = \lambda t - N_t$. It is well known that X is a martingale. Suppose that there exists a traded asset with a price process given by X . Under normal circumstances, one should not be able to make arbitrage profits by trading in this

asset since its price does not change over time on average. Indeed, if H is of the form (3), then $H \cdot X$ is a martingale with expected value zero so that the traders earn zero profit on average, as expected. Now consider another strategy $H = \mathbf{1}_{[0, T_1]}$, where T_1 is the time of the first jump of N . Since X is an FV process, $H \cdot X$ is well defined as a Stieltjes integral and is given by $(H \cdot X)_t = \lambda(t \wedge T_1) > 0$, a.s., being the value of the portfolio at time t . Thus, this trading strategy immediately accumulates arbitrage profits. A moment of reflection reveals that such a trading strategy is not feasible under usual circumstances since it requires the knowledge of the time of a market crash, time T_1 in this case, before it happens. If we use $H = \mathbf{1}_{[0, T_1]}$ instead, this problem disappears.

Naturally, one will want the stochastic integral to be linear. Given a linear integral operator, we can define $H \cdot X$ for integrands that are linear combinations of processes of the form (3).

Definition 3 *A process H is said to be simple predictable if H has a representation*

$$H_t = H_0 \mathbf{1}_{\{0\}}(t) + \sum_{i=1}^n H_i \mathbf{1}_{(T_i, T_{i+1}]}(t) \quad (5)$$

where $0 = T_1 \leq \dots \leq T_{n+1} < \infty$ is a finite sequence of stopping times, $H_0 \in \mathcal{F}_0$, $H_i \in \mathcal{F}_{T_i}$, $1 \leq i \leq n$ with $|H_i| < \infty$, a.s., $0 \leq i \leq n$. The collection of simple predictable processes is denoted by $\mathbf{S}$.

Let $\mathbf{L}^0$ be the space of finite-valued random variables endowed with the topology of convergence in probability. Define the linear mapping $I_X : \mathbf{S} \mapsto \mathbf{L}^0$ as

$$\begin{aligned} I_X(H) &= (H \cdot X)_\infty \\ &:= H_0 X_0 + \sum_{i=1}^n H_i (X_{T_{i+1}} - X_{T_i}) \end{aligned} \quad (6)$$

where H has the representation given in equation (5). Note that this definition does not depend on the particular choice of representation for H .

Another property that the operator I_X must have is that it should satisfy some version of the bounded convergence theorem. This will inevitably place some restrictions on the stochastic process X . Thus, to have a large enough class of integrators, we choose a

reasonably weak version. A particularly weak version of the bounded convergence theorem is that the *uniform* convergence of H^n to H in $\mathbf{S}$ implies the convergence of $I_X(H^n)$ to $I_X(H)$ only in *probability*. Let $\mathbf{S}_u$ be the space $\mathbf{S}$ topologized by uniform convergence and recall that for a process X and a stopping time T , the notation X^T denotes the process $(X_{t \wedge T})_{t \geq 0}$.

Definition 4 A process X is a *total semimartingale* if X is càdlàg, adapted and $I_X : \mathbf{S}_u \mapsto \mathbf{L}^0$ is continuous. X is a *semimartingale* (see **Semimartingale**) if, for each $t \in [0, \infty)$, X^t is a total semimartingale.

This continuity property of I_X allows us to extend the definition of stochastic integrals to a class of integrands that is larger than $\mathbf{S}$ when the integrator is a semimartingale.

It follows from the definition of a semimartingale that semimartingales form a vector space. One can also show that all square integrable martingales and all adapted FV processes are semimartingales (see **Semimartingale**). Therefore, the sum of a square integrable martingale and an adapted FV process would also be a semimartingale. The converse of this statement is also “essentially” true. The precise statement is the following theorem.

Theorem 2 (Bichteler–Dellacherie Theorem). Let X be a semimartingale. Then there exist processes M, A , with $M_0 = A_0 = 0$ such that

$$X_t = X_0 + M_t + A_t \quad (7)$$

where M is a local martingale and A is an adapted FV process.

Here, we emphasize that this decomposition is not necessarily unique. Indeed, suppose that X has the decomposition $X = X_0 + M + A$ and the space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ supports a Poisson process N with intensity λ . Then $Y_t = N_t - \lambda t$ will define a martingale, which is also an FV process. Therefore, X can also be written as $X = X_0 + (M + Y) + (A - Y)$. The reason for the nonuniqueness is the existence of martingales that are of finite variation. However, if X has a decomposition $X = X_0 + M + A$, where M is a local martingale and A is *predictable*^a and FV with $M_0 = A_0 = 0$, then such a decomposition is unique since all predictable local martingales that are of finite variation have to be constant.

Arguably, Brownian motion is the most well known of all semimartingales. In the following section, we develop stochastic integration with respect to a Brownian motion.

L^2 Theory of Stochastic Integration with Respect to Brownian Motion

We assume that there exists a Brownian motion, B , on $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ with $B_0 = 0$, and that $\mathcal{F}_0$ only contains the $(\mathcal{F}, \mathbb{P})$ -null sets. First, we define the notion of *predictability*, which is the key concept in defining the stochastic integral.

Definition 5 The *predictable σ -algebra* $\mathcal{P}$ on $[0, \infty) \times \Omega$ is defined to be the smallest σ -algebra on $[0, \infty) \times \Omega$ with respect to which every adapted càglàd process is measurable. A process is said to be *predictable* if it is a $\mathcal{P}$ -measurable map from $[0, \infty) \times \Omega$ to $\mathbb{R}$.

Clearly, $\mathbf{S} \subset \mathcal{P}$. Actually, there is more to this as is shown by the next theorem.

Theorem 3 Let $\mathbf{bS}$ be the set of elements of $\mathbf{S}$ that are bounded a.s. Then, $\mathcal{P} = \sigma(\mathbf{bS})$, that is, $\mathcal{P}$ is generated by the processes in $\mathbf{bS}$.

By linearity of the stochastic integral and Theorem 1 and using the fact that Brownian motion has increments independent from the past with a certain Gaussian distribution, we have the following.

Theorem 4 Let $H \in \mathbf{bS}$ and define $(H \cdot B)_t = (H \cdot B^t)_\infty$, that is, $(H \cdot B)_t$ is the stochastic integral of H with respect to B^t . Then $H \cdot B$ is a martingale and

$$\mathbb{E}[(H \cdot B)_t^2] = \int_0^t \mathbb{E}[H_s^2] \, ds \quad (8)$$

In the following, we construct the stochastic integral with respect to Brownian motion for a subset of predictable processes. To keep the exposition simple, we restrict our attention to a finite interval $[0, T]$, where T is arbitrary but deterministic. Define

$$\mathcal{L}^2(B^T) := \left\{ H \in \mathcal{P} : \int_0^T \mathbb{E}[H_s^2] \, ds < \infty \right\} \quad (9)$$

which is a Hilbert space. Note that $\mathbf{bS} \subset \mathcal{L}^2(B^T)$. Letting $L^2(\mathcal{F}_T)$ denote the space of square integrable

$\mathcal{F}_T$ -measurable random variables, Theorem 4 now implies the map

$$I_{B^T} : \mathbf{bS} \mapsto \mathcal{L}^2(\mathcal{F}_T) \quad (10)$$

defined by

$$I_{B^T}(H) = (H \cdot B)_T \quad (11)$$

is an isometry. Consequently, we can extend the definition of the stochastic integral uniquely to the closure of $\mathbf{bS}$ in $\mathcal{L}^2(B^T)$. An application of monotone class theorem along with Theorem 3 yields that the closure is the whole $\mathcal{L}^2(B^T)$.

Theorem 5 *Let $H \in \mathcal{L}^2(B^T)$. Then the Itô integral $(H \cdot B)_T$ of H with respect to B^T is the image of H under the extension of the isometry I_{B^T} to the whole of $\mathcal{L}^2(B^T)$. In particular,*

$$\mathbb{E}[(H \cdot B)_T^2] = \int_0^T \mathbb{E}[H_s^2] \, ds \quad (12)$$

Moreover, the process Y defined by $Y_t = (H \cdot B)_{t \wedge T}$ is a square integrable martingale.

The property (12) is often called the *Itô isometry*.

Stochastic Integration with Respect to General Semimartingales

In the previous section, we developed the stochastic integration for Brownian motion over the interval $[0, T]$. We need to mention here that the method employed works not only for Brownian motion but also for any martingale M that is square integrable over $[0, T]$, the latter case requiring some extra effort mainly for establishing the existence of the so-called *quadratic variation* process associated with M . This would, in turn, allow us to extend the definition of the stochastic integral with respect to X of the form $X = M + A$, where M is a square integrable martingale and A is a process of finite variation on compacts by defining, under some conditions on H ,

$$H \cdot X = H \cdot M + H \cdot A \quad (13)$$

where $H \cdot A$ can be computed as a path-by-path Lebesgue–Stieltjes integral. In this section, we establish the stochastic integral with respect to a general semimartingale. The idea would be similar to the construction of the stochastic integral with respect to

Brownian motion. We show that the integral operator is a continuous mapping from the set of simple predictable process into an appropriate space so that we can extend the set of possible integrands to the closure of $\mathbf{S}$ in a certain topology.

Definition 6 *A sequence of processes $(H^n)_{n \geq 1}$ converges to a process H uniformly on compacts in probability (UCP) if, for each $t > 0$, $\sup_{0 \leq s \leq t} |H_s^n - H_s|$ converges to 0 in probability.*

The following result is not surprising and one can refer to, for example, [7] for a proof.

Theorem 6 *The space $\mathbf{S}$ is dense in $\mathbb{L}$ under the UCP topology.*

The following mapping is key to defining the stochastic integral with respect to a general semimartingale.

Definition 7 *For $H \in \mathbf{S}$ and X being a càdlàg process, define the linear mapping $J_X : \mathbf{S} \mapsto \mathbb{D}$ by*

$$J_X(H) = H_0 X_0 + \int_{i=1}^n H_i (X^{T_{i+1}} - X^{T_i}) \quad (14)$$

where H has the representation as in equation (5).

Note the difference between J_X and I_X . I_X maps processes into random variables, whereas J_X maps processes into processes.

Definition 8 *For $H \in \mathbf{S}$ and X being an adapted càdlàg process, we call $J_X(H)$ the stochastic integral of H with respect to X .*

Observe that $J_X(H)_t = I_{X^t}(H)$. This property, combined with the definition of a semimartingale, yields the following continuity property for J_X .

Theorem 7 *Let X be a semimartingale and $\mathbf{S}_{\text{UCP}}$ (respectively $\mathbb{D}_{\text{UCP}}$) denote the space $\mathbf{S}$ (respectively, $\mathbb{D}$) endowed with the UCP topology. Then the mapping $J_X : \mathbf{S}_{\text{UCP}} \mapsto \mathbb{D}_{\text{UCP}}$ is continuous.*

Using Theorem 6, we can now extend the integration operator J_X from $\mathbf{S}$ to $\mathbb{L}$ by continuity, since $\mathbb{D}_{\text{UCP}}$ is a complete metric space^b.

Definition 9 *Let X be a semimartingale. The continuous linear mapping $J_X : \mathbb{L}_{\text{UCP}} \mapsto \mathbb{D}_{\text{UCP}}$ obtained as the extension of $J_X : \mathbf{S}_{\text{UCP}} \mapsto \mathbb{D}_{\text{UCP}}$ is called the stochastic integral.*

Note that, in contrast to the L^2 theory utilized in the previous section, we do not need to impose any integrability conditions on either X or H to establish the existence of the stochastic integral $H \cdot X$ as long as H remains in $\mathbb{L}$. The above continuity property of the stochastic integrals moreover allows us to approximate the $H \cdot X$ using the Riemann sums.

Definition 10 Let σ denote a finite sequence of finite stopping times:

$$0 = T_0 \leq T_1 \leq \dots \leq T_k < \infty. \quad (15)$$

The sequence of σ is called a random partition. A sequence of random partitions σ_n

$$\sigma_n : 0 = T_0^n \leq T_1^n \leq \dots \leq T_{k_n}^n \quad (16)$$

is said to tend to identity if

1. $\lim_{n \rightarrow \infty} \sup_j T_j^n = \infty$, a.s. and
2. $\sup_j |T_{j+1}^n - T_j^n|$ converges to 0 a.s.

Let Y be a process and σ be a random partition. Define the process

$$Y^\sigma := Y_0 \mathbf{1}_{\{0\}} + \sum_j Y_{T_j} \mathbf{1}_{(T_j, T_{j+1}]} \quad (17)$$

Consequently, if Y is in $\mathbb{D}$ or $\mathbb{L}$

$$Y^\sigma \cdot X = Y_0 X_0 + \sum_j Y_{T_j} (X^{T_{j+1}} - X^{T_j}) \quad (18)$$

for any semimartingale X .

Theorem 8 Let X be a semimartingale and let $\int_{0+}^t H_s dX_s$ denote $(H \cdot X)_t - H_0 X_0$ for any $H \in \mathbb{L}$. If Y is a process in $\mathbb{D}$ or in $\mathbb{L}$, and (σ_n) is a sequence of random partitions tending to identity, then the process $\left(\int_{0+}^t Y_s^{\sigma_n} dX_s \right)_{t \geq 0}$ converges to the stochastic integral $(Y_-) \cdot X$ in UCP, where Y_- is the process defined as $(Y_-)_s = \lim_{r \rightarrow s, r < s} Y_r$, for $s > 0$, and $(Y_-)_0 = 0$.

Example 2 As an application of the above theorem, we calculate $\int_0^t B_s dB_s$, where B is a standard Brownian motion with $B_0 = 0$. Let (σ_n) be a sequence of random partitions of the form (16) tending to identity and let $B^n = B^{\sigma_n}$. Note that

$$\begin{aligned} \int_0^t B_s^n dB_s &= \sum_{\substack{t_j \in \sigma_n \\ t_j < t}} B_{t_j} (B_{t \wedge t_{j+1}} - B_{t_j}) \\ &= \sum_{\substack{t_j \in \sigma_n \\ t_j < t}} \frac{1}{2} (B_{t \wedge t_{j+1}} + B_{t_j}) (B_{t \wedge t_{j+1}} - B_{t_j}) \\ &\quad - \sum_{\substack{t_j \in \sigma_n \\ t_j < t}} \frac{1}{2} (B_{t \wedge t_{j+1}} - B_{t_j})^2 \\ &= \frac{1}{2} B_{(t \wedge T_{k_n}^n)}^2 - \frac{1}{2} \sum_{\substack{t_j \in \sigma_n \\ t_j < t}} (B_{t_{j+1}} - B_{t_j})^2 \end{aligned} \quad (19)$$

As n tends to ∞ , the sum^c in equation (19) is known to converge to t . Obviously, $B_{T_{k_n}^n \wedge t}^2$ tends to B_t^2 since σ_n tends to identity. Thus, we conclude via Theorem 8 that

$$\int_0^t B_s dB_s = \frac{1}{2} B_t^2 - \frac{t}{2} \quad (20)$$

since B is continuous with $B_0 = 0$. Thus, the integration rules for a stochastic integral are quite different from those for an ordinary integral. Indeed, if A were a continuous process of finite variation with $A_0 = 0$, then the Riemann–Stieltjes integral of $A \cdot A$ will yield the following formula:

$$\int_0^t A_s dA_s = \frac{1}{2} A_t^2 \quad (21)$$

As in the case of Brownian motion, stochastic integration with respect to a semimartingale preserves the martingale property.

Theorem 9 Let $H \in \mathbb{L}$ such that $\lim_{t \downarrow 0} |H_t| < \infty$ and X be a local martingale (see **Martingales**). Then $H \cdot X$ is also a local martingale.

Next, we would like to weaken the restriction that an integrand must be in $\mathbb{L}$. If we want the stochastic integral to still preserve the martingale property with this extended class of integrands, we inevitably need to restrict our attention to predictable processes. To see this, consider the process $H = \mathbf{1}_{[0, T_1)}$ in Example 1. This process is not predictable since the jump times of a Poisson process are not predictable stopping times. As we have shown in Example 1, the

integral of H with respect to a particular martingale is not a martingale.

Before we allow more general predictable integrands in a stochastic integral, we need to develop the notion of *quadratic variation* of a semimartingale. This is discussed in the following section.

Properties of Stochastic Integrals

In this section, H denotes an element of $\mathbb{L}$ and X denotes a semimartingale. For a process $Y \in \mathbb{D}$, we define $\Delta Y_t = Y_t - Y_{t-}$, the jump at t . Recall that two processes Y and Z are said to be *indistinguishable* if $\mathbb{P}\{\omega : Y_t(\omega) = Z_t(\omega), \forall t\} = 1$.

Theorem 10 *Let T be a stopping time. Then $(H \cdot X)^T = H \mathbf{1}_{[0, T]} \cdot X = H \cdot (X^T)$.*

Theorem 11 *The jump process $(\Delta(H \cdot X))_{t \geq 0}$ is indistinguishable from $(H_t \Delta X_t)_{t \geq 0}$.*

In finance theory, one often needs to work under the so-called *risk-neutral* measure rather than the empirical or objective measure $\mathbb{P}$. Recall that definitions of a semimartingale and its stochastic integral are given in spaces topologized by convergence in probability. Thus, one may wonder whether the value of a stochastic integral remains unchanged under an equivalent change of measure. The following theorem shows that this is indeed the case. Let $\mathbb{Q}$ be another probability measure on $(\Omega, \mathcal{F})$ and let $H_{\mathbb{Q}} \cdot X$ denote the stochastic integral of H with respect to X computed under $\mathbb{Q}$.

Theorem 12 *Let $\mathbb{Q} \ll \mathbb{P}$. Then, $H_{\mathbb{Q}} \cdot X$ is indistinguishable from $H_{\mathbb{P}} \cdot X$.*

Theorem 13 *Let $\mathbb{G} = (\mathcal{G}_t)_{t \geq 0}$ be another filtration such that H is in both $\mathbb{L}(\mathbb{G})$ and $\mathbb{L}(\mathbb{F})$, and such that X is also a $\mathbb{G}$ -semimartingale. Then $H_{\mathbb{G}} \cdot X$ is indistinguishable from $H_{\mathbb{F}} \cdot X$.*

The following theorem shows the stochastic integral is an extension of the Lebesgue–Stieltjes integral.

Theorem 14 *If X is an FV process, then $H \cdot X$ is indistinguishable from the Lebesgue–Stieltjes integral, computed path by path. Consequently, $H \cdot X$ is an FV process.*

Theorem 15 *The stochastic integral is associative. That is, $H \cdot X$ is also a semimartingale and if $G \in \mathbb{L}$*

$$G \cdot (H \cdot X) = (GH) \cdot X \quad (22)$$

Definition 11 *The quadratic variation process of X , denoted by $[X, X] = ([X, X]_t)_{t \geq 0}$, is defined as*

$$[X, X] = X^2 - 2X_- \cdot X \quad (23)$$

Recall that $X_{0-} = 0$. Let Y be another semimartingale. The quadratic covariation of X and Y , denoted by $[X, Y]$, is defined as

$$[X, Y] = XY - Y_- \cdot X - X_- \cdot Y \quad (24)$$

Since X_- (and Y_-) belongs to $\mathbb{L}$, we can use Theorem 8 to deduce the following.

Theorem 16 *Let Y be a semimartingale. The quadratic covariation $[X, Y]$ of X and Y is an adapted càdlàg process that satisfies the following:*

1. $[X, Y]_0 = X_0 Y_0$ and $\Delta[X, Y] = \Delta X \Delta Y$.
2. If (σ_n) is a sequence of partitions tending to identity, then

$$\begin{aligned} X_0 Y_0 + \sum_j (X^{T_{j+1}^n} - X^{T_j^n}) \\ \times (Y^{T_{j+1}^n} - Y^{T_j^n}) \rightarrow [X, Y] \end{aligned} \quad (25)$$

with convergence in UCP, where σ_n is of the form (16).

3. If T is any stopping time, then $[X^T, Y] = [X, Y^T] = [X, Y]^T$. Moreover, $[X, X]$ is increasing.

Since $[X, X]$ is increasing and càdlàg by definition, we immediately deduce that $[X, X]$ is of finite variation. Moreover, the following polarization identity

$$[X, Y] = \frac{1}{2}([X + Y, X + Y] - [X, X] - [Y, Y]) \quad (26)$$

reveals that $[X, Y]$ is the difference of two increasing processes; therefore, $[X, Y]$ is an FV process as well. This, in turn, implies XY is also a semimartingale and yields the *integration by parts* formula:

$$X_t Y_t = (X_- \cdot Y)_t + (Y_- \cdot X)_t + [X, Y]_t \quad (27)$$

When X and Y are FV processes, the classical integration by parts formula reads as follows:

$$X_t Y_t = X_0 Y_0 + (X_- \cdot Y)_t + (Y_- \cdot X)_t + \sum_{0 < s \leq t} \Delta X_s \Delta Y_s \quad (28)$$

Therefore, if X or Y is a continuous processes of finite variation, then $[X, Y] = X_0 Y_0$. In particular, if X is a continuous FV process, then its quadratic variation is equal to X_0^2 .

Theorem 17 *Let X and Y be two semimartingales, and let H and K be two measurable processes. Then one has a.s.*

$$\begin{aligned} & \int_0^\infty |H_s| |K_s| \, d[X, Y]_s \\ & \leq \left(\int_0^\infty H_s^2 \, d[X, X]_s \right)^{\frac{1}{2}} \left(\int_0^\infty K_s^2 \, d[Y, Y]_s \right)^{\frac{1}{2}} \end{aligned} \quad (29)$$

The above inequality is called *Kunita–Watanabe inequality*. An immediate consequence of this inequality is that if X or Y has zero quadratic variation, then $[X, Y] = 0$. The following theorem follows from the definition of quadratic variation and Theorem 9.

Theorem 18 *Let X be a local martingale. Then, $X^2 - [X, X]$ is a local martingale. Moreover, $[X, X]$ is the unique adapted càdlàg and FV process A such that $X^2 - A$ is a local martingale and $\Delta A = (\Delta X)^2$ with $A_0 = X_0^2$.*

Note that the uniqueness in the above theorem is lost if we do not impose $\Delta A = (\Delta X)^2$. Roughly speaking, the above theorem infers $\mathbb{E}(X_t^2) = \mathbb{E}([X, X]_t)$ when X is a martingale. The following theorem formalizes this intuition.

Corollary 1 *Let X be a local martingale. Then, X is a martingale with $\mathbb{E}(X_t^2) < \infty$, for all $t \geq 0$, if and only if $\mathbb{E}([X, X]_t) < \infty$, for all $t \geq 0$. If $\mathbb{E}([X, X]_t) < \infty$, then $\mathbb{E}(X_t^2) = \mathbb{E}([X, X]_t)$.*

The following corollary to Theorem 18 is of fundamental importance in the theory of martingales.

Corollary 2 *Let X be a continuous local martingale, and $S \leq T \leq \infty$ be stopping times. If X has paths of finite variation on the stochastic interval (S, T) ,*

then X is constant on $[S, T]$. Moreover, if $[X, X]$ is constant on $[S, T] \cap [0, \infty)$, then X is also constant there.

The following result is quite handy when it comes to the calculation of the quadratic covariation of two stochastic integrals.

Theorem 19 *Let Y be a semimartingale and $K \in \mathbb{L}$. Then*

$$[H \cdot X, K \cdot Y]_t = \int_0^t H_s K_s \, d[X, Y]_s \quad (30)$$

In the following section, we define the stochastic integral for predictable integrals. However, we already have all the results to present the celebrated *Itô's formula*.

Theorem 20 (Itô's Formula). *Let X be a semimartingale and f be a C^2 real function. Then $f(X)$ is again a semimartingale and the following formula holds:*

$$\begin{aligned} f(X_t) - f(X_0) &= \int_{0+}^t f'(X_{s-}) \, dX_s \\ &+ \frac{1}{2} \int_{0+}^t f''(X_{s-}) \, d[X, X]_s \\ &+ \sum_{0 < s \leq t} \left\{ f(X_s) - f(X_{s-}) - f'(X_{s-}) \Delta X_s \right. \\ &\quad \left. - \frac{1}{2} f''(X_{s-}) (\Delta X_s)^2 \right\} \end{aligned} \quad (31)$$

Stochastic Integration for Predictable Integrands

In this section, we weaken the hypothesis that $H \in \mathbb{L}$ in order for $H \cdot X$ to be well defined for a semimartingale X . As explained earlier, we restrict our attention to predictable processes since we want the stochastic integral to preserve the martingale property. We will not be able to show the existence of stochastic integral $H \cdot X$ for all $H \in \mathcal{P}$ but, as in the section L^2 Theory of Stochastic Integration with Respect to Brownian Motion, we give a meaning to $H \cdot X$ for the appropriately integrable processes in $\mathcal{P}$. First, we assume that X is a special semimartingale, that is, there exist processes M and A such that M is a

local martingale and A is predictable and of finite variation with $M_0 = A_0 = 0$ and $X = X_0 + M + A$. This decomposition of a special semimartingale is unique and called the *canonical decomposition*. Without loss of generality, let us assume that $X_0 = 0$.

Definition 12 Let X be a special semimartingale with the canonical decomposition $X = M + A$. The $\mathcal{H}^2$ norm of X is defined as

$$\|X\|_{\mathcal{H}^2} := \| [M, M]_{\infty}^{1/2} \|_{L^2} + \left\| \int_0^{\infty} |dA_s| \right\|_{L^2} \quad (32)$$

The space of $\mathcal{H}^2$ semimartingales consists of special semimartingales with finite $\mathcal{H}^2$ norm. We write $X \in \mathcal{H}^2$ to indicate that X belongs to the space of $\mathcal{H}^2$ semimartingales.

One can show that the space of $\mathcal{H}^2$ semimartingales is a Banach space, which is the key property to extend the definition of stochastic integrals for a more general class of integrands. Let $\mathbf{b}\mathbb{L}$ denote the space of bounded adapted processes with càglàd paths and $\mathbf{b}\mathcal{P}$ denote the space of bounded predictable processes.

Definition 13 Let $X \in \mathcal{H}^2$ with the canonical decomposition $X = N + A$ and let $H, J \in \mathbf{b}\mathcal{P}$. We define the metric $d_X(H, J)$ as

$$d_X(H, J) := \left\| \left(\int_0^{\infty} (H_s - J_s)^2 d[M, M]_s \right)^{1/2} \right\|_{L^2} + \left\| \int_0^{\infty} |H_s - J_s| dA_s \right\|_{L^2} \quad (33)$$

From the monotone class theorem, we obtain the following.

Theorem 21 For $X \in \mathcal{H}^2$, the space $\mathbf{b}\mathbb{L}$ is dense in $\mathbf{b}\mathcal{P}$ under $d_X(\cdot, \cdot)$.

It is straightforward to show that if $H \in \mathbf{b}\mathbb{L}$ and $X \in \mathcal{H}^2$, then $H \cdot X \in \mathcal{H}^2$. The following is an immediate consequence of the definition of $d_X(\cdot, \cdot)$.

Theorem 22 Let $X \in \mathcal{H}^2$ and $(H^n) \subset \mathbf{b}\mathbb{L}$ such that (H^n) is Cauchy under $d_X(\cdot, \cdot)$. Then, $(H^n \cdot X)$ is Cauchy in $\mathcal{H}^2$.

Moreover, it is easy to show that if $(H^n) \subset \mathbf{b}\mathbb{L}$ and $(J^n) \subset \mathbf{b}\mathbb{L}$ converge to the same limit under $d_X(\cdot, \cdot)$, then $(H^n \cdot X)$ and $(J^n \cdot X)$ converge to the same limit in $\mathcal{H}^2$. Thus, we can now define the stochastic integral $H \cdot X$ for any $H \in \mathbf{b}\mathcal{P}$.

Definition 14 Let $X \in \mathcal{H}^2$ and $H \in \mathbf{b}\mathcal{P}$. Let $(H^n) \subset \mathbf{b}\mathbb{L}$ such that $\lim_{n \rightarrow \infty} d_X(H^n, H) = 0$. The stochastic integral $H \cdot X$ is the unique semimartingale $Y \in \mathcal{H}^2$ such that $\lim_{n \rightarrow \infty} H^n \cdot X = Y$ in $\mathcal{H}^2$.

Note that if B is a standard Brownian motion, B is not in $\mathcal{H}^2$ but $B^T \in \mathcal{H}^2$ for any deterministic and finite T . Therefore, for any $H \in \mathbf{b}\mathcal{P}$, $H \cdot B^T$ is well defined. Moreover, $H \in \mathbf{b}\mathcal{P}$ implies $H \in \mathcal{L}^2(B^T)$ where $\mathcal{L}^2(B^T)$ is the space defined in the section L^2 Theory of Stochastic Integration with Respect to Brownian Motion. One can easily check that the stochastic integral $H \cdot B^T$ defined by Definition 14 is indistinguishable from the stochastic integral $H \cdot B^T$ defined in the section L^2 Theory of Stochastic Integration with Respect to Brownian Motion. Clearly, $\mathbf{b}\mathcal{P}$ is strictly contained in $\mathcal{L}^2(B^T)$, and we know from the section L^2 Theory of Stochastic Integration with Respect to Brownian Motion that it is possible to define the stochastic integral with respect to B^T for any process in $\mathcal{L}^2(B^T)$. Thus, it is natural to ask whether we can extend the stochastic integral given by Definition 14 to integrands that satisfy a certain square integrability condition.

Definition 15 Let $X \in \mathcal{H}^2$ with the canonical decomposition $X = M + A$. We say that $H \in \mathcal{P}$ is $(\mathcal{H}^2, X)$ integrable if

$$\mathbb{E} \left(\int_0^{\infty} H_s^2 d[M, M]_s \right) + \mathbb{E} \left(\left\{ \int_0^{\infty} |H_s| dA_s \right\}^2 \right) < \infty \quad (34)$$

It can be shown that if $H \in \mathcal{P}$ is $(\mathcal{H}^2, X)$ integrable, $(H^n \cdot X)$ is a Cauchy sequence in $\mathcal{H}^2$ where $H^n = H \mathbf{1}_{\{|H| \leq n\}}$ is in $\mathbf{b}\mathcal{P}$, which means that we can define the stochastic integral for such H .

Definition 16 Let $X \in \mathcal{H}^2$ and let $H \in \mathcal{P}$ be $(\mathcal{H}^2, X)$ integrable. The stochastic integral $H \cdot X$ is defined to be the $\lim_{n \rightarrow \infty} H^n \cdot X$, with convergence in $\mathcal{H}^2$, where $H^n = H \mathbf{1}_{\{|H| \leq n\}}$.

In the case $X = B^T$, $M = B^T$, and $A = 0$; therefore, H being $(\mathcal{H}^2, X)$ integrable is equivalent to the condition

$$\int_0^T \mathbb{E}(H_s^2) ds < \infty \quad (35)$$

which gives exactly the elements of $\mathcal{L}^2(B^T)$.

So far, we have been able to define the stochastic integral with predictable integrands only for semimartingales in $\mathcal{H}^2$. This seems to be a major restriction. However, as the following theorem shows, it is not. Recall that for a stopping time T , $X^{T-} = X\mathbf{1}_{[0, T)} + X_T - \mathbf{1}_{[T, \infty]}$.

Theorem 23 *Let X be a semimartingale, $X_0 = 0$. Then X is prelocally in $\mathcal{H}^2$. That is, there exists a nondecreasing sequence of stopping times (T^n) , $\lim_{n \rightarrow \infty} T^n = \infty$ a.s., such that $X^{T^n-} \in \mathcal{H}^2$, for each $n \geq 1$.*

Definition 17 *Let X be a semimartingale and $H \in \mathcal{P}$. The stochastic integral $H \cdot X$ is said to exist if there exists a sequence of stopping times (T^n) increasing to ∞ a.s. such that $X^{T^n-} \in \mathcal{H}^2$, for each $n \geq 1$, and such that H is $(\mathcal{H}^2, X^{T^n-})$ integrable for each $n \geq 1$. In this case, we write $H \in L(X)$ and define the stochastic integral as*

$$H \cdot X = H \cdot (X^{T^n-}), \quad \text{on } [0, T^n] \quad (36)$$

for each n .

A particular case when $H \cdot X$ is well defined is when H is locally bounded.

Theorem 24 *Let X be a semimartingale and $H \in \mathcal{P}$ be locally bounded. Then, $H \in L(X)$.*

We also have the martingale preservation property.

Theorem 25 *Let M be a local martingale and $H \in \mathcal{P}$ be locally bounded. Then, $H \cdot M$ is a local martingale.*

The general result that M a local martingale and $H \in L(M)$ implies that $H \cdot M$ is a local martingale is not true. The following example is due to Emery and can be taken as a starting point for a study of *sigma-martingales* (see **Equivalent Martingale Measures**).

Example 3 Let T be an exponential random variable with parameter 1 and let U be an independent

random variable with $\mathbb{P}(U = 1) = \mathbb{P}(U = -1) = 1/2$, and set $X = U\mathbf{1}_{[T, \infty)}$. Then, X is a martingale in its own filtration. Let H be defined as $H_t = \frac{1}{t}\mathbf{1}_{\{t > 0\}}$. H is a deterministic predictable integral. Note that H is not locally bounded, being only continuous on $(0, \infty)$. $H \cdot X$ exists as a Lebesgue–Stieltjes integral since X has paths of finite variation. However, $H \cdot X$ is not a local martingale since, for any stopping time S with $P(S > 0) > 0$, $\mathbb{E}(|(H \cdot X)_S|) = \infty$.

When M is a continuous local martingale, the theory becomes nicer.

Theorem 26 *Let M be a continuous local martingale and let $H \in \mathcal{P}$ be such that $\int_0^t H_s^2 d[M, M]_s < \infty$, for each $t \geq 0$. Then $H \in L(M)$ and $H \cdot M$ is a continuous local martingale.*

The question may arise as to whether the properties of stochastic integral stated for left-continuous integrands in the section Properties of Stochastic Integrals continue to hold when we allow predictable integrands. The answer is positive except for Theorems 13 and 14. Still, if X is a semimartingale with paths of finite variation on compacts and if $H \in L(X)$ is such that the Stieltjes integral $\int_0^t |H_s| |dX_s|$ exists a.s. for each $t \geq 0$, then the stochastic integral $H \cdot X$ agrees with the Stieltjes integral computed path by path. However, $H \cdot X$ is not necessarily an FV process. See [7, Exercise 45 in Chapter IV] of [7] for a counterexample. The analogous result for Theorem 13 is the following, which is particularly useful when one needs to study asymmetric information in financial markets where some traders possess extra information compared to others.

Theorem 27 *Let $\mathbb{G}$ be another filtration satisfying the usual hypotheses and suppose that $\mathcal{F}_t \subset \mathcal{G}_t$, each $t \geq 0$, and that X remains a semimartingale with respect to $\mathbb{G}$. Let H be locally bounded and predictable for $\mathbb{F}$. Then H is locally bounded and predictable for $\mathbb{G}$, the stochastic integral $H_{\mathbb{G}} \cdot X$ exists and is equal to $H_{\mathbb{F}} \cdot X$.*

It is important to have H locally bounded in the above theorem; see [4] for a counterexample in the context of enlargement of filtrations.

We end this section with the *dominated convergence theorem for stochastic integrals*.

Theorem 28 *Let X be a semimartingale and $(H^n) \subset \mathcal{P}$ be a sequence converging a.s. to a limit $H \in \mathcal{P}$. If there exists a process $G \in L(X)$ such that $|H^n| \leq G$, for all n , then $H^n \in L(X)$ for all n , $H \in L(X)$ and $(H^n \cdot X)$ converges to $H \cdot X$ in UCP.*

Concluding Remarks

In this article, we used the approach of Protter [7] to define the semimartingale as a good integrator and construct its stochastic integral. Another approach that is closely related is given by Chou *et al.* [1], who developed the stochastic integration for general predictable integrands with respect to a semimartingale in a space endowed with the semimartingale topology. Historically, the stochastic integral was first proposed for Brownian motion by Itô [3], then for continuous martingales, then for square integrable martingales, and finally for càdlàg processes that can be written as the sum of a locally square integrable local martingale and an FV process by J.L. Doob, H. Kunita, S. Watanabe, P. Courrège, P.A. Meyer, and others. Later in 1970, Doléans-Dade and Meyer [2] showed that the local square integrability condition could be relaxed, which led to the traditional definition of a semimartingale as a sum of a local martingale and an FV process. A different theory of stochastic integration, the *Itô-belated integral*, was developed by McShane [5]. It imposed different restrictions on the integrators and the integrands and used a theory of “gauges” and appeared to be very different from the approach here. It turns out, however, that when the integral $\int H dX$ made sense both as a stochastic integral in the sense developed here and as an Itô-belated integral, they were indistinguishable. See [6] for a comparison of these two integrals. Another related stochastic integral is called the *Fisk–Stratonovich (FS)* integral that was developed by Fisk and Stratonovich independently. The FS integral obeys the integration by parts formula for FV

processes when at least one of the integrand or the integrator is continuous.

End Notes

^aSee Definition 5 for the definition of a predictable process.

^bFor a proof of the fact that $\mathbb{D}_{\text{UCP}}$ is metrizable and complete under that metric, see [7].

^cThis sum converges to the *quadratic variation* of B over the interval $[0, t]$ as we see in Theorem 16.

References

- [1] Chou, C.S., Meyer, P.A. & Stricker, C. (1980). Sur les intégrales stochastiques de processus prévisibles non bornés, *Séminaire de Probabilités, XIV*. Lecture Notes in Mathematics, 784, Springer, Berlin, pp. 128–139.
- [2] Doléans-Dade, C. & Meyer, P.-A. (1970). Intégrales stochastiques par rapport aux martingales locales, *Séminaire de Probabilités, IV*. Lecture Notes in Mathematics, 124, Springer, Berlin, pp. 77–107.
- [3] Itô, K. (1944). Stochastic integral, *Proceedings of the Imperial Academy of Tokyo* **20**, 519–524.
- [4] Jeulin, T. (1980). *Semi-martingales et Grossissement d'une Filtration*, Lecture Notes in Mathematics, Springer, Berlin, Vol. 833.
- [5] McShane, E.J. (1974). *Stochastic Calculus and Stochastic Models*, Probability and Mathematical Statistics, Academic Press, New York, Vol. 25.
- [6] Protter, P. (1979). A comparison of stochastic integrals, *The Annals of Probability* **7**(2), 276–289.
- [7] Protter, P. (2005). *Stochastic Integration and Differential Equations*, 2nd Edition, Version 2.1, Springer, Berlin.

Related Articles

Arbitrage Strategy; Complete Markets; Equivalent Martingale Measures; Filtrations; Itô's Formula; Martingale Representation Theorem; Semimartingale.

UMUT ÇETİN

Equivalence of Probability Measures

In finance it is often important to consider different probability measures. The statistical measure, commonly denoted by P , is supposed to (ideally) reflect the real-world dynamics of financial assets. A risk-neutral measure (see **Equivalent Martingale Measures**), often denoted by Q , is the measure of choice for the valuation of derivative securities. Prices of traded assets are supposed to be (local) Q -martingales, and hence their dynamics (as seen under Q) typically differs from their actual behavior (as modeled under P). How far can the dynamics with respect to these two measures be away in terms of qualitative behavior? We would not expect that events that do not occur in the real world, in the sense that they have P -probability zero, like a stock price exploding to infinity, would have positive Q -probability in the risk-neutral world. This discussion leads to the notion of absolute continuity.

Definition 1 Let P, Q be two probability measures defined on a measurable space $(\Omega, \mathcal{F})$. We say that Q is absolutely continuous with respect to P , denoted by $Q \ll P$, if all P -zero sets are also Q -zero sets. If $Q \ll P$ and $P \ll Q$ we say that P and Q are equivalent, denoted by $P \sim Q$. In other words, two equivalent measures have the same zero sets.

Let $Q \ll P$. By the Radon–Nikodym theorem there exists a density $Z = dQ/dP$ so that for $f \in L^1(Q)$ we can calculate its expectation with respect to Q by

$$E_Q[f] = E_P[Zf] \quad (1)$$

Note that if Q is absolutely continuous, but not equivalent to P , then we have $P(Z = 0) > 0$.

We now look at a dynamic picture, and assume that we also have a filtration $(\mathcal{F}_t)_{0 \leq t \leq T}$ at our disposal where T is some fixed finite time horizon. For $t \leq T$ let

$$Z_t = E_P[Z | \mathcal{F}_t] \quad (2)$$

We call the martingale $Z = (Z_t)$ the *density process* of Q . The *Bayes formula* tells us how to calculate conditional expectations with respect to Q in terms of P . Let $0 \leq s \leq t \leq T$ and f be $\mathcal{F}_t$ -measurable and

in $L^1(Q)$. We then have

$$Z_s E_Q[f | \mathcal{F}_s] = E_P[Z_t f | \mathcal{F}_s] \quad (3)$$

As consequence of Bayes' formula, we get that if M is a Q -martingale then ZM is a P -martingale and *vice versa*. Hence, we can turn any Q -martingale into a P -martingale by just multiplying it with the density process. It follows that the martingale property is *not* invariant under equivalent measure changes.

There are, however, a couple of important objects like stochastic integrals and quadratic variations which do remain invariant under equivalent measure changes although they depend, by their definition, *a priori* on some probability measure. Let us illustrate this in case of the quadratic variation of a semimartingale S . This is defined to be the limit in P -probability of the sum of the squared S -increments over a time grid, for vanishing mesh size. It is elementary that convergence in P -probability implies convergence in Q -probability if $Q \ll P$, and thus convergence in P -probability is equivalent to the convergence in Q -probability when P and Q are equivalent. This implies, for example, that quadratic variations remain the same under a change to an equivalent probability measure.

The compensator or angle bracket process, however, is *not* invariant with respect to equivalent measure changes. It is defined (for reasonable processes S) as the process $\langle S \rangle$ one has to subtract from the quadratic variation process $[S]$ to turn it into a local martingale. But, as we have seen, the martingale property typically gets lost by switching the measure. As an example, consider a Poisson process N with intensity λ . We have $[N] = N$, so the compensator equals λt . As we shall see below, the effect of an equivalent measure change is that the intensity changes as well, to μ , say, so the compensator under the new measure would be μt .

Girsanov's Theorem

As we have discussed above, the martingale property is not preserved under measure changes. Fortunately, it turns out that at least the semimartingale property is preserved. Moreover, it is possible to state the precise semimartingale decomposition under the new measure Q . This result is known in the literature as the *Girsanov's theorem*, although it was rather Cameron and Martin who proved a first version of

it in a Wiener space setting. Later on it was extended in various levels of generality by Girsanov, Meyer, and Lenglart, among many others.

Let us first give some examples. They are all the consequences of the general formulation of Girsanov's theorem to be given below.

1. Let B be a P -Brownian motion, $\mu \in \mathbb{R}$, and define an equivalent measure Q by the stochastic exponential

$$\frac{dQ}{dP} = \mathcal{E}(-\mu B)_T = \exp\left(-\mu B_T - \frac{1}{2}\mu^2 T\right) \quad (4)$$

Then $\widehat{B} = B + \mu t$ is a Q -Brownian motion (up to time T). Alternatively stated, the semimartingale decomposition of B under Q is $B = \widehat{B} - \mu t$. Hence the effect of the measure change is to add a drift term to the Brownian motion.

2. Let $N_t - \lambda t$ be a compensated Poisson process on an interval $[0, T]$ with P -intensity $\lambda > 0$, and let $\kappa > 0$. Define an equivalent measure Q by

$$\begin{aligned} \frac{dQ}{dP} &= e^{-\kappa\lambda T} \prod_{0 < s \leq T} (1 + \kappa \Delta N_s) \\ &= e^{-\kappa\lambda T} (1 + \kappa)^{N_T} \\ &= \exp(N_T \ln(1 + \kappa) - \kappa\lambda T) \end{aligned} \quad (5)$$

Then N is a Poisson process on $[0, T]$ under Q with intensity $(1 + \kappa)\lambda$. The process $N_t - (1 + \kappa)\lambda t$ is a compensated Poisson process under Q and thus a Q -martingale. Hence the effect of the measure change is to change the intensity of the Poisson process, or in other words, to add a drift term to the compensated Poisson process.

One of the most important applications of measure changes in mathematical finance is to find martingale measures for the price process S of some risky asset.

Definition 2 A martingale measure for S is a probability measure Q such that S is a Q -local martingale.

Let us now state a general form of Girsanov's theorem. It is not the most general setting, though, since we will assume that Q is equivalent to P which suffices for most applications in finance. This is due to the fact that one would often choose Q

to be a martingale measure for the price process, and then equivalence is a necessary condition to exclude arbitrage opportunities [1]. There is, however, also a result which covers the case where Q is only absolutely continuous, but not equivalent to P , and which has been proven by Lenglart [2].

Theorem 1 (Girsanov's Theorem: Standard Version). Let $P \sim Q$, with density process given by

$$Z_t = E\left[\frac{dQ}{dP} \middle| \mathcal{F}_t\right] \quad (6)$$

If S is a semimartingale under P with decomposition $S = M + A$ (here M is a local martingale, and A a process of locally finite variation), then S is a semimartingale under Q as well and has decomposition

$$\begin{aligned} S &= \left(M - \int \frac{1}{Z} d[Z, M]\right) \\ &\quad + \left(A + \int \frac{1}{Z} d[Z, M]\right) \end{aligned} \quad (7)$$

In particular, $M - \int \frac{1}{Z} d[Z, M]$ is a local Q -martingale.

In situations where the process S may exhibit jumps, it is often more convenient to apply a version of Girsanov which uses the angle bracket instead of the quadratic covariation.

Theorem 2 (Girsanov's Theorem: Predictable Version). Let $P \sim Q$, with density process as above, and $S = M + A$ be a P -semimartingale. Given that $\langle Z, M \rangle$ exists (with respect to P), then the decomposition of S under Q is

$$\begin{aligned} S &= \left(M - \int \frac{1}{Z_-} d\langle Z, M \rangle\right) \\ &\quad + \left(A + \int \frac{1}{Z_-} d\langle Z, M \rangle\right) \end{aligned} \quad (8)$$

Here Z_- denotes the left-continuous version of Z .

Whereas the standard version of Girsanov's theorem always works, we need an integrability condition (existence of $\langle Z, M \rangle$) for the predictable version.

However, in case $S = M + A$ for a local martingale M and a finite variation process A , it is rarely the case in a discontinuous framework that $dA \ll d[M]$, whereas it is quite natural in financial applications that $dA \ll d\langle M \rangle$ (see below).

In mathematical finance, these results are often applied to find a martingale measure for the price process S . Consider, for example, the Bachelier model where $S = B + \mu t$ is a Brownian motion plus drift. If we now take as above the measure change as given by a density process $Z_t = \exp\left(-\mu B_t - \frac{1}{2}\mu^2 t\right)$, then we have (since $dZ = -\mu Z dB$)

$$\begin{aligned} A + \int \frac{1}{Z} d[Z, M] &= \mu t + \int \frac{1}{Z} d\left[-\mu \int Z dB, B\right] \\ &= \mu t + \int \frac{1}{Z} d\left(-\mu \int Z dt\right) \\ &= 0 \end{aligned} \quad (9)$$

According to Girsanov's theorem (here the standard version coincides with the predictable one since S is continuous), the price process S is therefore a Q -local martingale (and, in fact, a Brownian motion according to Lévy's characterization), and hence Q is a martingale measure for S .

More generally, Girsanov's theorem implies an important structural result for the price process S in an arbitrage-free market. As has been mentioned above, it is essentially true that some no-arbitrage property implies the existence of an equivalent martingale measure Q for $S = M + A$, with density process Z . Therefore, we must have by the predictable version (8), given that $\langle Z, M \rangle$ exists, that

$$A = - \int \frac{1}{Z_-} d\langle Z, M \rangle \quad (10)$$

to get that S is a local Q -martingale. As it follows from the so-called Kunita-Watanabe inequality that

$$d\langle Z, M \rangle \ll d\langle M \rangle \quad (11)$$

(here $\langle Z, M \rangle$ respectively $\langle M \rangle$ are interpreted as the associated measures on the nonnegative real line), we conclude that

$$dA \ll d\langle M \rangle \quad (12)$$

and hence there exists some predictable process λ such that

$$S = M + \int \lambda d\langle M \rangle \quad (13)$$

For example, in the Bachelier model $S = B + \mu t$ we have that $\langle B \rangle_t = t$, and hence λ equals the constant μ .

The predictable version of Girsanov's theorem can now be applied to remove the drift $\int \lambda d\langle M \rangle$ as follows: we define a probability measure Q via

$$\frac{dQ}{dP} = \mathcal{E}\left(-\int \lambda dM\right)_T \quad (14)$$

where $\mathcal{E}$ denotes the Doléans-Dade stochastic exponential, assuming that $\mathcal{E}\left(-\int \lambda dM\right)$ is a martingale. The corresponding density process Z therefore satisfies the stochastic differential equation

$$dZ = -Z_- \lambda dM \quad (15)$$

It follows that

$$\langle Z, M \rangle = \left\langle -\int Z_- \lambda dM, M \right\rangle = -\int Z_- \lambda d\langle M \rangle \quad (16)$$

and

$$S = M + \int \lambda d\langle M \rangle = M - \int \frac{1}{Z_-} d\langle Z, M \rangle \quad (17)$$

is by the (predictable version) of the Girsanov theorem a local Q -martingale: the drift has been removed by the measure change.

This representation of S has an important consequence for the structure of martingale measures, provided the so-called *structure condition* holds:

$$\int_0^T \lambda_s^2 d\langle M \rangle_s < \infty \quad P\text{-a.s.} \quad (18)$$

In that case, the remarkable conclusion we can draw from (13) is that the existence of an equivalent martingale measure for S implies that S is a *special semimartingale*, for example, its finite variation part is predictable and therefore the semimartingale decomposition (13) is unique. Moreover, the following result holds.

Proposition 1 *Let Q be an equivalent martingale measure for S , and the structure condition (18) hold. Then the density process Z of Q with respect to P is given by the stochastic exponential*

$$Z = \mathcal{E}\left(-\int \lambda dM + L\right) \quad (19)$$

for some process L such that L as well as $[M, L]$ are local P -martingales. The converse statement is true as well, assuming that all involved processes are locally bounded: if Q is a probability measure whose density process can be written like in equation (19) with L as above, then Q is a martingale measure for S .

This result is fundamental in incomplete markets (see **Complete Markets**), where there are many equivalent martingale measures for the price process S . Indeed, any choice of L as in the statement of the proposition gives one particular pricing measure.

In applications in finance, the density process Z can also be interpreted in terms of a **change of numeraire**.

References

- [1] Delbaen, F. & Schachermayer, W. (2006). *The Mathematics of Arbitrage*, Springer, Berlin.
- [2] Protter, P.E. (2005). *Stochastic Integration and Differential Equations*, 2nd Edition, Version 2.1, Springer, Heidelberg.

Related Articles

Change of Numeraire; Equivalent Martingale Measures; Semimartingale; Stochastic Exponential; Stochastic Integrals.

THORSTEN RHEINLÄNDER

Skorokhod Embedding

Analysis of a random evolution focuses initially on the behavior at a fixed deterministic, or random, time. The process and time horizon are known and we investigate the marginal law of the process. If we reverse this point of view, we face the embedding problem. We fix a probability distribution and a (well-understood) stochastic process and we try to design a random time such that the process at this time behaves according to the specified distribution. In other words, we know *what* we want to see and we ask *when* to look for it.

This *Skorokhod embedding problem* (SEP) or the *Skorokhod stopping problem*, first formulated and solved by A.V. Skorokhod in 1961 (English translation in 1965 [20]), is thus the problem of representing a given distribution μ as the distribution of a given stochastic process (such as a Brownian motion) at some stopping time. It has stimulated research in probability theory for over 40 years now—the problem has been changed, generalized, or specialized in various ways. We discuss some key results in the domain, along with the applications in quantitative finance, namely to the computation of robust market-consistent prices and hedges of exotic derivatives.

The Skorokhod Embedding Problem

The SEP problem can be stated as follows:

Given a stochastic process $(X_t : t \geq 0)$ and a probability measure μ , find a minimal stopping time τ such that X_τ has the law $\mu : X_\tau \sim \mu$.

At first, there seems to be a trivial solution to the SEP when $X_t = B_t$ is a Brownian motion. Write Φ and F_μ for the cumulative distribution function of the standard normal distribution and of μ , respectively. Then $F_\mu^{-1}(\Phi(B_1))$ has law μ and hence the stopping time $\tau = \inf\{t \geq 2 : B_t = F_\mu^{-1}(\Phi(B_1))\}$ satisfies $B_\tau \sim \mu$. However, this solution is intuitively “too large”, in particular $\mathbb{E}\tau = \infty$. A meaningful solution needs to be “small”. To express this, Skorokhod [20] imposed $\mathbb{E}\tau < \infty$ and solved the problem explicitly for any centered target measure with finite variance. To avoid the restriction on the set of target measures, in general, one requires τ to be *minimal*. Minimality of τ signifies that if a stopping time ρ satisfies $\rho \leq \tau$ and $X_\rho \sim X_\tau$ then $\rho = \tau$. When $\mathbb{E}B_\tau = 0$,

minimality of τ is equivalent to $(B_{t \wedge \tau} : t \geq 0)$ being a uniformly integrable martingale (see [6, 12]) and, in consequence, when $\mathbb{E}B_\tau^2 < \infty$, it is further equivalent to $\mathbb{E}\tau < \infty$. Note that we can have many, in fact, infinitely many, minimal stopping times all of which embed the same distribution μ .

We want τ to be “small” to enable us to iterate the embedding procedure. In this way, Skorokhod [20] represented a random walk as a Brownian motion stopped at an increasing sequence of stopping times and deduced properties of the random walk from the well-understood behavior of Brownian motion. As a simple example, one can use the representation to deduce the central limit theorem from the strong law of large numbers (cf. [14, Sec. 11.2]). The ideas of embedding processes into Brownian motion were extended and finally led to the celebrated work of Monroe [13], who proved that any semimartingale is a time-changed Brownian motion.

The SEP, as stated above, does not necessarily have a solution—existence of a solution depends greatly on X and μ . This can be seen already for real-valued diffusions [6]. However, for Brownian motion on $\mathbb{R}$, or any continuous local martingale (X_t) with $\langle X \rangle_\infty = \infty$ a.s., there is always a solution to the SEP and there are numerous explicit constructions (typically, for the case of centered μ), of which we give two examples below (cf. [14]).

Explicit Solutions

Skorokhod [20] and Dubins [8] solved the SEP for Brownian motion and arbitrary centered^a probability measure μ . However, the search for new solutions continued and was, to a large extent, motivated by the properties of the stopping times. Researchers sought simple explicit solutions that would have additional optimal properties. Several solutions were obtained using stopping times of the form

$$\tau = \inf\{t : (A_t, B_t) \in \Gamma\}, \quad \Gamma = \Gamma(\mu) \subset \mathbb{R}^2 \quad (1)$$

which is a first hitting time for the Markov process (A_t, B_t) , where (A_t) is some auxiliary increasing process. We now give two examples.

Consider $A_t = t$ and let τ_R be the resulting stopping time in (1). Root [17] proved that for any centered μ there is a *barrier* $\Gamma = \Gamma(\mu)$ such that $B_\tau \sim \mu$, where a *barrier* is a set in $\mathbb{R}_+ \times \mathbb{R}$ (time–space) such that if a point is in Γ , then all points to the right of it are also in Γ (see Figure 1).

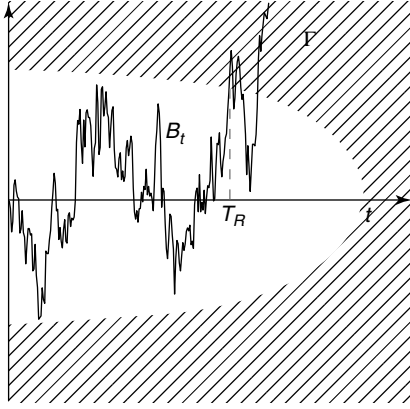


Figure 1 The barrier Γ and Root stopping time τ_R embedding a uniform law

Later Rost (cf. [14]) proved an analogous result replacing $\Gamma(\mu)$ with a reversed barrier $\tilde{\Gamma} = \tilde{\Gamma}(\mu)$, which is a set in time–space such that if a point is in $\tilde{\Gamma}$, then all the points to the left of it are also in $\tilde{\Gamma}$. We denote $\tilde{\tau}_R$ the first hitting of $\tilde{\Gamma}(\mu)$. Rost (cf. [14, 19]) proved that for any other solution τ to the SEP and any positive convex function f , we have

$$\mathbb{E}f(\tau_R) \leq \mathbb{E}f(\tau) \leq \mathbb{E}f(\tilde{\tau}_R) \quad (2)$$

In financial terms, as we will see, this implies bounds on the prices of volatility derivatives. Given a measure μ , the barrier Γ and the reversed barrier $\tilde{\Gamma}$ are not known explicitly. However, using techniques of partial differential equations, they can be computed numerically together with the bounds in equation (2) (see [9]).

Consider now $A_t = \bar{B}_t = \sup_{u \leq t} B_u$ in equation (1). Azéma and Yor [1] proved that, for a probability measure μ satisfying $\int x \mu(dx) = B_0$, the stopping time

$$\tau_{AY} = \inf\{t : \Psi_\mu(B_t) \leq \bar{B}_t\},$$

where
$$\Psi_\mu(x) = \frac{1}{\mu([x, \infty))} \int_{[x, \infty)} u \mu(du) \quad (3)$$

is minimal and $B_{\tau_{AY}} \sim \mu$. The Azéma–Yor stopping time is also optimal as it stochastically maximizes the maximum: $\mathbb{P}(\bar{B}_\tau \geq \alpha) \leq \mathbb{P}(\bar{B}_{\tau_{AY}} \geq \alpha)$, for all $\alpha \geq 0$ and any minimal τ with $B_\tau \sim \mu$. Later, Perkins [16] developed a stopping time τ_p , which, in turn, stochastically minimizes the maximum. As we will

see, these two solutions induce upper and lower bounds on the price of a one-touch option.

Applications

Robust Price Bounds

In the standard approach to pricing and hedging, one postulates a model for the underlying, calibrates it to the market prices of liquidly traded vanilla options (see **Call Options**), and then uses the model to derive prices and associated hedges for exotic over-the-counter products (such as **Barrier Options**; **Look-back Options**; **Foreign Exchange Options**). Prices and hedges will be correct only if the model describes the real world perfectly, which is not very likely. The SEP-driven approach uses the market data to deduce bounds on the prices consistent with *no-arbitrage* and the associated super-replicating strategies (see **Superhedging**), which are robust to model misspecification.

Assume *absence of arbitrage* (see **Fundamental Theorem of Asset Pricing**) and work under a risk-neutral measure (see **Risk-neutral Pricing**) so that the forward price process (see **Forwards and Futures**) $(S_t : t \leq T)$ is a martingale. Equivalently, under a simplifying assumption of zero interest rates, S_t is simply the stock price process. We are interested in pricing an exotic option with payoff given by a path-dependent functional $F(S)_T$. Our main example considered below is a *one-touch* option struck at α that pays 1 if the stock price reaches α before maturity T : $O^\alpha(S)_T = \mathbf{1}_{\bar{S}_T \geq \alpha}$, where $\bar{S}_T = \sup_{t \leq T} S_t$. It follows from Monroe's theorem that $S_t = B_{\rho_t}$, for a Brownian motion (B_t) with $B_0 = S_0$ and some increasing sequence of stopping times $\rho_t : t \leq T$ (possibly relative to an enlarged filtration). We make no other assumptions about the dynamics of the underlying. Instead, we propose to investigate the restrictions induced by the market data.

Suppose, first, that we know the market prices of calls and puts (see **Call Options**) for all strikes at one maturity T . This is equivalent to knowing the distribution μ of S_T (cf. [3]). Thus, we can see the stopping time $\rho = \rho_T$ as a solution to the SEP for μ . Conversely, given a solution τ to the SEP for μ , the process $\tilde{S}_t = B_{\tau \wedge \frac{t}{T-\tau}}$ is a model for the stock-price process consistent with the observed prices of calls and puts at maturity T . In this way, we obtain

a correspondence that allows us to identify market models with solutions to the SEP and *vice versa*. In consequence, to estimate the fair price of the exotic option $\mathbb{E}F(S)_T$, it suffices to bound $\mathbb{E}F(B)_\tau$ among all solutions τ to the SEP. More precisely, if $F(S)_T = F(B)_{\rho_T}$ a.s., then we have

$$\inf_{\tau: B_\tau \sim \mu} \mathbb{E}F(B)_\tau \leq \mathbb{E}F(S)_T \leq \sup_{\tau: B_\tau \sim \mu} \mathbb{E}F(B)_\tau \quad (4)$$

where all stopping times τ are minimal. Consider, for example, a volatility derivative^b paying $F(S)_T = f(\langle S \rangle_T)$, for some positive convex function f , and suppose that the underlying (S_t) is continuous. Then, by Dubins–Schwarz theorem, we can take the time change $\rho_t = \langle S \rangle_t$ so that $f(\langle S \rangle_T) = f(\rho_T) = F(B)_{\rho_T}$. Using inequality (2), inequality (4) becomes

$$\mathbb{E}f(\tau_R) \leq \mathbb{E}f(\langle S \rangle_T) \leq \mathbb{E}f(\tilde{\tau}_R) \quad (5)$$

where $B_{\tau_R} \sim S_T \sim B_{\tilde{\tau}_R}$ (cf. [9]).

When (S_t) has jumps typically one of the bounds in inequality (4) remains true and the other degenerates. In the example of a one-touch option, one sees that $O^\alpha(S)_T \leq O^\alpha(B)_{\rho_T}$ and the fair price is always bounded above by $\sup_\tau \{\mathbb{P}(\bar{B}_\tau \geq \alpha) : B_\tau \sim \mu\}$. Furthermore, the supremum is attained by the Azéma–Yor construction discussed above. The best lower bound on the price in the presence of jumps is the obvious bound $\mu([\alpha, \infty))$. In consequence, the price of a one-touch option $\mathbb{E}O^\alpha(S)_T = \mathbb{P}(\bar{S}_T \geq \alpha)$ is bounded by

$$\begin{aligned} \mu([\alpha, \infty)) &\leq \mathbb{P}(\bar{S}_T \geq \alpha) \leq \mathbb{P}(\bar{B}_{\tau_{AY}} \geq \alpha) \\ &= \mu([\Psi_\mu^{-1}(\alpha))) \end{aligned} \quad (6)$$

and the lower bound can be improved to $\mathbb{P}(\bar{B}_{\tau_P} \geq \alpha)$ under the hypothesis that (S_t) is continuous, where τ_P is Perkins' stopping time (see [5] for detailed discussion and numerical examples). Selling a one-touch option for a lower price than the upper bound in equation (6) necessarily involves some risk. If additional modeling assumptions are made, then a lower price can be justified, but this new price is not necessarily robust to model misspecification.

The above analysis can be extended if we know more market data. For example, knowing prices of puts and calls at some earlier expiry $T_1 < T$ would lead to solving the SEP, constrained by embedding an intermediate law μ_1 before μ . This was achieved by Brown *et al.* [4] who gave an explicit construction

of an embedding that maximizes the maximum. As we have seen, in financial terms, this amounts to obtaining the least upper bound on the price of a one-touch option.

In practice, we do not observe the prices of calls and puts for all strikes but only for a finite family of strikes. As a result, the terminal law of S_T is not specified entirely and one needs to optimize among possible terminal laws (cf. [5, 10]). In general, different sets of market prices lead to embedding problems with different constraints. The resulting problems can be complex. In particular, to our best knowledge, there are no known optimal solutions to the SEP with multiple intermediate law constraints.

Robust Hedging

Once we know the price-range for an option, we want to understand model-free super-replicating strategies (see **Superhedging**). In general, to achieve this, we need to develop a pathwise approach to the SEP. Following [5], we treat the example of a one-touch option. We develop a super-replicating portfolio with the initial wealth equal to the upper bound displayed in equation (6).

The key observation lies in the following simple inequality:

$$\mathbf{1}_{\bar{S}_T \geq \alpha} \leq \frac{(S_T - K)^+}{\alpha - K} + \frac{S_{\bar{S} \wedge T} - S_T}{\alpha - K} \mathbf{1}_{\bar{S}_T \geq \alpha} \quad (7)$$

where $\alpha > S_0, K$ and $\bar{S} = \inf\{t : S_t \geq \alpha\}$. Taking expectations yields $\mathbb{P}(\bar{S}_T \geq \alpha) \leq C(K)/(\alpha - K)$, where $C(K)$ denotes the price of a European call with strike K and maturity T . Taking the optimal $K = K^*$ such that $C(K^*) = (\alpha - K^*)|C'(K^*)|$ we find $\mathbb{P}(\bar{S}_T \geq \alpha) \leq |C'(K^*)| = \mathbb{P}(S_T \geq K^*)$. On the other hand, using $|C'(K)| = \mu([K, \infty))$, where $\mu \sim S_T$, we have

$$C(K) = \int_K^\infty (u - K)\mu(du) = |C'(K)|(\Psi_\mu(K) - K) \quad (8)$$

The equation for K^* implies readily that $K^* = \Psi_\mu^{-1}(\alpha)$ and the bound we have derived coincides with equation (6).

Inequality (7) encodes the super-replicating strategy. The first term of the right-hand side means we buy $1/(\alpha - K^*)$ calls with strike K^* . The second

term is a simple dynamic trading: if the price reaches level α , we sell $1/(\alpha - K^*)$ forwards on the stock. At the cost of $C_1 = C(K^*)/(\alpha - K^*)$ we are then guaranteed to super-replicate the one-touch regardless of the dynamics of the underlying. In consequence, selling the one-touch for $C_2 > C_1$ would be an arbitrage opportunity as we would make a riskless profit of $C_2 - C_1$. Finally, note that our derivation of the superhedge is pathwise and makes no assumptions about the existence (or uniqueness) of the pricing measure.

Other Resources

The arguments for robust pricing and hedging of lookback (see **Lookback Options**) and barrier (see **Barrier Options**) options can be found in the pioneering work of Hobson [10] and in [5]. Dupire [9] investigated volatility derivatives using the SEP. Cox *et al.* [7] designed pathwise inequalities to derive price range and robust super-replicating strategies for derivatives paying a convex function of the local time (see **Local Times; Corridor Variance Swap**). The idea of *no-arbitrage* bounds on the prices goes back to Merton [11] (see **Arbitrage Bounds**). This was refined in *no-good deals* (see **Good-deal Bounds**) pricing, where one postulates that markets not only exclude arbitrage opportunities but also any *highly desirable investments*. *No-good deals* pricing yields tighter bounds on the prices but requires an arbitrary choice of utility function.

We refer to [14] for an extended survey of the SEP, including its history and overview of its applications. We have not discussed here the SEP for processes other than Brownian motion. Rost [18] investigated the problem for a general Markov process and has a necessary and sufficient condition on the target measure μ for existence of an embedding. Bertoin and Le Jan [2] then developed an explicit solution, in a broad class of Markov processes, which was based on additive functionals. More recently, the approach of Vallois [21] was extended to provide explicit solutions for classes of discontinuous processes including Azéma's martingale [15].

Acknowledgments

This research was supported by a Marie Curie Intra-European Fellowship at Imperial College London within the 6th European Community Framework Programme.

End Notes

^aWhen modeling the stock price process, implicitly we shift both B and μ by a constant S_0 .

^bHere, written on the realized quadratic variation of the stock itself and not the log process.

References

- [1] Azéma, J. & Yor, M. (1979). Une solution simple au problème de Skorokhod, in *Séminaire de Probabilités, XIII*, Lecture Notes in Mathematics, Springer, Berlin, Vol. 721, pp. 90–115.
- [2] Bertoin, J. & Le Jan, Y. (1992). Representation of measures by balayage from a regular recurrent point, *Annals of Probability* **20**(1), 538–548.
- [3] Breeden, D.T. & Litzenberger, R.H. (1978). Prices of state-contingent claims implicit in option prices, *The Journal of Business* **51**(4), 621–651.
- [4] Brown, H., Hobson, D. & Rogers, L.C.G. (2001). The maximum maximum of a martingale constrained by an intermediate law, *Probability Theory and Related Fields* **119**(4), 558–578.
- [5] Brown, H., Hobson, D. & Rogers, L.C.G. (2001). Robust hedging of barrier options, *Mathematical Finance* **11**(3), 285–314.
- [6] Cox, A. & Hobson, D. (2006). Skorokhod embeddings, minimality and non-centered target distributions, *Probability Theory and Related Fields* **135**(3), 395–414.
- [7] Cox, A., Hobson, D. & Oblój, J. (2008). Pathwise inequalities for local time: applications to Skorokhod embeddings and optimal stopping, *Annals of Applied Probability* **18**(5), 1870–1896.
- [8] Dubins, L.E. (1968). On a theorem of Skorokhod, *The Annals of Mathematical Statistics* **39**, 2094–2097.
- [9] Dupire, B. (2005). *Arbitrage Bounds for Volatility Derivatives as a Free Boundary Problem*, http://www.math.kth.se/pde_finance/presentations/Bruno.pdf.
- [10] Hobson, D. (1998). Robust hedging of the lookback option, *Finance and Stochastics* **2**, 329–347.
- [11] Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.
- [12] Monroe, I. (1972). On embedding right continuous martingales in Brownian motion, *The Annals of Mathematical Statistics* **43**, 1293–1311.
- [13] Monroe, I. (1978). Processes that can be embedded in Brownian motion, *The Annals of Probability* **6**(1), 42–56.
- [14] Oblój, J. (2004). The Skorokhod embedding problem and its offspring, *Probability Surveys* **1**, 321–392.
- [15] Oblój, J. (2007). An explicit solution to the Skorokhod embedding problem for functionals of excursions of Markov processes, *Stochastic Process and their Application* **117**(4), 409–431.

-
- [16] Perkins, E. (1986). The Cereteli-Davis solution to the H^1 -embedding problem and an optimal embedding in Brownian motion, in *Seminar on stochastic processes, 1985 (Gainesville, Fla., 1985)*, Progress in Probability and Statistics, Birkhäuser Boston, Boston, Vol. 12, pp. 172–223.
 - [17] Root, D.H. (1969). The existence of certain stopping times on Brownian motion, *The Annals of Mathematical Statistics* **40**, 715–718.
 - [18] Rost, H. (1971). The stopping distributions of a Markov Process, *Inventiones Mathematicae* **14**, 1–16.
 - [19] Rost, H. (1976). Skorokhod stopping times of minimal variance, in *Séminaire de Probabilités, X*, Lecture Notes in Mathematics, Springer, Berlin, Vol. 511, pp. 194–208.
 - [20] Skorokhod, A.V. (1965). *Studies in the Theory of Random Processes*, Addison-Wesley Publishing Co., Reading, Translated from the Russian by Scripta Technica, Inc.
 - [21] Vallois, P. (1983). Le problème de Skorokhod sur $\mathbf{R}$: une approche avec le temps local, in *Séminaire de Probabilités, XVII*, Lecture Notes in Mathematics, Springer, Berlin, Vol. 986, pp. 227–239.

Related Articles

Arbitrage Bounds; Arbitrage: Historical Perspectives; Arbitrage Pricing Theory; Arbitrage Strategy; Barrier Options; Complete Markets; Convex Risk Measures; Good-deal Bounds; Hedging; Implied Volatility Surface; Martingales; Model Calibration; Static Hedging; Superhedging.

JAN OBLÓJ

Markov Processes

A *Markov process* is a process that evolves in a memoryless way: its future law depends on the past only through the present position of the process. This property can be formalized in terms of conditional expectations: a process $(X_t, t \geq 0)$ adapted to the **filtration** $(\mathcal{F}_t)_{t \geq 0}$ (representing the information available at time t) is a Markov process if

$$\mathbb{E}(f(X_{t+s}) \mid \mathcal{F}_t) = \mathbb{E}(f(X_{t+s}) \mid X_t) \quad (1)$$

for all $s, t \geq 0$ and f bounded and measurable.

The interest of such a process in financial models becomes clear when one observes that the price of an option, or more generally, the value at time t of any future claim with maturity T , is given by the general formula (see **Risk-neutral Pricing**)

$$\begin{aligned} V_t &= \text{value at time } t \\ &= \mathbb{E}(\text{discounted payoff at time } T \mid \mathcal{F}_t) \end{aligned} \quad (2)$$

where the expectation is computed with respect to a *pricing measure* (see **Equivalent Martingale Measures**). The Markov property is a frequent assumption in financial models because it provides powerful tools (semigroup, theory of partial differential equations (PDEs), etc.) for the quantitative analysis of such problems.

Assuming the Markov property (1) for $(S_t, t \geq 0)$, the value V_t of the option can be expressed as

$$\begin{aligned} V_t &= \mathbb{E}(e^{-r(T-t)} f(S_T) \mid \mathcal{F}_t) \\ &= \mathbb{E}(e^{-r(T-t)} f(S_T) \mid S_t) \end{aligned} \quad (3)$$

so V_t can be expressed as a (deterministic) function of t, S_t : $u(t, S_t) = \mathbb{E}(e^{-r(T-t)} f(S_T) \mid S_t)$. Furthermore, this function u is shown to be the solution of a parabolic PDE, the Kolmogorov backward equation.

The goal of this article is to present the Markov processes and their relation with PDEs, and to illustrate the role of Markovian models in various financial problems. We give a general overview of the links between Markov processes and PDEs without giving more details and we focus on the case of Markov processes solution to stochastic differential equations (SDEs).

We will restrict ourselves to $\mathbb{R}^d$ -valued Markov processes. The set of Borel subsets of $\mathbb{R}^d$ is denoted

by $\mathcal{B}$. In the following, we will denote a Markov process by $(X_t, t \geq 0)$, or simply X when no confusion is possible.

Markov Property and Transition Semigroup

A Markov process retains no memory of where it has been in the past. Only the current state of the process influences its future dynamics. The following definition formalizes this notion:

Definition 1 *Let $(X_t, t \geq 0)$ be a stochastic process defined on a probability filtered space $(\Omega, \mathcal{F}_t, \mathbb{P})$ with values in $\mathbb{R}^d$. X is a Markov process if*

$$\mathbb{P}(X_{t+s} \in \Gamma \mid \mathcal{F}_t) = \mathbb{P}(X_{t+s} \in \Gamma \mid X_t) \quad \mathbb{P}\text{-a.s.} \quad (4)$$

for all $s, t \geq 0$ and $\Gamma \in \mathcal{B}$. Equation (4) is called the *Markov property* of the process X . The Markov process is called *time homogeneous* if the law of X_{t+s} conditionally on $X_t = x$ is independent of t .

Observe that equation (4) is equivalent to equation (1) and that X is a time-homogeneous Markov process if there exists a positive function P defined on $\mathbb{R}_+ \times \mathbb{R}^d \times \mathcal{B}$ such that

$$P(s, X_t, \Gamma) = \mathbb{P}(X_{t+s} \in \Gamma \mid \mathcal{F}_t) \quad (5)$$

holds $\mathbb{P}$ -a.s. for all $t, s \geq 0$ and $\Gamma \in \mathcal{B}$. P is called the *transition function* of the time homogeneous Markov process X .

For the moment, we restrict ourselves to the time-homogeneous case.

Proposition 1 *The transition function P of a time-homogeneous Markov process X satisfies*

1. $P(t, x, \cdot)$ is a probability measure on $\mathbb{R}^d$ for any $t \geq 0$ and $x \in \mathbb{R}^d$,
2. $P(0, x, \cdot) = \delta_x$ (unit mass at x) for any $x \in \mathbb{R}^d$,
3. $P(\cdot, \cdot, \Gamma)$ is measurable for any $\Gamma \in \mathcal{B}$, and for any $s, t \geq 0, x \in \mathbb{R}^d, \Gamma \in \mathcal{B}$, P satisfies the Chapman–Kolmogorov property

$$P(t+s, x, \Gamma) = \int_{\mathbb{R}^d} P(s, y, \Gamma) P(t, x, dy) \quad (6)$$

From an analytical viewpoint, we can think of the transition function as a Markov semigroup^a $(P_t, t \geq 0)$, defined by

$$\begin{aligned} P_t f(x) &:= \int_{\mathbb{R}^d} P(t, x, dy) f(dy) \\ &= \mathbb{E}(f(X_t) \mid X_0 = x) \end{aligned} \quad (7)$$

in which case the Chapman–Kolmogorov equation becomes the *semigroup property*

$$P_s P_t = P_{t+s}, \quad s, t \geq 0 \quad (8)$$

Conversely, given a Markov semigroup $(P_t, t \geq 0)$ and a probability measure ν on $\mathbb{R}^d$, it is always possible to construct a Markov process X with initial law ν that satisfies equation (7) (see [9, Th.4.1.1]). The links between PDEs and Markov processes are based on this equivalence between semigroups and Markov processes. This can be expressed through a single object: the infinitesimal generator.

Strong Markov Property, Feller Processes

Recall that a random time τ is called a $\mathcal{F}_t$ -*stopping time* if $\{\tau \leq t\} \in \mathcal{F}_t$ for any $t \geq 0$.

Definition 2 A Markov process $(X_t, t \geq 0)$ with transition function $P(t, x, \Gamma)$ is *strong Markov* if, for any $\mathcal{F}_t$ -stopping time τ ,

$$\mathbb{P}(X_{\tau+t} \in \Gamma \mid \mathcal{F}_\tau) = P(t, X_\tau, \Gamma) \quad (9)$$

for all $t \geq 0$ and $\Gamma \in \mathcal{B}$.

Let $C_0(\mathbb{R}^d)$ denote the space of bounded continuous functions on $\mathbb{R}^d$, which vanish at infinity, equipped with the L^∞ norm denoted by $\|\cdot\|$.

Definition 3 A Feller semigroup^b is a strongly continuous,^c positive, Markov semigroup $(P_t, t \geq 0)$ such that $P_t : C_0(\mathbb{R}^d) \rightarrow C_0(\mathbb{R}^d)$ and

$$\begin{aligned} \forall f \in C_0(\mathbb{R}^d), \quad 0 \leq f &\Rightarrow 0 \leq P_t f \\ \forall f \in C_0(\mathbb{R}^d) \quad \forall x \in \mathbb{R}^d, \quad P_t f(x) &\rightarrow f(x) \text{ as } t \rightarrow 0 \end{aligned} \quad (10)$$

For a Feller semigroup, the corresponding Markov process can be constructed as a strong Markov process.

Theorem 1 ([9] Th.4.2.7). *Given a Feller semigroup $(P_t, t \geq 0)$ and any probability measure ν on $\mathbb{R}^d$, there exists a filtered probability space $(\Omega, \mathcal{F}_t, \mathbb{P})$ and a strong Markov process $(X_t, t \geq 0)$ on this space with values in $\mathbb{R}^d$ with initial law ν and with transition function P_t . A strong Markov process whose semigroup is Feller is called a Feller process.*

Infinitesimal Generator

We are now in a position to introduce the key notion of *infinitesimal generator* of a Feller process.

Definition 4 For a Feller process $(X_t, t \geq 0)$, the *infinitesimal generator* of X is the (generally unbounded) linear operator $L : \mathcal{D}(L) \rightarrow C_0(\mathbb{R}^d)$ defined as follows. We write $f \in \mathcal{D}(L)$ if, for some $g \in C_0(\mathbb{R}^d)$, we have

$$\frac{\mathbb{E}(f(X_t) \mid X_0 = x) - f(x)}{t} \rightarrow g(x) \quad (11)$$

when $t \rightarrow 0$ for the norm $\|\cdot\|$, and we then define $Lf = g$.

By Theorem 1, an equivalent definition can be obtained by replacing X by its Feller semigroup $(P_t, t \geq 0)$. In particular, for all $f \in \mathcal{D}(L)$,

$$Lf(x) = \lim_{t \rightarrow 0} \frac{P_t f(x) - f(x)}{t} \quad (12)$$

An important property of the infinitesimal generator is that it allows one to construct fundamental martingales associated with a Feller process.

Theorem 2 ([21], III.10). *Let X be a Feller process on $(\Omega, \mathcal{F}_t, \mathbb{P})$ with infinitesimal generator L such that $X_0 = x \in \mathbb{R}^d$. For all $f \in \mathcal{D}(L)$,*

$$f(X_t) - f(x) - \int_0^t Lf(X_s) ds \quad (13)$$

defines a $\mathcal{F}_t$ -martingale. In particular,

$$\mathbb{E}(f(X_t)) = f(x) + \mathbb{E}\left(\int_0^t Lf(X_s) ds\right) \quad (14)$$

As explained earlier, the law of a Markov process is characterized by its semigroup. In most cases, a Feller semigroup can be itself characterized by its infinitesimal generator (the precise conditions for

this to hold are given by the Hille–Yosida theorem, see [21, Th.III.5.1]). For almost all Markov financial models, these conditions are well established and always satisfied (see Examples 1, 2, 3, and 4). As illustrated by equation (14), when $\mathcal{D}(L)$ is large enough, the infinitesimal generator captures the law of the whole dynamics of a Markov process and provides an analytical tool to study the Markov process. The other major mathematical tool used in finance is the stochastic calculus (see **Stochastic integral**, **Itô formula**), which applies to **Semimartingales** (see [18]). It is therefore crucial for applications to characterize under which conditions a Markov process is a semimartingale. This question is answered for very general processes in [5]. We mention that this is always the case for Feller diffusions, defined later.

Feller Diffusions

Let us consider the particular case of continuous Markov processes, which include the solutions of stochastic differential equations (SDEs).

Definition 5 A Feller diffusion on $\mathbb{R}^d$ is a Feller process X on $\mathbb{R}^d$ that has continuous paths, and such that the domain $\mathcal{D}(L)$ of the generator L of X contains the space $C_K^\infty(\mathbb{R}^d)$ of infinitely differentiable functions of compact support.

Feller diffusions are Markov processes admitting a second-order differential operator as infinitesimal generator.

Theorem 3 For any $f \in C_K^\infty(\mathbb{R}^d)$, the infinitesimal generator L of a Feller diffusion has the form

$$Lf(x) = \frac{1}{2} \sum_{i,j=1}^d a_{ij}(x) \frac{\partial^2 f}{\partial x_i \partial x_j}(x) + \sum_{i=1}^d b_i(x) \frac{\partial f}{\partial x_i}(x) \quad (15)$$

where the functions $a_{ij}(\cdot)$ and $b_i(\cdot)$, $1 \leq i, j \leq d$ are continuous and the matrix $a = (a_{ij}(x))_{1 \leq i, j \leq d}$ is nonnegative definite symmetric for all $x \in \mathbb{R}^d$.

Kolmogorov Equations

Observe by equation (12) that the semigroup P_t of a Feller process X satisfies the following differential

equation; for all $f \in \mathcal{D}(L)$,

$$\frac{d}{dt} P_t f = L P_t f \quad (16)$$

This equation is called *Kolmogorov's backward equation*. In particular, if L is a differential operator (e.g., if X is a Feller diffusion), the function $u(t, x) = P_t f(x)$ is the solution of the PDE

$$\begin{cases} \frac{\partial u}{\partial t} = Lu \\ u(0, x) = f(x) \end{cases} \quad (17)$$

Conversely, if this PDE admits a unique solution, then its solution is given by

$$u(t, x) = \mathbb{E}(f(X_t) \mid X_0 = x) \quad (18)$$

This is the simplest example of a probabilistic interpretation of the solution of a PDE in terms of a Markov process.

Moreover, because Feller semigroups are strongly continuous, it is easy to check that the operators P_t and L commute. Therefore, equation (16) may be rewritten as

$$\frac{d}{dt} P_t f = P_t L f \quad (19)$$

This equation is known as Kolmogorov's forward equation. It is the weak formulation of the equation

$$\frac{d}{dt} \mu_t^x = L^* \mu_t^x \quad (20)$$

where the probability measure μ_t^x on $\mathbb{R}^d$ denotes the law of X_t conditioned on $X_0 = x$ and where L^* is the adjoint operator of L . In particular, with the notation of Theorem 3, if X is a Feller diffusion and if $\mu_t^x(dy)$ admits a density $q(x; t, y)$ with respect to Lebesgue's measure on $\mathbb{R}^d$ (which holds, e.g., if the functions $b_i(x)$ and $a_{ij}(x)$ are bounded and locally Lipschitz, if the functions $a_{ij}(x)$ are globally Hölder and if the matrix $a(x)$ is uniformly positive definite [10, Th.6.5.2]), the forward Kolmogorov equation is the weak form (in the sense of the distribution theory) of the PDE

$$\begin{aligned} \frac{\partial}{\partial t} q(x; t, y) = & - \sum_{i=1}^d \frac{\partial}{\partial y_i} (b_i(y) q(x; t, y)) \\ & + \sum_{i,j=1}^d \frac{\partial^2}{\partial y_i \partial y_j} (a_{ij}(y) q(x; t, y)) \end{aligned} \quad (21)$$

This equation is known as *Fokker–Planck equation* and gives another family of PDEs that have probabilistic interpretations. Fokker–Planck equation has applications in finance for quantiles, Value at Risk, or risk measure computations [22], whereas Kolmogorov’s backward equation (17) is more suited to financial problems related to the hedging of derivatives products or portfolio allocation (see the section “Parabolic PDEs Associated to Markov Processes”, and sequel).

Time-inhomogeneous Markov Processes

The law of a time-inhomogeneous Markov process is described by the doubly indexed family of operators $(P_{s,t}, 0 \leq s \leq t)$ where, for any bounded measurable f and any $x \in \mathbb{R}^d$,

$$P_{s,t}f(x) = \mathbb{E}(f(X_t) \mid X_s = x) \quad (22)$$

Then, the semigroup property becomes, for $s \leq t \leq r$,

$$P_{s,t}P_{t,r} = P_{s,r} \quad (23)$$

Definition 3 of Feller semigroups can be generalized to time-inhomogeneous processes as follows. The time-inhomogeneous Markov process X is called a *Feller time-inhomogeneous process* if $(P_{s,t}, 0 \leq s \leq t)$ is a family of positive, Markov linear operators on $C_0(\mathbb{R}^d)$ which is strongly continuous in the sense

$$\forall s \geq 0, \quad x \in \mathbb{R}^d, f \in C_0(\mathbb{R}^d), \quad \|P_{s,t}f - f\| \rightarrow 0 \quad (24)$$

as $t \rightarrow s$

In this case, it is possible to generalize the notion of infinitesimal generator. For any t , let

$$\begin{aligned} L_t f(x) &= \lim_{s \rightarrow 0} \frac{P_{t,t+s}f(x) - f(x)}{s} \\ &= \lim_{s \rightarrow 0} \frac{\mathbb{E}[f(X_{t+s}) \mid X_t = x] - f(x)}{s} \end{aligned} \quad (25)$$

for any $f \in C_0(\mathbb{R}^d)$ such that $L_t f \in C_0(\mathbb{R}^d)$ and the limit above holds in the sense of the norm $\|\cdot\|$. The set of such $f \in C_0(\mathbb{R}^d)$ is called the *domain* $\mathcal{D}(L_t)$ of the operator L_t . $(L_t, t \geq 0)$ is called the *family of*

time-inhomogeneous infinitesimal generators of the process X .

All the results on Feller processes stated earlier can be easily transposed to the time-inhomogeneous case, observing that if $(X_t, t \geq 0)$ is a time-inhomogeneous Markov process on $\mathbb{R}^d$, then $(\tilde{X}_t, t \geq 0)$, where $\tilde{X}_t = (t, X_t)$ is a time-homogeneous Markov process on $\mathbb{R}_+ \times \mathbb{R}^d$. Moreover, if X is time-inhomogeneous Feller, it is elementary to check that the process $\tilde{X}$ is time-homogeneous Feller as defined in Definition 3. Its semigroup $(\tilde{P}_t, t \geq 0)$ is linked to the time-inhomogeneous semigroup by the relation

$$\begin{aligned} \tilde{P}_t f(s, x) &= \mathbb{E}[f(s+t, X_{s+t}) \mid X_s = x] \\ &= (P_{s,s+t}f(s+t, \cdot))(x) \end{aligned} \quad (26)$$

for all bounded and measurable $f : \mathbb{R}_+ \times \mathbb{R}^d \rightarrow \mathbb{R}$. If $\tilde{L}$ denotes the infinitesimal generator of the process $\tilde{X}$, it is elementary to check that, for any $f(t, x) \in \mathcal{D}(\tilde{L})$ that is differentiable with respect to t , with derivative uniformly continuous in (t, x) , $x \mapsto f(t, x)$ belongs to $\mathcal{D}(L_t)$ for any $t \geq 0$ and

$$\tilde{L}f(t, x) = \frac{\partial f}{\partial t}(t, x) + (L_t f(t, \cdot))(x) \quad (27)$$

On this observation, it is possible to apply Theorem 3 to time-inhomogeneous Feller diffusions, defined as continuous time-inhomogeneous Feller processes with infinitesimal generators $(L_t, t \geq 0)$ such that $C_K^\infty(\mathbb{R}^d) \subset \mathcal{D}(L_t)$ for any $t \geq 0$. For such processes, there exist continuous functions b_i and a_{ij} , $1 \leq i, j \leq d$ from $\mathbb{R}_+ \times \mathbb{R}^d$ to $\mathbb{R}$ such that the matrix $a(t, x) = (a_{i,j}(t, x))_{1 \leq i, j \leq d}$ is symmetric nonnegative definite and

$$\begin{aligned} L_t f(x) &= \frac{1}{2} \sum_{i,j=1}^d a_{ij}(t, x) \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \\ &\quad + \sum_{i=1}^d b_i(t, x) \frac{\partial f}{\partial x_i}(x) \end{aligned} \quad (28)$$

for all $t \geq 0$, $x \in \mathbb{R}^d$ and $f \in C_K^\infty(\mathbb{R}^d)$.

For more details on time-inhomogeneous Markov processes, we refer to [10].

Example 1 Brownian Motion The standard one-dimensional Brownian motion $(B_t, t \geq 0)$ is a Feller diffusion in $\mathbb{R}$ ($d = 1$) such that $B_0 = 0$ and for

which the parameters of Theorem 3 are $b = 0$ and $a = 1$. The Brownian motion is the fundamental prototype of Feller diffusions. Other diffusions are inherited from this process because they can be expressed as solutions to SDEs driven by independent Brownian motions (see later). Similarly, the standard d -dimensional Brownian motion is a vector of d independent standard one-dimensional Brownian motions and corresponds to the case $b_i = 0$ and $a_{ij} = \delta_{ij}$ for $1 \leq i, j \leq d$, where δ_{ij} is the Kronecker delta function ($\delta_{ij} = 1$ if $i = j$ and 0 otherwise).

Example 2 Black–Scholes Model In the Black–Scholes model, the underlying asset price S_t follows a geometric Brownian motion with constant drift μ and volatility σ .

$$S_t = S_0 \exp((\mu - \sigma^2/2)t + \sigma B_t) \quad (29)$$

where B is a standard Brownian motion. With Itô's formula, it is easily checked that S is a Feller diffusion with infinitesimal generator

$$Lf(x) = \mu x f'(x) + \frac{1}{2} \sigma^2 x^2 f''(x) \quad (30)$$

Itô's formula also yields

$$S_t = S_0 + \mu \int_0^t S_s ds + \sigma \int_0^t S_s dB_s \quad (31)$$

which can be written as the SDE

$$dS_t = \mu S_t dt + \sigma S_t dB_t \quad (32)$$

The correspondence between the SDE and the second-order differential operator L appears below as a general fact.

Example 3 Stochastic Differential Equations SDEs are probably the most used Markov models in finance. Solutions of SDEs are examples of Feller diffusions. When the parameters b_i and a_{ij} of Theorem 3 are sufficiently regular, a Feller process X with generator equation (15) can be constructed as the solution of the SDE

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t \quad (33)$$

where $b(x) \in \mathbb{R}^d$ is $(b_1(x), \dots, b_d(x))$, where the $d \times r$ matrix $\sigma(x)$ satisfies $a_{ij}(x) = \sum_{k=1}^r \sigma_{ik}(x)\sigma_{jk}(x)$

(i.e., $a = \sigma\sigma'$) and where B_t is a r -dimensional standard Brownian motion. For example, when $d = r$, one can take for $\sigma(x)$ the symmetric square root matrix of the matrix $a(x)$.

The construction of Markov solutions to the SDE (33) with generator (15) is possible if b and σ are globally Lipschitz with linear growth [13, Th.5.2.9], or if b and a are bounded and continuous functions [13, Th.5.4.22]. In the second case, the SDE has a solution in a weaker sense. Uniqueness (at least in law) and the strong Markov property hold if b and σ are locally Lipschitz [13, Th.5.2.5], or if b and a are Hölder continuous and the matrix a is uniformly positive definite [13, Rmk.5.4.30, Th.5.4.20]. In the one-dimensional case, existence and uniqueness for the SDE (32) can be proved under weaker assumptions [13, Sec.5.5].

In all these cases, the Markov property allows one to identify the SDE (33) with its generator (15). This will allow us to make the link between parabolic PDEs and the corresponding SDE in the section “Parabolic PDEs Associated to Markov Processes” and sequel.

Similarly, one can associate to the time-inhomogeneous SDE

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t \quad (34)$$

the time-inhomogeneous generators (28). Existence for this SDE holds if b_t and σ_{ij} are globally Lipschitz in x and locally bounded (uniqueness holds if b_t and σ_{ij} are only locally Lipschitz in x). As earlier, in this case, a solution to equation (34) is strong Markov. We refer the reader to [16] for more details.

Example 4 Backward Stochastic Differential Equations Backward stochastic differential equations are SDEs where a random variable is given as a terminal condition. Let us motivate the definition of a backward SDE (BSDE) by continuing the study of the elementary example of the introduction of this article.

Consider an asset S_t modeled by the Black–Scholes SDE (32) and assume that it is possible to borrow and lend cash at a constant risk-free interest rate r . A self-financed trading strategy is determined by an initial portfolio value and the amount π_t of the portfolio value placed in the risky asset at time t . Given the stochastic process $(\pi_t, t \geq 0)$, the portfolio

value V_t at time t solves the SDE

$$dV_t = rV_t dt + \pi_t(\mu - r) dt + \sigma\pi_t dB_t \quad (35)$$

where B is the Brownian motion driving the dynamics (32) of the risky asset S .

Assume that this portfolio serves to hedge a call option with strike K and maturity T . This problem can be expressed as finding a couple of processes (V_t, π_t) adapted to the Brownian filtration $\mathcal{F}_t = \sigma(B_s, s \leq t)$ such that

$$V_t = (S_T - K)^+ - \int_t^T (rV_s + \pi_s(\mu - r)) ds - \int_t^T \sigma\pi_s dB_s \quad (36)$$

Such SDEs with terminal condition and with unknown process driving the Brownian integral are called *BSDEs*. This particular BSDE admits a unique solution (see the section “Quasi- and Semilinear PDEs and BSDEs”) and can be explicitly solved. Because V_0 is $\mathcal{F}_0$ adapted, it is nonrandom and therefore V_0 is the usual free arbitrage price of the option. In particular, choosing $\mu = r$, we recover the usual formula for the free arbitrage price $V_0 = \mathbb{E}[e^{-rT}(S_T - K)^+]$, and the quantity of risky asset π_t/S_t in the portfolio is given by the Black–Scholes Δ -hedge $\partial u / \partial x(t, S_t)$, where $u(t, x)$ is the solution of the Black–Scholes PDE (see **Exchange Options**)

$$\begin{cases} \frac{\partial u}{\partial t} + rx \frac{\partial u}{\partial x} + \frac{\sigma^2}{2} x^2 \frac{\partial^2 u}{\partial x^2} - ru = 0 \\ \quad \forall (t, x) \in [0, T) \times (0, +\infty) \\ u(T, x) = f(x) \quad \forall x \in (0, +\infty) \end{cases} \quad (37)$$

Applying Itô formula to $u(t, S_t)$, an elementary computation shows that $u(t, S_t)$ solves the same SDE (35) with $\mu = r$ as V_t , with the same terminal condition. Therefore, by uniqueness, $V_t = u(t, S_t)$.

Usually, for more general BSDEs, $(\pi_t, t \geq 0)$ is an implicit process given by the martingale representation theorem. In the section “Quasi- and Semilinear PDEs and BSDEs”, we give results on the existence and uniqueness of solutions of BSDEs, and on their links with nonlinear PDEs.

Discontinuous Markov Processes

In financial models, it is sometimes natural to consider discontinuous Markov processes, for example, when one wants to take into account jumps in prices. This can sometimes be done by modeling the dynamics using **Poisson processes**, **Lévy processes** or other jump processes (see **Jump Processes**). In particular, it is possible to define SDEs where the Brownian motion is replaced by a Lévy process (see **CGMY model**, **NIG model**, or **Generalized hyperbolic model** for examples). In this situation, the generator is an integro-differential operator and the parabolic PDE is replaced by **Partial integro-differential Equations**.

Dimension of the State Space

In many pricing/hedging problems, the dimension of the pricing PDE is greater than the state space of the underlyings. In such cases, the financial problem is apparently related to non-Markov stochastic processes. However, it can usually be expressed in terms of Markov processes if one increases the dimension of the process considered. For example, in the context of Markov **short rates** $(r_t, t \geq 0)$, the pricing of a **zero-coupon bond** is expressed in terms of the process $R_t = \int_0^t r_s ds$ which is not Markovian, whereas the couple (r_t, R_t) is Markovian. For **Asian options** on a Markov asset, the couple formed by the asset and its integral is Markovian. If the asset involves a stochastic volatility solution to a SDE (see **Heston model** and **SABR model**), then the couple formed by the asset value and its volatility is Markov. As mentioned earlier, another important example is given by time-inhomogeneous Markov processes that become time homogeneous when one considers the couple formed by the current time and the original process.

In some cases, the dimension of the system can be reduced while preserving the Markovian nature of the problem. In the case of the portfolio management of multidimensional Black–Scholes prices with deterministic volatility matrix, mean return vector and interest rate, the dimension of the problem is actually reduced to one (see **Merton problem**). When the volatility matrix, the mean return vector, and the interest rate are Markov processes of dimension d' , the dimension of the problem is reduced to $d' + 1$.

Parabolic PDEs Associated to Markov Processes

Computing the value of any future claim with fixed maturity (for example, the price of an European option on an asset solution to a SDE), or solving an optimal portfolio management problem, amounts to solve a parabolic second-order PDE, that is a PDE of the form

$$\begin{aligned} \frac{\partial u}{\partial t}(t, x) + L_t u(t, x) \\ = f(t, x, u(t, x), \nabla u(t, x)), \quad (t, x) \in \mathbb{R}_+ \times \mathbb{R}^d \end{aligned} \quad (38)$$

where $\nabla u(t, x)$ is the gradient of $u(t, x)$ with respect to x and the linear differential operators L_t has the form equation (28).

The goal of this section is to explain the links between these PDEs and the original diffusion process, or some intermediate Markov process. We will distinguish between *linear* parabolic PDEs, where the function $f(t, x, y, z)$ does not depend on z and is linear in y , *semilinear* parabolic PDEs, where the function $f(t, x, y, z)$ does not depend on z but is nonlinear in y , and *quasi-linear* parabolic PDEs, where the function $f(t, x, y, z)$ is nonlinear in (y, z) . We will also discuss the links between diffusion processes and some *fully nonlinear* PDEs (Hamilton–Jacobi–Bellman (HJB) equations or variational inequalities) of the form

$$\begin{aligned} F\left(t, \frac{\partial u}{\partial t}(t, x), u(t, x), \nabla u(t, x), Hu(t, x)\right) = 0, \\ (t, x) \in \mathbb{R}_+ \times \mathbb{R}^d \end{aligned} \quad (39)$$

for some nonlinear function F , where Hu denotes the Hessian matrix of u with respect to the space variable x .

Such problems involve several notions of solutions discussed in the literature (*see viscosity solution*). In the sections “Brownian Motion, Ornstein–Uhlenbeck Process, and the Heat Equation” and “Linear Case”, we consider *classical solutions*, that is, solutions that are continuously differentiable with respect to the time variable, and twice continuously differentiable with respect to the space variables. In the sections “Quasi- and Semilinear PDEs and BSDEs” and

“Optimal Control, Hamilton–Jacobi–Bellman Equations, and Variational Inequalities”, because of the nonlinearity of the problem, classical solutions may not exist, and one must consider the weaker notion of viscosity solutions.

In the section “Brownian Motion, Ornstein–Uhlenbeck Process, and the Heat Equation”, we consider heat-like equations where the solution can be explicitly computed. The section “Linear Case” deals with linear PDEs, the section “Quasi- and Semilinear PDEs and BSDEs” deals with quasi- and semilinear PDEs and their links with BSDEs, and the section “Optimal Control, Hamilton–Jacobi–Bellman Equations, and Variational Inequalities” deals with optimal control problems.

Brownian Motion, Ornstein–Uhlenbeck Process, and the Heat Equation

The *heat equation* is the first example of a parabolic PDE with basic probabilistic interpretation (for which there is no need of stochastic calculus).

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = \frac{1}{2} \Delta u(t, x), & (t, x) \in (0, +\infty) \times \mathbb{R}^d \\ u(0, x) = f(x), & x \in \mathbb{R}^d \end{cases} \quad (40)$$

where Δ denotes the Laplacian operator of $\mathbb{R}^d$. When f is a bounded measurable function, it is well known that the solution of this problem is given by the formula

$$u(t, x) = \int_{\mathbb{R}^d} f(y) g(x; t, y) dy \quad (41)$$

where

$$g(x; t, y) = \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{|x - y|^2}{2t}\right) \quad (42)$$

$|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$. g is often called the *fundamental solution* of the heat equation. We recognize that $g(x; t, y) dy$ is the law of $x + B_t$ where B is a standard d -dimensional Brownian motion. Therefore, equation (41) may be rewritten as

$$u(t, x) = \mathbb{E}[f(x + B_t)] \quad (43)$$

which provides a simple probabilistic interpretation of the solution of the heat equation in $\mathbb{R}^d$ as a particular case of equation (18). Note that equation (40)

involves the infinitesimal generator of the Brownian motion $(1/2)\Delta$.

Let us mention two other cases where the link between PDEs and stochastic processes can be done without stochastic calculus. The first one is the Black–Sholes model, solution to the SDE

$$dS_t = S_t(\mu dt + \sigma dB_t) \quad (44)$$

When $d = 1$, its infinitesimal generator is $Lf(x) = \mu xf'(x) + (\sigma^2/2)x^2 f''(x)$ and its law at time t when $S_0 = x$ is $l(x; t, y)$ dy where

$$l(x; t, y) = \frac{1}{\sigma y \sqrt{2\pi t}} \times \exp \left[-\frac{1}{2\sigma^2 t} \left(\log \frac{y}{x} - \left(\mu - \frac{\sigma^2}{2} \right) t \right)^2 \right] \quad (45)$$

Then, for any bounded and measurable f , elementary computations show that

$$u(t, x) = \int_0^\infty f(y) l(x; t, y) dy \quad (46)$$

satisfies

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = Lu(t, x), & (t, x) \in (0, +\infty)^2 \\ u(0, x) = f(x), & x \in (0, +\infty) \end{cases} \quad (47)$$

Here again, this formula gives immediately the probabilistic interpretation

$$u(t, x) = \mathbb{E}[f(S_t) \mid S_0 = x] \quad (48)$$

The last example is the **Ornstein–Uhlenbeck** process in $\mathbb{R}$

$$dX_t = \beta X_t dt + \sigma dB_t \quad (49)$$

with $\beta \in \mathbb{R}$, $\sigma > 0$ and $X_0 = x$. The infinitesimal generator of this process is $Af(x) = \beta xf'(x) + (\sigma^2/2)f''(x)$. It can be easily checked that X_t is a Gaussian random variable with mean $x \exp(\beta t)$ and variance $\sigma^2(\exp(2\beta t) - 1)/2\beta$ with the convention that $(\exp(2\beta t) - 1)/2\beta = t$ if $\beta = 0$. Therefore, its probability density function is given by

$$h(x; t, y) = \sqrt{\frac{\beta}{\sigma^2 \pi (\exp(2\beta t) - 1)}}$$

$$\times \exp \left[-\frac{2\beta(y - x \exp(\beta t))^2}{\sigma^2(\exp(2\beta t) - 1)} \right] \quad (50)$$

Then, for any bounded and measurable f ,

$$u(t, x) = \int_{\mathbb{R}} f(y) h(x; t, y) dy = \mathbb{E}[f(X_t) \mid X_0 = x] \quad (51)$$

is solution of

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) = Au(t, x), & (t, x) \in (0, +\infty) \times \mathbb{R} \\ u(0, x) = f(x), & x \in \mathbb{R} \end{cases} \quad (52)$$

Linear Case

The probabilistic interpretations of the previous PDEs can be generalized to a large class of linear parabolic PDEs with arbitrary second-order differential operator, interpreted as the infinitesimal generator of a Markov process. Assume that the vector $b(t, x) \in \mathbb{R}^d$ and the $d \times r$ matrix $\sigma(t, x)$ are uniformly bounded and locally Lipschitz functions on $[0, T] \times \mathbb{R}^d$ and consider the SDE in $\mathbb{R}^d$

$$dX_t = b(t, X_t) dt + \sigma(t, X_t) dB_t \quad (53)$$

where B is a standard r -dimensional Brownian motion. Set $a = \sigma\sigma'$ and assume also that the $d \times d$ matrix $a(t, x)$ is uniformly Hölder and satisfies the uniform ellipticity condition: there exists $\gamma > 0$ such that for all $(t, x) \in [0, T] \times \mathbb{R}^d$ and $\xi \in \mathbb{R}^d$,

$$\sum_{i,j=1}^d a_{ij}(t, x) \xi_i \xi_j \geq \gamma |\xi|^2 \quad (54)$$

Let $(L_t)_{t \geq 0}$ be the family of time-inhomogeneous infinitesimal generators of the Feller diffusion X_t solution to the SDE (53), given by equation (28). Consider the Cauchy problem

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) + L_t u(t, x) + c(t, x)u(t, x) = f(t, x), & (t, x) \in [0, T] \times \mathbb{R}^d \\ u(T, x) = g(x), & x \in \mathbb{R}^d \end{cases} \quad (55)$$

where $c(t, x)$ is uniformly bounded and locally Hölder on $[0, T] \times \mathbb{R}^d$, $f(t, x)$ is locally Hölder on $[0, T] \times \mathbb{R}^d$, $g(x)$ is continuous on $\mathbb{R}^d$ and

$$\begin{aligned} |f(t, x)| + |g(x)| &\leq A \exp(a|x|), \\ \forall (t, x) \in [0, T] \times \mathbb{R}^d \end{aligned} \quad (56)$$

for some constants $A, a > 0$. Under these conditions, it follows easily from Theorems 6.4.5 and 6.4.6 of [10] that equation (55) admits a unique classical solution u such that

$$|u(t, x)| \leq A' \exp(a|x|) \quad \forall (t, x) \in [0, T] \times \mathbb{R}^d \quad (57)$$

for some constant $A' > 0$.

The following result is known as *Feynman–Kac formula* and can be deduced from equation (57) using exactly the same method as for [10, Th.6.5.3] and using the fact that, under our assumptions, X_t has finite exponential moments [10, Th.6.4.5].

Theorem 4 *Under the previous assumptions, the solution of the Cauchy problem (55) is given by*

$$\begin{aligned} u(t, x) &= \mathbb{E} \left[g(X_T) \exp \left(\int_t^T c(s, X_s) ds \right) \mid X_t = x \right] \\ &\quad - \mathbb{E} \left[\int_t^T f(s, X_s) \right. \\ &\quad \times \exp \left(\int_t^s c(\alpha, X_\alpha) d\alpha \right) ds \mid X_t = x \Big] \end{aligned} \quad (58)$$

Let us mention that this result can be extended to parabolic linear PDEs on bounded domains [10, Th.6.5.2] and to elliptic linear PDEs on bounded domains [10, Th.6.5.1].

Example 5 European Options The Feynman–Kac formula has many applications in finance. Let us consider the case of an European option on a one-dimensional Markov asset $(S_t, t \geq 0)$ with payoff

$g(S_u, 0 \leq u \leq T)$. The free arbitrage value at time t of this option is

$$V_t = \mathbb{E}[e^{-r(T-t)} g(S_u, t \leq u \leq T) \mid \mathcal{F}_t] \quad (59)$$

By the Markov property (1), this quantity only depends on S_t and t [10, Th.2.1.2]. The Feynman–Kac formula (58) allows one to characterize V in the case where g depends only on S_T and S is a Feller diffusion.

Most often, the asset SDE

$$dS_t = S_t(\mu(t, S_t) dt + \sigma(t, S_t) dB_t) \quad (60)$$

cannot satisfy the uniform ellipticity assumption (54) in the neighborhood of 0. Therefore, Theorem 4 does not apply directly. This is a general difficulty for financial models. However, in most cases (and in all the examples below), it can be overcome by taking the logarithm of the asset price. In our case, we assume that the process $(\log S_t, 0 \leq t \leq T)$ is a Feller diffusion on $\mathbb{R}$ with time-inhomogeneous generator

$$L_t \phi(y) = \frac{1}{2} a(t, y) \phi''(y) + b(t, y) \phi'(y) \quad (61)$$

that satisfy the assumptions of Theorem 4. This holds for example for the Black–Scholes model (32). This assumption implies that S is a Feller diffusion on $(0, +\infty)$ whose generator takes the form

$$\tilde{L}_t \phi(x) = \frac{1}{2} \tilde{a}(t, x) x^2 \phi''(x) + \tilde{b}(t, x) x \phi'(x) \quad (62)$$

where $\tilde{a}(t, x) = a(t, \log x)$ and $\tilde{b}(t, x) = b(t, \log x) + a(t, \log x)/2$.

Assume also that $g(x)$ is continuous on $\mathbb{R}_+$ with polynomial growth when $x \rightarrow +\infty$. Then, by Theorem 4, the function

$$v(t, y) = \mathbb{E}[e^{-r(T-t)} g(S_T) \mid \log S_t = y] \quad (63)$$

is solution to the Cauchy problem

$$\begin{cases} \frac{\partial v}{\partial t}(t, y) + L_t v(t, y) \\ -rv(t, y) = 0, & (t, y) \in [0, T] \times \mathbb{R} \\ v(T, y) = g(\exp(y)), & y \in \mathbb{R} \end{cases} \quad (64)$$

Making the change of variable $x = \exp(y)$, $u(t, x) = v(t, \log x)$ is solution to

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) + \tilde{b}(t, x)x \frac{\partial u}{\partial x}(t, x) + \frac{1}{2}\tilde{a}(t, x)x^2 \frac{\partial^2 u}{\partial x^2}(t, x) - rv(t, x) = 0, & (t, x) \in [0, T) \times (0, +\infty) \\ u(T, x) = g(x), & x \in (0, +\infty) \end{cases} \quad (65)$$

and $V_t = u(t, S_t)$. The Black–Scholes PDE (37) is a particular case of this result.

Example 6 An Asian Option We give an example of a path-dependent option for which the uniform ellipticity condition of the matrix a does not hold. An Asian option is an option where the payoff is determined by the average of the underlying price over the period considered. Consider the Asian call option

$$\left(\frac{1}{T} \int_0^T S_u du - K \right)^+ \quad (66)$$

on a Black–Scholes asset $(S_t, t \geq 0)$ following

$$dS_t = rS_t dt + \sigma S_t dB_t \quad (67)$$

$$\begin{cases} \frac{\partial u}{\partial t} + \frac{\sigma^2}{2}x^2 \frac{\partial^2 u}{\partial x^2} + rx \frac{\partial u}{\partial x} + \frac{1}{T}x \frac{\partial u}{\partial y} - ru = 0, & (t, x, y) \in [0, T) \times (0, +\infty) \times \mathbb{R} \\ u(T, x, y) = (y/T - K)^+, & (x, y) \in (0, +\infty) \times \mathbb{R} \end{cases} \quad (72)$$

where B is a standard one-dimensional Brownian motion. The free arbitrage price at time t is

$$\mathbb{E} \left[e^{-r(T-t)} \left(\frac{1}{T} \int_0^T S_u du - K \right)^+ \middle| S_t \right] \quad (68)$$

To apply the Feynman–Kac formula, one must express this quantity as the (conditional) expectation

$$\begin{cases} \frac{\partial \varphi}{\partial t}(t, z) + \frac{\sigma^2}{2}z^2 \frac{\partial^2 \varphi}{\partial z^2}(t, z) - \left(\frac{1}{T} + rz \right) \frac{\partial \varphi}{\partial z}(t, z) + r\varphi(t, z) = 0, & (t, z) \in [0, T) \times \mathbb{R} \\ \varphi(T, z) = -(z)^+/T, & z \in \mathbb{R} \end{cases} \quad (74)$$

of the value at time T of some Markov quantity. This can be done by introducing the process

$$A_t = \int_0^t S_u du, \quad 0 \leq t \leq T \quad (69)$$

It is straightforward to check that (S, A) is a Feller diffusion on $(0, +\infty)^2$ with infinitesimal generator

$$\begin{aligned} Lf(x, y) &= rx \frac{\partial f}{\partial x}(x, y) + \frac{\sigma^2}{2}x^2 \frac{\partial^2 f}{\partial x^2}(x, y) \\ &\quad + \frac{1}{T}x \frac{\partial f}{\partial y}(x, y) \end{aligned} \quad (70)$$

Although considering the change of variable $(\log S, A)$, Theorem 4 does not apply to this process because the infinitesimal generator is degenerated (without second-order derivative in y). Formally, the Feynman–Kac formula would give that

$$\begin{aligned} u(t, x, y) &:= \mathbb{E}[e^{-r(T-t)}(A_T/T - K)^+ \mid S_t = x, A_t = y] \\ &\quad (71) \end{aligned}$$

is solution to the PDE

Actually, it is possible to justify the previous statement in the specific case of a one-dimensional Black–Scholes asset: u can be written as

$$u(t, x, y) = e^{-r(T-t)} x \varphi \left(t, \frac{KT - y}{x} \right) \quad (73)$$

(see [20]) where $\varphi(t, z)$ is the solution of the one-dimensional parabolic PDE

From this, it is easy to check that u solves equation (72).

Note that this relies heavily on the fact that the underlying asset follows the Black–Scholes model. As far as we know, no rigorous justification of

Feynman–Kac formula is available for Asian options on more general assets.

Quasi- and Semilinear PDEs and BSDEs

The link between quasi- and semilinear PDEs and BSDEs is motivated by the following formal argument. Consider the semilinear PDE

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) + L_t u(t, x) = f(u(t, x)) \\ (t, x) \in (0, T) \times \mathbb{R} \\ u(T, x) = g(x) \quad x \in \mathbb{R} \end{cases} \quad (75)$$

where (L_t) is the family of infinitesimal generators of a time-inhomogeneous Feller diffusion $(X_t, t \geq 0)$. Assume that this PDE admits a classical solution $u(t, x)$. Assume also that we can find a unique adapted process $(Y_t, 0 \leq t \leq T)$ such that

$$Y_t = \mathbb{E}[g(X_T) - \int_t^T f(Y_s) ds \mid \mathcal{F}_t] \quad \forall t \in [0, T] \quad (76)$$

Now, by Itô's formula applied to $u(t, X_t)$,

$$u(t, X_t) = \mathbb{E}[g(X_T) - \int_t^T f(u(s, X_s)) ds \mid \mathcal{F}_t] \quad (77)$$

Therefore, $Y_t = u(t, X_t)$ and the stochastic process Y provides a probabilistic interpretation of the solution of the PDE (75). Now, by the martingale decomposition theorem, if Y satisfies (76), there exists an adapted process $(Z_t, 0 \leq t \leq T)$ such that

$$Y_t = g(X_T) - \int_t^T f(Y_s) ds - \int_t^T Z_s dB_s \quad \forall t \in [0, T] \quad (78)$$

solution of the SDE $dY_t = f(Y_t) dt + Z_t dB_t$ with terminal condition $Y_T = g(X_T)$.

The following definition of a BSDE generalizes the previous situation. Given functions $b_i(t, x)$ and $\sigma_{ij}(t, x)$ that are globally Lipschitz in x and locally bounded ($1 \leq i, j \leq d$) and a standard d -dimensional Brownian motion B , consider the unique solution X of the time-inhomogeneous SDE

$$dX_t = b(t, X_t) dt + \sigma(t, X_t) dB_t \quad (79)$$

with initial condition $X_0 = x$. Consider also two functions $f : [0, T] \times \mathbb{R}^d \times \mathbb{R}^k \times \mathbb{R}^{k \times d} \rightarrow \mathbb{R}^k$ and $g : \mathbb{R}^d \rightarrow \mathbb{R}^k$. We say that $((Y_t, Z_t), t \geq 0)$ solve the BSDE

$$dY_t = f(t, X_t, Y_t, Z_t) dt + Z_t dB_t \quad (80)$$

with terminal condition $g(X_T)$ if Y and Z are progressively measurable processes with respect to the Brownian filtration $\mathcal{F}_t = \sigma(B_s, s \leq t)$ such that, for any $0 \leq t \leq T$,

$$Y_t = g(X_T) - \int_t^T f(s, X_s, Y_s, Z_s) ds - \int_t^T Z_s dB_s \quad (81)$$

Example 4 corresponds to $g(x) = (x - K)^+$, $f(t, x, y, z) = -ry + z(\mu - r)/\sigma$ and $Z_t = \sigma \pi_t$. Note that the role of the implicit unknown process Z is to make Y adapted.

The existence and uniqueness of (Y, Z) solving equation (81) hold under the assumptions that $g(x)$ is continuous with polynomial growth in x , $f(t, x, y, z)$ is continuous with polynomial growth in x and linear growth in y and z , and f is uniformly Lipschitz in y and z . Let us denote by (A) all these assumptions. We refer to [17] for the proof of this result and the general theory of BSDEs (*see also forward-backward SDEs*).

Consider the quasi-linear parabolic PDE

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) + L_t u(t, x) = f(t, x, u(t, x), \nabla_x u(t, x) \sigma(t, x)), & (t, x) \in (0, T) \times \mathbb{R}^d \\ u(T, x) = g(x), & x \in \mathbb{R}^d \end{cases} \quad (82)$$

where B is the same Brownian motion as the one driving the Feller diffusion X . In other words, Y is

The following results give the links between the BSDE (80) and the PDE (82).

Theorem 5 ([15], Th.4.1). Assume that $b(t, x)$, $\sigma(t, x)$, $f(t, x, y, z)$, and $g(x)$ are continuous and differentiable with respect to the space variables x, y, z with uniformly bounded derivatives. Assume also that b, σ , and f are uniformly bounded and that $a = \sigma \sigma'$ is uniformly elliptic. Then equation (82) admits a unique classical solution u and

$$Y_t = u(t, X_t) \quad \text{and} \quad Z_t = \nabla_x u(t, X_t) \sigma(t, X_t) \quad (83)$$

Theorem 6 ([17], Th.2.4). Assume (A) and that $b(t, x)$ and $\sigma(t, x)$ are globally Lipschitz in x and locally bounded. Define the function $u(t, x) = Y_t^{t,x}$, where $Y^{t,x}$ is the solution to the BSDE (82) on the time interval $[t, T]$, where X is solution to the SDE (79) with initial condition $X_t = x$. Then u is a viscosity solution of equation (82).

Theorem 5 gives an interpretation of the solution of a BSDE in terms of the solution of a quasi-linear PDE. In particular, in Example 4, it gives the usual interpretation of the hedging strategy $\pi_t = Z_t/\sigma$ as the Δ -hedge of the option price. Note also that Theorem 5 implies that the process (X, Y, Z) is Markov—a fact which is not obvious from the definition. Conversely, Theorem 6 shows how to construct a viscosity solution of a quasi-linear PDE from BSDEs.

BSDEs provide an indirect tool to compute quantities related to a solution X of the SDE (such as the hedging price and strategy of an option based on the process X). BSDEs also have links with general stochastic control problems, that we will not mention (see **BSDEs**). Here, we give an example of application to the pricing of an American put option.

Example 7 Pricing of an American Put Option Consider a Black–Scholes underlying asset S and assume for simplicity that the risk-free interest rate r is zero. The price of an **American put option** on S with strike K and maximal exercise policy T is given by

$$\sup_{0 \leq \tau \leq T} \mathbb{E}^*[(K - S_\tau)^+] \quad (84)$$

where τ is a stopping time and where $\mathbb{P}^*$ is the risk-neutral probability measure, under which the process S is simply a Black–Scholes asset with zero drift.

In the case of an European put option, the price is given by the solution of the BSDE

$$Y_t = (K - S_T)^+ - \int_t^T Z_s dB_s \quad (85)$$

by a similar argument as in Example 4. In the case of an American put option, the price at time t is necessarily bigger than $(K - S_t)^+$. It is therefore natural to include this condition by considering the BSDE (85) reflected on the obstacle $(K - S_t)^+$. Mathematically, this corresponds to the problem of finding adapted processes Y, Z , and R such that

$$\begin{cases} Y_t = (K - S_T)^+ - \int_t^T Z_s dB_s + R_T - R_t \\ Y_t \geq (K - S_t)^+ \\ R \text{ is continuous, increasing, } R_0 = 0 \text{ and} \\ \int_0^T [Y_t - (K - S_t)^+] dR_t = 0 \end{cases} \quad (86)$$

The process R increases only when $Y_t = (K - S_t)^+$ in such a way that Y cannot cross this obstacle. The existence of a solution of this problem is a particular case of general results, (see [7]). As a consequence of the following theorem, this reflected BSDE gives a way to compute the price of the American put option.

Theorem 7 ([7], Th.7.2). The American put option has the price Y_0 , where (Y, Z, R) solves the reflected BSDE (86).

The essential argument of the proof is the following. Fix $t \in [0, T)$ and a stopping time $\tau \in [t, T]$. Since

$$Y_\tau - Y_t = R_t - R_\tau + \int_t^\tau Z_s dB_s \quad (87)$$

and because R is increasing, $Y_t = \mathbb{E}^*[Y_\tau + R_\tau - R_t | \mathcal{F}_t] \geq \mathbb{E}^*[(K - S_\tau)^+ | \mathcal{F}_t]$. Conversely, if $\tau_t^* = \inf\{u \in [t, T] : Y_u = (K - S_u)^+\}$, because $Y > (K - S)^+$ on $[t, \tau_t^*)$, R is constant on this interval and

$$Y_t = \mathbb{E}^*[Y_{\tau_t^*} + R_{\tau_t^*} - R_t | \mathcal{F}_t] = \mathbb{E}^*[(K - S_{\tau_t^*})^+] \quad (88)$$

Therefore,

$$Y_t = \text{ess sup}_{t \leq \tau \leq T} \mathbb{E}^*[(K - S_\tau)^+ | \mathcal{F}_t] \quad (89)$$

which gives another interpretation for the solution Y of the reflected BSDE. Applying this for $t = 0$ yields $Y_0 = \sup_{\tau \leq T} \mathbb{E}^*[(K - S_\tau)^+]$ as stated.

Moreover, as shown by the previous computation, the process Y provides an interpretation of the optimal exercise policy as the first time where Y hits the obstacle $(K - S)^+$. This fact is actually natural from equation (89); the optimal exercise policy is the first time where the current payoff equals the maximal future expected payoff.

As it will appear in the next section, as the solution of an optimal stopping problem, if $S_0 = x$, the price of this American put option is $u(0, x)$, where u is the solution of the nonlinear PDE

$$\begin{cases} \min \left\{ u(t, x) - (K - x)^+; -\frac{\partial u}{\partial t}(t, x) - \frac{\sigma^2}{2} x^2 \frac{\partial^2 u}{\partial x^2}(t, x) \right\} = 0, & (t, x) \in (0, T) \times (0, +\infty) \\ u(T, x) = (K - x)^+, & x \in (0, +\infty) \end{cases} \quad (90)$$

Therefore, similarly as in Theorem 6, the reflected BSDE (84) provides a probabilistic interpretation of the solution of this PDE.

The (formal) essential argument of the proof of this result can be summarized as follows (for details, see [14, Section V.3.1]). Consider the solution u of equation (90) and apply Itô's formula to $u(t, S_t)$. Then, for any stopping time $\tau \in [0, T]$,

$$\begin{aligned} u(0, x) &= \mathbb{E}[u(\tau, S_\tau)] - \mathbb{E} \left[\int_0^\tau \left(\frac{\partial u}{\partial t}(t, S_t) \right. \right. \\ &\quad \left. \left. + \frac{\sigma^2}{2} S_t^2 \frac{\partial^2 u}{\partial x^2}(t, S_t) \right) ds \right] \end{aligned} \quad (91)$$

Because u is solution of equation (90), $u(0, x) \geq \mathbb{E}[u(\tau, S_\tau)] \geq \mathbb{E}[(K - S_\tau)^+]$. Hence, $u(0, x) \geq \sup_{0 \leq \tau \leq T} \mathbb{E}[(K - S_\tau)^+]$.

Conversely, if $\tau^* = \inf\{0 \leq t \leq T : u(t, S_t) = (K - S_t)^+\}$, then

$$\frac{\partial u}{\partial t}(t, S_t) + \frac{\sigma^2}{2} S_t^2 \frac{\partial^2 u}{\partial x^2}(t, S_t) = 0 \quad \forall t \in [0, \tau^*] \quad (92)$$

Therefore, for $\tau = \tau^*$, all the inequalities in the previous computation are equalities and $u(0, x) = \sup_{0 \leq \tau \leq T} \mathbb{E}[(K - S_\tau)^+]$.

Optimal Control, Hamilton–Jacobi–Bellman Equations, and Variational Inequalities

We discuss only two main families of **stochastic control problems**: finite horizon and the optimal stopping problems. Other classes of optimal problems appearing in finance are mentioned in the end of this section.

Finite Horizon Problems

The study of optimal control problems with finite horizon is motivated, for example, by the ques-

tions of portfolio management, quadratic hedging of options, or **super-hedging** cost for **uncertain volatility models**.

Let us consider a controlled diffusion X^α in $\mathbb{R}^d$ solution to the SDE

$$dX_t^\alpha = b(X_t^\alpha, \alpha_t) dt + \sigma(X_t^\alpha) dB_t \quad (93)$$

where B is a standard r -dimensional Brownian motion and the control α is a given progressively measurable process taking values in some compact metric space A . Such a control is called *admissible*. For simplicity, we consider the time-homogeneous case and we assume that the control does not act on the diffusion coefficient σ of the SDE. Assume that $b(x, a)$ is bounded, continuous, and Lipschitz in the variable x , uniformly in $a \in A$. Assume also that σ is Lipschitz and bounded. For any $a \in A$, we introduce the linear differential operator

$$\begin{aligned} L^a \varphi &= \frac{1}{2} \sum_{i,j=1}^d \left(\sum_{k=1}^d \sigma_{ik}(x) \sigma_{jk}(x) \right) \frac{\partial^2 \varphi}{\partial x_i \partial x_j} \\ &\quad + \sum_{i=1}^d b_i(x, a) \frac{\partial \varphi}{\partial x_i} \end{aligned} \quad (94)$$

which is the infinitesimal generator of X^α when α is a constant equal to $a \in A$.

A typical form of finite horizon optimal control problems in finance consists in computing

$$u(t, x) = \inf_{\alpha \text{ admissible}} \mathbb{E} \left[e^{-rT} g(X_T^\alpha) + \int_t^T e^{-rt} f(X_t^\alpha, \alpha_t) dt \mid X_t^\alpha = x \right] \quad (95)$$

where f and g are continuous and bounded functions and to find an optimal control α^* that realizes the minimum. Moreover, it is desirable to find a *Markov* optimal control, that is, an optimal control having the form $\alpha_t^* = \psi(t, X_t)$. Indeed, in this case, the controlled diffusion X^{α^*} is a Markov process.

In the case of nondegenerate diffusion coefficient, we have the following link between the optimal control problems and a semilinear PDEs.

Theorem 8 *Under the additional assumption that σ is uniformly elliptic, u is the unique bounded classical solution of the Hamilton–Jacobi–Bellman (HJB) equation*

$$\begin{cases} \frac{\partial u}{\partial t}(t, x) + \inf_{a \in A} \{L^a u(t, x) + f(x, a)\} - ru(t, x) = 0, & (t, x) \in (0, T) \times \mathbb{R}^d \\ u(T, x) = g(x), & x \in \mathbb{R}^d \end{cases} \quad (96)$$

Furthermore, a Markov control $\alpha_t^* = \psi(t, X_t)$ is optimal for a fixed initial condition x and initial time $t = 0$ if and only if

$$\begin{aligned} L^{\psi(t, x)} u(t, x) + f(x, \psi(t, x)) \\ = \inf_{a \in A} \{L^a u(t, x) + f(x, a)\} \end{aligned} \quad (97)$$

for almost every $(t, x) \in [0, T] \times \mathbb{R}^d$.

This is Theorem III.2.3 of [3] restricted to the case of precise controls (see later).

Here again, the essential argument of the proof can be easily (at least formally) written: consider any admissible control α and the corresponding controlled diffusion X^α with initial condition $X_0 = x$. By Itô's formula applied to $e^{-rt} v(t, X_t^\alpha)$, where v is the solution of equation (96),

$$\mathbb{E}[e^{-rT} v(T, X_T^\alpha)] = v(0, x) + \int_0^T e^{-rt} \left[\frac{\partial v}{\partial t}(t, X_t^\alpha) + L^{\alpha_t} v(t, X_t^\alpha) + r v(t, X_t^\alpha) \right] ds$$

$$\times \mathbb{E} \left[\frac{\partial v}{\partial t}(t, X_t^\alpha) + L^{\alpha_t} v(t, X_t^\alpha) + r v(t, X_t^\alpha) \right] ds \quad (98)$$

Therefore, by equation (96),

$$\begin{aligned} v(0, x) \\ \leq \mathbb{E} \left[e^{-rT} g(X_T^\alpha) + \int_0^T e^{-rt} f(X_t^\alpha, \alpha_t) dt \mid X_0^\alpha = x \right] \end{aligned} \quad (99)$$

for any admissible control α . Now, for the Markov control α^* defined in Theorem 8, all the inequalities in the previous computation are equalities. Hence $v = u$.

The cases where σ is not uniformly elliptic or where σ is also dependent on the current control α_t are much more difficult. In both cases, it is necessary to enlarge the set of admissible control by considering *relaxed controls*, that is, controls that belong to the set $\mathcal{P}(A)$ of probability measures on A . For such a control α , the terms $b(x, \alpha_t)$ and

$f(x, \alpha_t)$ in equations (93) and (95) are replaced by $\int b(x, a) \alpha_t(da)$ and $\int f(x, a) \alpha_t(da)$, respectively. The admissible controls of the original problem correspond to relaxed controls that are Dirac masses at each time. These are called *precise controls*.

The value $\tilde{u}$ of this new problem is defined as in equation (95), but the infimum is taken over all progressively measurable processes α taking values in $\mathcal{P}(A)$. It is possible to prove under general assumptions that both problems give the same value: $\tilde{u} = u$ (cf. [3, Cor.I.2.1] or [8, Th.2.3]).

In these cases, one usually cannot prove the existence of a classical solution of equation (96). The weaker notion of **viscosity solution** is generally the correct one. In all the cases treated in the literature, $u = \tilde{u}$ solves the same HJB equation as in Theorem 8, except that the infimum is taken over $\mathcal{P}(A)$ instead of A (cf. [3, Th.IV.2.2] for the case without control on σ). However, it is not trivial at all in general to obtain a result on precise controls from the result on relaxed controls. This is due to the fact that

usually no result is available on the existence and the characterization of a Markov-relaxed optimal control. The only examples where it has been done require restrictive assumptions (cf. [8, Cor.6.8]). However, in most of the financial applications, the value function u is the most useful information. In practice, one usually only needs to compute a control that give an expected value arbitrarily close to the optimal one.

Optimal Stopping Problems

Optimal stopping problems arise in finance, for example, for the **American options** pricing (when

$$\begin{cases} \max \left\{ u(t, x) - g(t, x); -\frac{\partial u}{\partial t}(t, x) - L_t u(t, x) + ru(t, x) - f(t, x) \right\} = 0, & (t, x) \in (0, T) \times \mathbb{R}^d \\ u(T, x) = g(T, x) & x \in \mathbb{R}^d \end{cases} \quad (103)$$

to sell a claim, an asset?) or in production models (when to extract or product a good? when to stop production?).

Let us consider a Feller diffusion X in $\mathbb{R}^d$ solution to the SDE

$$dX_t = b(t, X_t) dt + \sigma(t, X_t) dB_t \quad (100)$$

where B is a standard d -dimensional Brownian motion. As in equation (28), let $(L_t)_{t \geq 0}$ denote its family of time-inhomogeneous infinitesimal generators. Denote by $\Pi(t, T)$ the set of stopping times valued in $[t, T]$.

A typical form of optimal stopping problems consists in computing

$$\begin{aligned} u(t, x) = & \inf_{\tau \in \Pi(t, T)} \mathbb{E} \left[e^{-r(\tau-t)} g(\tau, X_\tau) \right. \\ & \left. + \int_t^\tau e^{-r(s-t)} f(s, X_s) ds \mid X_t = x \right] \end{aligned} \quad (101)$$

and to characterize an optimal stopping time.

Assume that $b(t, x)$ is bounded and continuously differentiable with bounded derivatives and that $\sigma(t, x)$ is bounded, continuously differentiable with respect to t and twice continuously differentiable with respect to x with bounded derivatives. Assume also that σ is uniformly elliptic. Finally,

assume that $g(t, x)$ is differentiable with respect to t and twice differentiable with respect to x and that

$$|f(t, x)| + \left| \frac{\partial g}{\partial t}(t, x) \right| + \sum_{i=1}^d \left| \frac{\partial g}{\partial x_i}(t, x) \right| \leq C e^{\mu|x|} \quad (102)$$

for positive constants C and μ .

Theorem 9 ([2], Sec.III.4.9). *Under the previous assumptions, $u(t, x)$ admits first-order derivatives with respect to t and second-order derivatives with respect to x that are L^p for all $1 \leq p < \infty$. Moreover, u is the solution of the variational inequality*

The proof of this result is based on a similar (formal) justification as the one we gave for equation (90). We refer to [12] for a similar result under weaker assumptions more suited to financial models when $f = 0$ (this is in particular the case for American options).

In some cases (typically with $f = 0$, see [11]), it can be shown that the infimum in equation (101) is attained for the stopping time

$$\tau^* = \inf \{ t \leq s \leq T : u(s, X_s^{t,x}) = g(s, X_s^{t,x}) \} \quad (104)$$

where $X^{t,x}$ is the solution of the SDE (100) with initial condition $X_t^{t,x} = x$.

Generalizations and Extensions

An optimal control problem can also be solved through the optimization of a family of BSDEs related to the laws of the controlled diffusions. On this question, we refer to [19] and **BSDEs**.

In this section, we considered only very specific optimal control problems. Other important families of optimal control problems are given by impulse control problems, where the control may induce a jump of the underlying stochastic process, or ergodic control problems, where the goal is to optimize a quantity related to the stationary behavior of the controlled

diffusion. Impulse control has applications, for example, in stock or resource management problems. In the finite horizon case, when the underlying asset follows a model with stochastic or elastic volatility or when the market is incomplete, other optimal control problems can be considered, such as characterizing the superhedging cost, or minimizing some risk measure. Various constraints can be included in the optimal control problem, such as maximizing the expectation of an utility with the constraint that this utility has a fixed volatility, or minimizing the volatility for a fixed expected utility. One can also impose Gamma constraints on the control. Another important extension of optimal control problems arises when one wants to solve numerically an HJB equation. Usual discretization methods require to restrict to a bounded domain and to fix artificial boundary conditions. The numerical solution can be interpreted as the solution of an optimal control problem in a bounded domain. In this situation, a crucial question is to quantify the impact on the discretized solution of an error on the artificial boundary condition (which usually cannot be computed exactly).

On Numerical Methods

The Feynman–Kac formula for linear PDEs allows one to use **Monte Carlo methods** to compute the solution of the PDE. They are especially useful when the solution of the PDE has to be computed at a small number of points, or when dimension is large (typically larger or equal to 4), since they provide a rate of convergence independent of the dimension.

Concerning quasi- or semilinear PDEs and some optimal control problems (e.g., American put options in the section “Quasi- and Semilinear PDEs and BSDEs”), interpretations in terms of BSDEs provide indirect Monte Carlo methods of numerical computation (see [1] for **Bermudan options** or [4, 6] for general BSDEs schemes). These methods have the advantage that they do not require to consider artificial boundary conditions. However, their speed of convergence to the exact solution is still largely unknown, and could depend on the dimension of the problem.

For high dimensional HJB equations, the analytical discretization methods lead to important numerical problems. First, these methods need to solve an optimization problem at each node of the discretization grid, which can be very costly in high dimension

or difficult depending on the particular constraints imposed on the control. Moreover, these methods require to localize the problem, that is, to solve the problem in a bounded domain with artificial boundary conditions, which are usually difficult to compute precisely. This localization problem can be solved by computing the artificial boundary condition with a Monte Carlo method based on BSDEs. However, the error analysis of this method is based on the probabilistic interpretation of HJB equations in bounded domains, which is a difficult problem in general.

End Notes

^aA Markov semigroup family $(P_t, t \geq 0)$ on $\mathbb{R}^d$ is a family of bounded linear operators of norm 1 on the set of bounded measurable functions on $\mathbb{R}^d$ equipped with the L^∞ norm, which satisfies equation (8).

^bThis is not the most general definition of Feller semigroups (see [21, Def.III.6.5]). In our context, because we only introduce analytical objects from stochastic processes, the semigroup (P_t) is naturally defined on the set of bounded measurable functions.

^cThe strong continuity of a semigroup is usually defined as $\|P_t f - f\| \rightarrow 0$ as $t \rightarrow 0$ for all $f \in C_0(\mathbb{R}^d)$. However, in the case of Feller semigroups, this is equivalent to the weaker formulation (10) (see [21, Lemma III.6.7]).

References

- [1] Bally, V. & Pagès, G. (2003). Error analysis of the optimal quantization algorithm for obstacle problems, *Stochastic Processes and their Applications* **106**(1), 1–40.
- [2] Bensoussan, A. & Lions, J.-L. (1982). *Applications of Variational Inequalities in Stochastic Control*, *Studies in Mathematics and its Applications*, North-Holland Publishing, Amsterdam, Vol. 12 (Translated from the French).
- [3] Borkar, V.S. (1989). *Optimal Control of Diffusion Processes*, *Pitman Research Notes in Mathematics Series*, Longman Scientific & Technical, Harlow, Vol. 203.
- [4] Bouchard, B. & Touzi, N. (2004). Discrete-time approximation and Monte-Carlo simulation of backward stochastic differential equations, *Stochastic Processes and their Applications* **111**(2), 175–206.
- [5] Çinlar, E. & Jacod, J. (1981). Representation of semimartingale Markov processes in terms of Wiener processes and Poisson random measures, in *Seminar on Stochastic Processes, 1981 (Evanston, Ill., 1981)*, *Progress in Probability and Statistics*, Birkhäuser, Boston, Vol. 1, pp. 159–242.
- [6] Delarue, F. & Menozzi, S. (2006). A forward-backward stochastic algorithm for quasi-linear PDEs, *Annals of Applied Probability* **16**(1), 140–184.

- [7] El Karoui, N., Kapoudjian, C., Pardoux, E., Peng, S. & Quenez, M.C. (1997). Reflected solutions of backward SDE's, and related obstacle problems for PDE's, *Annals of Probability* **25**(2), 702–737.
- [8] El Karoui, N., Nguyen, D. & Huu Jeanblanc-Picqué, M. (1987). Compactification methods in the control of degenerate diffusions: existence of an optimal control, *Stochastics* **20**(3), 169–219.
- [9] Ethier, S.N. & Kurtz, T.G. (1986). *Markov Processes: Characterization and Convergence*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, New York.
- [10] Friedman, A. (1975). *Stochastic Differential Equations and Applications*, Vol. 1, Probability and Mathematical Statistics, Academic Press [Harcourt Brace Jovanovich Publishers], New York, Vol. 28.
- [11] Jacka, S.D. (1993). Local times, optimal stopping and semimartingales, *Annals of Applied Probability* **21**(1), 329–339.
- [12] Jalliet, P., Lamberton, D. & Lapeyre, B. (1990). Variational inequalities and the pricing of American options, *Acta Applicandae Mathematicae* **21**(3), 263–289.
- [13] Karatzas, I. & Shreve, S.E. (1988). *Brownian Motion and Stochastic Calculus*, Graduate Texts in Mathematics, Springer-Verlag, New York, Vol. 113.
- [14] Lamberton, D. & Lapeyre, B. (1996). *Introduction to Stochastic Calculus Applied to Finance*, Chapman & Hall, London (Translated from the 1991 French original by Nicolas Rabeau and François Mantion).
- [15] Ma, J., Protter, P. & Yong, J.M. (1994). Solving forward-backward stochastic differential equations explicitly—a four step scheme, *Probability Theory and Related Fields* **98**(3), 339–359.
- [16] Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*, 6th Edition, Universitext, Springer-Verlag, Berlin.
- [17] Pardoux, E. (1998). Backward stochastic differential equations and viscosity solutions of systems of semilinear parabolic and elliptic PDEs of second order, in *Stochastic Analysis and Related Topics: The Geilo Workshop*, B.O.L. Decreusefond, J. Gjerd & A. Ustunel, eds, Birkhäuser, pp. 79–127.
- [18] Protter, P. (2001). A partial introduction to financial asset pricing theory, *Stochastic Processes and Their Applications* **91**(2), 169–203.
- [19] Quenez, M.C. (1997). Stochastic control and BSDEs, in *Backward Stochastic Differential Equations (Paris, 1995–1996)*, Pitman Research Notes in Mathematics Series, Longman, Harlow, Vol. 364, pp. 83–99.
- [20] Rogers, L.C.G. & Shi, Z. (1995). The value of an Asian option, *Journal of Applied Probability* **32**(4), 1077–1088.
- [21] Rogers, L.C.G. & Williams, D. (1994). *Diffusions, Markov Processes, and Martingales*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, 2nd Edition, John Wiley & Sons, Chichester, Vol. 1.
- [22] Talay, D. & Zheng, Z. (2003). Quantiles of the Euler scheme for diffusion processes and financial applications, *Mathematical Finance* **13**(1) 187–199, Conference on Applications of Malliavin Calculus in Finance (Rocquencourt, 2001).

MIREILLE BOSSY & NICOLAS CHAMPAGNAT

Doob–Meyer Decomposition

Submartingales are processes that grow on average. Subject to some condition of uniform integrability, they can be written uniquely as the sum of a martingale and a predictable increasing process. This result is known as the *Doob–Meyer decomposition*.

Consider a filtered probability space $(\Omega, \mathcal{F}, \mathbf{F}, P)$. It consists of a probability space $(\Omega, \mathcal{F}, P)$ and a *filtration* $\mathbf{F} = (\mathcal{F}_t)_{t \geq 0}$, that is, an increasing family of sub- σ -fields of $\mathcal{F}$. The σ -field $\mathcal{F}_t$ stands for the information available at time t . A random event A belongs to $\mathcal{F}_t$, if we know at time t , whether it will take place or not, that is, A does not depend on randomness in the future. For technical reasons, one typically assumes right continuity, that is, $\mathcal{F}_t = \bigcap_{s > t} \mathcal{F}_s$.

A *martingale* (see **Martingales**) (respectively *submartingale*, *supermartingale*) is an adapted, integrable process $(X_t)_{t \in \mathbb{R}_+}$ satisfying

$$E(X_t | \mathcal{F}_s) = X_s \quad (1)$$

(respectively $\geq X_s$, $\leq X_s$) for $s \leq t$. Moreover, we require these processes to be a.s. càdlàg, that is, right-continuous with left-hand limits. *Adaptedness* means that X_t is $\mathcal{F}_t$ -measurable, that is, the random value X_t is known at the latest at time t . *Integrability* $E(|X_t|) < \infty$ is needed for the conditional expectation to be defined. The crucial martingale equality (1) means that the best prediction of future values of X is the current value, that is, X will stay on the current level on average. In other words, it does not exhibit any positive or negative trend. If X denotes the price of a security, this asset does not produce profits or losses on average. Submartingales, on the other hand, grow on average. Put differently, they show an upward trend compared to a martingale. This loose statement is made precise in terms of the Doob–Meyer decomposition.

As a starting point, consider a discrete-time process $X = (X_t)_{t=0,1,2,\dots}$. In discrete time, a process X is called *predictable* if X_t is $\mathcal{F}_{t-1}$ -measurable for $t = 1, 2, \dots$. This means that the value X_t is known already one period ahead. The *Doob decomposition* states that any submartingale X can be written uniquely as

$$X_t = M_t + A_t \quad (2)$$

with a martingale M and an increasing predictable process A satisfying $A_0 = 0$. While the intuitive meaning of M and A may not be obvious, the corresponding decomposition of the increments $\Delta X_t := X_t - X_{t-1}$ is easier to understand.

$$\Delta X_t = \Delta M_t + \Delta A_t \quad (3)$$

can be interpreted in the sense that the increment ΔX_t consists of a predictable trend ΔA_t and a random deviation ΔM_t from that trend. Its implication $\Delta A_t = E(\Delta X_t | \mathcal{F}_{t-1})$ means that ΔA_t is the best prediction of ΔX_t in a mean-square sense and based on the information up to time $t - 1$.

The natural decomposition (3) does not make sense for continuous time processes but an analog of equation (2) still exists. To this end, the notion of predictability must be extended to continuous time. A process $X = (X_t)_{t \in \mathbb{R}_+}$ is called *predictable* if—viewed as a mapping on $\Omega \times \mathbb{R}_+$ —it is measurable with respect to the σ -field generated by all adapted, left-continuous processes. Intuitively, this rather abstract definition means that X_t is known slightly ahead of time t . In view of the discrete-time case, it may seem more natural to require that X_t be $\mathcal{F}_{t-}$ -measurable, where $\mathcal{F}_{t-}$ stands for the smallest sub- σ -field containing all $\mathcal{F}_s$, $s < t$. However, this slightly weaker condition turns out to be too weak for the general theory.

In order for a decomposition (2) into a martingale M and a predictable increasing process A to exist, one must assume some uniform integrability of X . The process X must belong to the so-called *class (D)*, which amounts to a rather technical condition implying $\sup_{t \geq 0} E(|X_t|) \leq \infty$ but being itself implied by $E(\sup_{t \geq 0} |\bar{X}_t|) \leq \infty$. For its precise definition, we need to introduce the concept of a stopping time, which is not only an indispensable tool for the general theory of stochastic processes but also interesting for applications, for example, in mathematical finance. A $[0, \infty]$ -valued random variable T is called *stopping time* if $\{T \leq t\} \in \mathcal{F}_t$ for any $t \geq 0$. Intuitively, T stands for a random time, which is generally not known in advance but at the latest once it has happened (e.g., the time of a phone call, the first time when a stock hits 100, the time when you crash your car into a tree). In financial applications, it appears, for example, as the exercise time of an American option.

Stopping times can be classified by their degree of suddenness. *Predictable* stopping times do not come

entirely as a surprise because one anticipates them. Formally, a stopping time T is called *predictable* if it allows for an *announcing sequence*, that is, for a sequence $(T_n)_{n \in \mathbb{N}}$ of stopping times satisfying $T_0 < T_1 < T_2 < \dots$ on $\{T > 0\}$ and $T_n \rightarrow T$ as $n \rightarrow \infty$. This is the case for a continuous stock price hitting 100 or for the car crashing into a tree, because you can literally see the level 100 or the tree coming increasingly closer. Phone calls, strikes of lightning, or jumps of Lévy process, on the other hand, are of an entirely different kind because they happen completely out of the blue. Such stopping times T are called *totally inaccessible*, which formally means that $P(S = T < \infty) = 0$ for all predictable stopping times S .

Coming back to our original theme, a process X is said to be of *class (D)* if the set $\{X_T : T \text{ finite stopping time}\}$ is uniformly integrable, which in turn means that

$$\lim_{c \rightarrow \infty} \sup_{T \text{ finite stopping time}} E(1_{\{|X_T| > c\}} |X_T|) = 0$$

The *Doob–Meyer decomposition* can now be stated as follows:

Theorem 1 *Any submartingale X of class (D) allows for a unique decomposition*

$$X_t = M_t + A_t \quad (4)$$

with a martingale M and some predictable increasing process A satisfying $A_0 = 0$.

The martingale M turns out to be of class (D) as well, which implies that it converges a.s. and in L^1 to some terminal random variable M_∞ . Since the whole martingale M can be recovered from its limit via $M_t = E(M_\infty | \mathcal{F}_t)$, one can formally identify such *uniformly integrable martingales* with their limit.

In the case of an Itô process

$$dX_t = H_t dW_t + K_t dt \quad (5)$$

the Doob–Meyer decomposition is easily obtained. Indeed, we have $M_t = X_0 + \int_0^t H_s dW_s$ and $A_t = \int_0^t K_s ds$. However, a general Itô process need not, of course, be a submartingale. However, equation (5) suggests that a similar decomposition exists for more general processes. This is indeed the case. For a generalization covering all Itô processes we relax both the martingale property of M and the

monotonicity of A . In general, A is only required to be of finite variation, that is, the difference of two increasing processes. In the Itô process example, these are $A_t^{(+)} = \int_0^t \max(K_s, 0) ds$ and $A_t^{(-)} = \int_0^t \max(-K_s, 0) ds$. Put differently, the trend may change its direction every now and then.

To cover all Itô processes, one must also allow for local martingales rather than martingales. M is said to be a *local martingale* if there exists a sequence of stopping times $(T_n)_{n \in \mathbb{N}}$, which increases to ∞ almost surely such that M^{T_n} is a martingale for any n . Here, the *stopped process* M^{T_n} is defined as $M_t^{T_n} := M_{\min(T_n, t)}$, that is, it stays constant after time T_n (as e.g., your wealth does if you sell an asset at T_n). This rather technical concept appears naturally in the general theory of stochastic processes. For example, stochastic integrals $M_t = \int_0^t H_s dN_s$ relative to martingales N generally fail to be martingales but are typically local martingales or a little less, namely, *σ -martingales*.

A local martingale is a uniformly integrable martingale, if and only if it is of class (D). Nevertheless, one should be careful with thinking that local martingales behave basically as martingales up to some integrability. For example, there exist local martingales $M_t = \int_0^t H_s dW_s$ with $M_0 = 0$ and $M_1 = 1$ a.s. and such that $E(|M_t|) < \infty$, $t \geq 0$. Even though such a process has no trend in a local sense, it behaves entirely differently from a martingale on a global scale. The difference between local martingales and martingales leads to many technical problems in mathematical finance. For example, the previous example may be interpreted in the sense that dynamic investment in a perfectly reasonable martingale may lead to arbitrage unless the set of trading strategies is restricted to some admissible subset.

Let us come back to generalizing the Doob–Meyer decomposition. Without class (D) it reads as follows:

Theorem 2 *Any submartingale X allows for a unique decomposition (4) with a local martingale M and some predictable increasing process A satisfying $A_0 = 0$.*

For a considerably larger class of processes X , there exists a *canonical decomposition* (4) with a local martingale M and some predictable process A of finite variation, which starts in 0. These processes are called *special semimartingales* and they play a key role in stochastic calculus. The slightly larger

class of *semimartingales* is obtained, if A is only required to be adapted rather than predictable. This class is, in some sense, the largest one that allows for the definition of a stochastic integral $\int_0^t H_s dX_s$ satisfying a mild continuity property. In the general **semimartingale** case, decomposition (4) should not be called canonical because it is not unique. Moreover, A should not be regarded as a trend unless it is predictable. On the other hand, if the jumps of a semimartingale X are sufficiently integrable (e.g., bounded), then X is special and hence allows for a canonical decomposition resembling the Doob–Meyer decomposition of a submartingale.

Further Reading

Protter, P. (2004). *Stochastic Integration and Differential Equations*, 2nd Edition, Version 2.1, Springer, Berlin.

Related Articles

American Options; Martingales; Semimartingale.

JAN KALLSEN

Forward–Backward Stochastic Differential Equations (SDEs)

A *forward–backward stochastic differential equation* (FBSDE) is a system of two Itô-type stochastic differential equations (SDEs) over $[0, T]$ taking the following form:

$$\begin{cases} dX_t = b(t, \omega, X_t, Y_t, Z_t)dt \\ \quad + \sigma(t, \omega, X_t, Y_t, Z_t)dW_t, & X_0 = x; \\ dY_t = -f(t, \omega, X_t, Y_t, Z_t)dt + Z_t dW_t, \\ Y_T = g(\omega, X_T) \end{cases} \quad (1)$$

Here W is a standard Brownian motion defined on a complete probability space $(\Omega, \mathcal{F}, P)$, and $\mathbf{F} \triangleq \{\mathcal{F}_t\}_{0 \leq t \leq T}$ is the filtration generated by W augmented with all the null sets. The coefficients b, σ, f, g are progressively measurable; b, σ, f are $\mathbf{F}$ -adapted for fixed (x, y, z) ; and g is $\mathcal{F}_T$ -measurable for fixed x . The first equation is forward because the initial value X_0 is given, while the second one is backward because the terminal condition Y_T is given. The solution to FBSDE (1) consists of three $\mathbf{F}$ -adapted processes (X, Y, Z) that satisfy equation (1) for any t , P almost surely (a.s.), and

$$\begin{aligned} \|(X, Y, Z)\|^2 &\triangleq E \left\{ \sup_{0 \leq t \leq T} [|X_t|^2 + |Y_t|^2] \right. \\ &\quad \left. + \int_0^T |Z_t|^2 dt \right\} < \infty \end{aligned} \quad (2)$$

BSDEs can be traced back to the 1973 paper by Bismut [7], where a linear BSDE is introduced as an adjoint equation for a stochastic control problem. Bensoussan [6] proved the well posedness of general linear BSDEs by using the martingale representation theorem. The general theory of nonlinear BSDEs, however, originated from the seminal work of Pardoux and Peng [37]. Their motivation was to study the general Pontryagin-type maximum principle for stochastic optimal controls; see, for example, [40]. Independent of the development of this theory, Duffie and Epstein [19, 20] proposed the concept of stochastic recursive utility, and it turns out that

BSDEs provide exactly the right mathematical tool for it.

Peng [41], and Pardoux and Peng [38], then studied *decoupled FBSDEs*, that is, b and σ do not depend on (y, z) . They discovered the deep relation between Markovian FBSDEs (i.e., FBSDEs with deterministic coefficients) and PDEs, *via* the so called *nonlinear Feynman–Kac formula*. Soon after that, people found that such FBSDEs had very natural applications in option pricing theory, and thus extended the Black–Scholes formula to a much more general framework. In particular, the solution triplet (X, Y, Z) can be interpreted as the underlying asset price, the option price, and the hedging portfolio, respectively. El Karoui *et al.* [22] further introduced *reflected BSDEs*, which are appropriate for pricing American options, again, in a general framework. See a survey paper [24] and the section Applications for such applications.

The theory of coupled FBSDEs was originally motivated by Black’s consol rate conjecture. Antonelli [1] proved the first well-posedness result, when the time duration T is small. For arbitrary T , there are three typical approaches, each with its limit. The most famous one is the *four-step scheme*, proposed by Ma *et al.* [34]. On the basis of this scheme, Duffie *et al.* [21] confirmed Black’s conjecture. The theory has also been applied to various areas, especially in finance and in stochastic control.

There have been numerous publications on the subject. We refer interested readers to the books [23, 35], and the references therein for the general theory and applications.

Decoupled FBSDEs

Since b and σ do not depend on (y, z) , one can first solve the forward SDE and then the backward one. The main idea in [37] to solve BSDEs is to apply the Picard iteration, or equivalently, the contraction mapping theorem.

Theorem 1 ([38]). *Assume that b, σ do not depend on (y, z) ; that b, σ, f, g are uniformly Lipschitz continuous in (x, y, z) , uniformly on (ω, t) ; and that*

$$\begin{aligned} I_0 &\triangleq E \left\{ \int_0^T \left[|b(t, \cdot, 0)|^2 + |\sigma(t, \cdot, 0)|^2 \right. \right. \\ &\quad \left. \left. + |f(t, \cdot, 0, 0, 0)|^2 \right] dt + |g(\cdot, 0)|^2 \right\} < \infty \end{aligned} \quad (3)$$

Then FBSDE (1) admits a unique solution (X, Y, Z) , and there exists a constant C , depending only on T , the dimensions, and the Lipschitz constant, such that $\|(X, Y, Z)\|^2 \leq C[|x_0|^2 + I_0]$.

When $\dim(Y) = 1$, we have the following comparison result for the BSDE. For $i = 1, 2$, assume (b, σ, f_i, g_i) satisfy the assumptions in Theorem 1 and let (X, Y^i, Z^i) denote the corresponding solutions to equation (1). If $f^1 \leq f^2$, $g^1 \leq g^2$, P a.s., for any (t, x, y, z) , then, $Y_t^1 \leq Y_t^2$, $\forall t$, P a.s.; see, for example, [24]. On the basis of this result, Lepeltier and San Martín [31] constructed solutions to BSDEs with non-Lipschitz coefficients. Moreover, Kobylanski [30] and Briand and Hu [10] proved the well posedness of BSDEs whose generator f has quadratic growth in Z . Such BSDEs are quite useful in practice.

When the coefficients are deterministic, the decoupled FBSDE (1) becomes

$$\begin{cases} dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, & X_0 = x; \\ dY_t = -f(t, X_t, Y_t, Z_t)dt + Z_t dW_t, \\ Y_T = g(X_T) \end{cases} \quad (4)$$

In this case, the FBSDE is associated with the following system of parabolic PDEs:

$$\begin{cases} u_t^i + \frac{1}{2} \text{tr} [u_{xx}^i \sigma \sigma^*(t, x)] + u_x^i b(t, x) \\ \quad + f^i(t, x, u, u_x \sigma(t, x)) = 0, \\ i = 1, \dots, m; \\ u(T, x) = g(x) \end{cases} \quad (5)$$

Theorem 2 ([38]). Assume b, σ, f, g satisfy all the conditions in Theorem 1.

- (i) If PDE (5) has a classical solution $u \in C^{1,2}([0, T] \times \mathbb{R}^n)$, then

$$Y_t = u(t, X_t), \quad Z_t = u_x \sigma(t, X_t) \quad (6)$$

- (ii) In general, define

$$u(t, x) \triangleq E\{Y_t | X_t = x\} \quad (7)$$

Then u is deterministic and $Y_t = u(t, X_t)$. Moreover, when $m = 1$, u is the unique viscosity solution to the PDE (5).

In this case, X is a Markov process; then by equation (6) the solution (X, Y, Z) is Markovian. For this

reason we call equation (4) a *Markovian FBSDE*. We note that in the Black–Scholes model, as we see in the section Applications, the PDE (5) is linear and one can solve for u explicitly. Then equation (6) in fact gives us the well known Black–Scholes formula. Moreover, the hedging portfolio $Z_t \sigma^{-1}(t, X_t)$ is the sensitivity of the option price Y_t with respect to the underlying asset price X_t . This is exactly the idea of the Δ -hedging. On the other hand, when f is linear in (y, z) , equation (7) actually is equivalent to the Feynman–Kac formula. In general, when $m = 1$, equation (7) provides a probabilistic representation for the viscosity solution to the PDE (5), and thus is called a *nonlinear Feynman–Kac formula*. Such a type of representation formula is also available for u_x [36].

The link between FBSDEs and PDEs opens the door to efficient Monte Carlo methods for high-dimensional PDEs and FBSDEs, and thus also for many financial problems. This approach can effectively overcome the *curse of dimensionality*; see, for example, [3–5, 8, 27, 45], and [12]. There are also some numerical algorithms for non-Markovian BSDEs and coupled FBSDEs; see, for example, [2, 9, 18, 33], and [17].

Coupled FBSDEs

The theory of coupled FBSDEs is much more complex and is far from complete. There are mainly three approaches for its well posedness, each with its limit. Since the precise statements of the results require complicated notation and technical conditions, we refer readers to the original research papers and focus only on the main ideas here.

Method 1 Contraction Mapping This method works very well for BSDEs and decoupled FBSDEs. However, to ensure the constructed mapping is a contraction one, for coupled FBSDEs one has to assume some stronger conditions. The first well-posedness result was by Antonelli [1], which has been extended further by Pardoux and Tang [39]. Roughly speaking, besides the standard Lipschitz conditions, FBSDE (1) is well posed in one of the following three cases: (i) T is small and either σ_z or g_x is small; (ii) X is weakly coupled into the BSDE (i.e., g_x and f_x are small) or (Y, Z) are weakly coupled into the FSDE (i.e., $b_y, b_z, \sigma_y, \sigma_z$ are small); or (iii) b is deeply decreasing in x (i.e., $[b(\cdot, x_1, \cdot) -$

$b(\cdot, x_2, \cdot)[x_1 - x_2] \leq -C|x_1 - x_2|^2$ for some large C) or f is deeply decreasing in y . Antonelli [1] also provides a counterexample to show that, under Lipschitz conditions only, equation (1) may have no solution.

Method 2 Four-step Scheme This is the most popular method for coupled FBSDEs with deterministic coefficients, proposed by Ma *et al.* [34]. The main idea is to use the close relationship between Markovian FBSDEs and PDEs, in the spirit of Theorem 2. Step 1 in [34] deals with the dependence of σ on z , which works only in very limited cases. The more interesting case is that σ does not depend on z . Then the other three steps read as follows:

Step 2. Solve the following PDE with $u(T, x) = g(x)$: for $i = 1, \dots, m$,

$$\begin{aligned} u_t^i + \frac{1}{2} \text{tr}[u_{xx}^i \sigma \sigma^*(t, x, u)] \\ + u_x^i b(t, x, u, u_x \sigma(t, x, u)) \\ + f^i(t, x, u, u_x \sigma(t, x, u)) = 0 \end{aligned} \quad (8)$$

Step 3. Solve the following FSDE:

$$\begin{aligned} X_t = x + \int_0^t b(s, X_s, u(s, X_s), u_x(s, X_s)) \\ \times \sigma(s, X_s, u(s, X_s)) ds \\ + \int_0^t \sigma(s, X_s, u(s, X_s)) dW_s \end{aligned} \quad (9)$$

Step 4. Set

$$\begin{aligned} Y_t &\triangleq u(t, X_t), Z_t \triangleq u_x(t, X_t) \\ &\times \sigma(t, X_t, u(t, X_t)) \end{aligned} \quad (10)$$

The main result in [34] is essentially the following theorem.

Theorem 3 Assume (i) b, σ, f, g are deterministic, uniformly Lipschitz continuous in (x, y, z) , and σ does not depend on z ; (ii) PDE (8) has a classical solution u with bounded derivatives. Then FBSDE (1) has a unique solution.

This result has been improved by Delarue [16] and Zhang [46], by weakening the requirement on u to only uniform Lipschitz continuity in x . Delarue [16]

assumes some sufficient conditions on the deterministic coefficients to ensure such Lipschitz continuity. In particular, one key condition is that the coefficient σ be uniformly nondegenerate. Zhang [46] allows the coefficients to be random and σ to be degenerate, but assumes all processes are one-dimensional along with some special compatibility condition on the coefficients, so that a similarly defined random field $u(t, \omega, x)$ is uniformly Lipschitz continuous in x .

Method 3 Method of Continuation The idea is that, if an FBSDE is well-posed, then a new FBSDE with slightly modified coefficients is also well-posed. The problem is then to find sufficient conditions so that this modification procedure can go arbitrarily long. This method allows the coefficients to be random and σ to be degenerate. However, it requires some monotonicity conditions; see for example, [29, 42], and [43]. For example, [29] assumes that, for some constant $\beta > 0$ and for any $\theta_i = (x_i, y_i, z_i)$, $i = 1, 2$,

$$\begin{aligned} [b(t, \omega, \theta_1) - b(t, \omega, \theta_2)][y_1 - y_2] \\ + [\sigma(t, \omega, \theta_1) - \sigma(t, \omega, \theta_2)][z_1 - z_2] \\ - [f(t, \omega, \theta_1) - f(t, \omega, \theta_2)][x_1 - x_2] \\ \geq \beta[|x_1 - x_2|^2 + |y_1 - y_2|^2 \\ + |z_1 - z_2|^2] \end{aligned} \quad (11)$$

$$\begin{aligned} [g(\omega, x_1) - g(\omega, x_2)][x_1 - x_2] \\ \leq -\beta|x_1 - x_2|^2 \end{aligned} \quad (12)$$

Applications

We now present some typical applications of FBSDEs.

1. Option pricing and hedging

Let us consider the standard Black–Scholes model. The financial market consists of two underlying assets, a riskless one B_t and a risky one S_t . Assume an investor holds a portfolio $(x_t, \pi_t)_{0 \leq t \leq T}$, with its wealth $V_t \triangleq x_t B_t + \pi_t S_t$. We say the portfolio is self-financing if $dV_t = x_t dB_t + \pi_t dS_t$; that is, the change

of the wealth is solely due to the change of the underlying assets' prices.

Now consider a European call option with terminal payoff $g(S_T) \triangleq (S_T - K)^+$. We say a self-financing portfolio (x_t, π_t) is a perfect hedge of the option if $V_T = g(S_T)$. Under a no-arbitrage assumption, V_t is the unique fair option price at t . Let r denote the interest rate of B , μ the appreciation rate, and σ the volatility of S . Then (S, V, π) satisfy the following linear FBSDE:

$$\begin{cases} dS_t = S_t[\mu dt + \sigma dW_t], & S_0 = s_0; \\ dV_t = [r(V_t - \pi_t S_t) + \mu \pi_t S_t]dt \\ \quad + \pi_t S_t \sigma dW_t, & V_T = g(S_T) \end{cases} \quad (13)$$

If the borrowing interest rate R is greater than the lending interesting rate r , then the drift term of dV_t becomes $r(V_t - \pi_t S_t)^+ - R(V_t - \pi_t S_t)^- + \mu \pi_t S_t$, and thus the BSDE becomes nonlinear. The coupled FBSDE gives a nice framework for the large investor problem, where the investment may affect the value of S_t . Assume $dS_t = \mu(t, S_t, V_t, \pi_t)dt + \sigma(t, S_t, V_t, \pi_t)dW_t$. Then the system becomes coupled. We refer to [24] and [15] for more detailed exposure.

2. American option and reflected FBSDEs

Consider an American option with generator f , terminal payoff function g , and early exercise payoff L_t . Let X denote the underlying asset price, Y the option price, and $Z\sigma^{-1}$ the hedging portfolio. Then the American option solves the following reflected FBSDE with an extra component K , which is continuous and increasing, with $K_0 = 0$:

$$\begin{cases} dX_t = b(t, \omega, X_t)dt + \sigma(t, \omega, X_t)dW_t, \\ \quad X_0 = x_0; \\ dY_t = -f(t, \omega, X_t, Y_t, Z_t)dt \\ \quad + Z_t dW_t - dK_t, \quad Y_T = g(\omega, X_T); \\ \quad Y_t \geq L_t; \quad [Y_t - L_t]dK_t = 0 \end{cases} \quad (14)$$

Here $K_T - K_t$ can be interpreted as the time value of the option. Moreover, the optimal exercise time is $\tau \triangleq \inf\{t \geq 0 : Y_t = L_t\} \wedge T$. See [22] for more details.

In the Markovian case with $L_t = h(t, X_t)$, the RFBSDE (14) is associated with the following obstacle problem of PDE with $u(T, x) = g(x)$, in the spirit

of Theorem 2:

$$\min \left(u - h(t, x), -u_t - \frac{1}{2} \text{tr}(u_{xx} \sigma \sigma^*(t, x)) - u_x b(t, x) - f(t, x, u, u_x \sigma) \right) = 0 \quad (15)$$

3. Some further extensions

The previous two models consider complete markets. El Karoui and Quenez [26] studied superhedging problems in incomplete markets. They have shown that the superhedging price of a contingent claim is the increasing limit of solutions of a sequence of BSDEs. Cvitanic *et al.* [14] also studied superhedging problems, but in the case that there is a constraint on the portfolio part Z . It turns out that the superhedging price is the minimum solution to an FBSDE with reflection/constraint on Z . Buckdahn and Hu [11] studied a similar problem, but using coupled FBSDE with reflections.

Another application is the zero-sum Dynkin game. The value process Y is the solution to a BSDE with double barriers of Y : $L_t \leq Y_t \leq U_t$. In this case, besides (Y, Z) , the solution consists of two increasing processes K^+, K^- satisfying $[Y_t - L_t]dK_t^+ = [U_t - Y_t]dK_t^- = 0$, and an equilibrium of the game is a pair of stopping times: $\tau_1^* \triangleq \inf\{t : Y_t = L_t\} \wedge T$, $\tau_2^* \triangleq \inf\{t : Y_t = U_t\} \wedge T$. The work in [13, 28] and [32] is along this line.

4. Black's consol rate conjecture

Let r denote the short-rate process and $Y_t \triangleq E_t \left\{ \int_t^\infty \exp \left(- \int_t^s r_l dl \right) ds \right\}$ be the consol price. Assume

$$dr_t = \mu(r_t, Y_t)dt + \alpha(r_t, Y_t)dW_t \quad (16)$$

for some deterministic functions μ, α . The question is whether Y satisfies certain SDEs. Black conjectured that there exists a function A , depending on μ and α , such that $dY_t = [r_t Y_t - 1]dt + A(r_t, Y_t)dW_t$.

The conjecture is confirmed in [21] by using FBSDEs. Assume r is "hidden Markovian," that is, $r_t = h(X_t)$ for some deterministic function h and some Markov process X . Consider the following

FBSDE over infinite horizon:

$$\begin{cases} dX_t = b(X_t, Y_t)dt + \sigma(X_t, Y_t)dW_t, \\ X_0 = x; \\ Y_t = [h(X_t)Y_t - 1]dt + Z_t dW_t, \\ Y_t \text{ is bounded a.s. uniformly in } t \in [0, \infty) \end{cases}$$

The above FBSDE is associated with the following elliptic PDE

$$\frac{1}{2}\sigma^2(x, u)u''(x) + b(x, u)u'(x) - h(x)u(x) + 1 = 0 \quad (17)$$

Assume equation (17) has a bounded classical solution u . Then the Black's conjecture is true with $A(x, y) = \sigma(x, y)u'(x)$.

5. Stochastic control

This is the original motivation to study BSDEs. The classical results in the literature assumed that the diffusion coefficient σ was independent of the control; then the problem was essentially parallel to a deterministic control problem. With the help of BSDEs, one can derive necessary conditions for stochastic control problems in a general framework. To illustrate the idea, we show a very simple example here. We refer readers to [7, 25, 40], and [44] for more details in this aspect.

Assume the state process is

$$X_t = x + \int_0^t \sigma(s, a_s) dW_s \quad (18)$$

where a is the control in some admissible set $\mathcal{A}$. The goal is to find optimal a^* to maximize the utility (or minimize the cost) $J(a) \triangleq E\{g(X_T) + \int_0^T h(t, a_t)dt\}$; that is, we want to find $a^* \in \mathcal{A}$ such that $J(a^*) \geq J(a)$, for all $a \in \mathcal{A}$.

Define an adjoint equation which is a BSDE:

$$Y_t = g'(X_T) - \int_t^T Z_s dW_s \quad (19)$$

Then for any Δa , one can show that

$$\begin{aligned} \nabla J(a, \Delta a) &\triangleq \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} [J(a + \varepsilon \Delta a) - J(a)] \\ &= E \left\{ \int_0^T [\sigma'(t, a_t) Z_t + h'(t, a_t)] \Delta a_t dt \right\} \end{aligned}$$

where σ', h' are derivatives with respect to a . If a^* is optimal, then $\nabla J(a^*, \Delta a) \leq 0$ for any Δa . As a necessary condition, we obtain the *stochastic maximum principle*:

$$\sigma'(t, a_t^*) Z_t + h'(t, a_t^*) = 0 \quad (20)$$

Under certain technical conditions, we get $a_t^* = I(t, Z_t)$ for some deterministic function I . Plugging this into equations (18) and (19) we obtain a coupled FBSDE.

References

- [1] Antonelli, F. (1993). Backward-forward stochastic differential equations, *The Annals of Applied Probability* **3**(3), 777–793.
- [2] Bally, V. (1997). Approximation scheme for solutions of BSDE, in *Backward Stochastic Differential Equations (Paris 1995–1996)*, N. El Karoui & L. Mazliak, eds, Pitman Research Notes in Mathematics Series, Longman, Harlow, Paris, Vol. 364, pp. 177–191.
- [3] Bally, V. & Pagès, G. (2003). Error analysis of the quantization algorithm for obstacle problems, *Stochastic Processes and their Applications* **106**, 1–40.
- [4] Bender, C. & Denk, R. (2007). A forward scheme for backward SDEs, *Stochastic Processes and their Applications* **117**(12), 1793–1823.
- [5] Bender, C. & Zhang, J. (2008). Time discretization and Markovian iteration for coupled FBSDEs, *The Annals of Applied Probability* **18**(1), 143–177.
- [6] Bensoussan, A. (1983). Stochastic maximum principle for distributed parameter systems, *Journal of the Franklin Institute* **315**(5–6), 387–406.
- [7] Bismut, J.M. (1973). *Théorie Probabiliste du Contrôle des Diffusions*, Memoirs of the American Mathematical Society, Providence, Rhode Island, Vol. 176.
- [8] Bouchard, B. & Touzi, N. (2004). Discrete-time approximation and Monte-Carlo simulation of backward stochastic differential equations, *Stochastic Processes and their Applications* **111**, 175–206.
- [9] Briand, P., Delyon, B. & Mémoin, J. (2001). Donsker-type theorem for BSDEs, *Electronics Communications in Probability* **6**, 1–14.
- [10] Briand, P. & Hu, Y. (2006). BSDE with quadratic growth and unbounded terminal value, *Probability Theory and Related Fields* **136**(4), 604–618.
- [11] Buckdahn, R. & Hu, Y. (1998). Hedging contingent claims for a large investor in an incomplete market, *Advances in Applied Probability* **30**(1), 239–255.

- [12] Cheridito, P., Soner, M., Touzi, N. & Victoir, N. (2006). Second order backward stochastic differential equations and fully non-linear parabolic PDEs, *Communications in Pure and Applied Mathematics* **60**, 1081–1110.
- [13] Cvitanic, J. & Karatzas, I. (1996). Backward SDE's with reflection and Dynkin games, *The Annals of Probability* **24**, 2024–2056.
- [14] Cvitanic, J., Karatzas, I. & Soner, M. (1998). Backward stochastic differential equations with constraints on the gains-process, *The Annals of Probability* **26**(4), 1522–1551.
- [15] Cvitanic, J. & Ma, J. (1996). Hedging options for a large investor and forward-backward SDE's, *The Annals of Applied Probability* **6**(2), 370–398.
- [16] Delarue, F. (2002). On the existence and uniqueness of solutions to FBSDEs in a non-degenerate case, *Stochastic Processes and their Applications* **99**(2), 209–286.
- [17] Delarue, F. & Menozzi, S. (2006). A forward backward stochastic algorithm for quasi-linear PDEs, *The Annals of Applied Probability* **16**, 140–184.
- [18] Douglas, J., Ma, J. & Protter, P. (1996). Numerical methods for forward backward stochastic differential equations, *The Annals of Applied Probability* **6**, 940–968.
- [19] Duffie, D. & Epstein, L. (1992). Stochastic differential utility, *Econometrica* **60**, 353–394.
- [20] Duffie, D. & Epstein, L. (1992). Asset pricing with stochastic differential utility, *Review of Financial Studies* **5**, 411–436.
- [21] Duffie, D., Ma, J. & Yong, J. (1995). Black's consol rate conjecture, *The Annals of Applied Probability* **5**(2), 356–382.
- [22] El Karoui, N., Kapoudjian, C., Pardoux, E., Peng, S. & Quenez, M.C. (1997). Reflected solutions of backward SDE's, and related obstacle problems for PDE's, *The Annals of Probability* **25**(2), 702–737.
- [23] El Karoui, N. & Mazliak, L. (1997). *Backward Stochastic Differential Equations*, Pitman Research Notes in Mathematics Series, Longman, Harlow, Vol. 364.
- [24] El Karoui, N., Peng, S. & Quenez, M.C. (1997). Backward stochastic differential equations in finance, *Mathematical Finance* **7**, 1–72.
- [25] El Karoui, N., Peng, S. & Quenez, M.C. (2001). A dynamic maximum principle for the optimization of recursive utilities under constraints, *The Annals of Applied Probability* **11**(3), 664–693.
- [26] El Karoui, N. & Quenez, M.C. (1995). Dynamic programming and pricing of contingent claims in an incomplete market, *SIAM Journal on Control and Optimization* **33**(1), 29–66.
- [27] Gobet, E., Lemor, J.-P. & Warin, X. (2005). A regression-based Monte-Carlo method to solve backward stochastic differential equations, *The Annals of Applied Probability* **15**, 2172–2202.
- [28] Hamadene, S. & Lepeltier, J.-P. (1995). Zero-sum stochastic differential games and backward equations, *Systems and Control Letters* **24**(4), 259–263.
- [29] Hu, Y. & Peng, S. (1995). Solution of forward-backward stochastic differential equations, *Probability Theory and Related Fields* **103**(2), 273–283.
- [30] Kobylanski, M. (2000). Backward stochastic differential equations and partial differential equations with quadratic growth, *The Annals of Probability* **28**(2), 558–602.
- [31] Lepeltier, J.P. & San Martín, J. (1997). Backward stochastic differential equations with continuous coefficients, *Statistics and Probability Letters* **32**, 425–430.
- [32] Ma, J. & Cvitanic, J. (2001). Reflected forward-backward SDEs and obstacle problems with boundary conditions, *Journal of Applied Mathematics and Stochastic Analysis* **14**(2), 113–138.
- [33] Ma, J., Protter, P., San Martín, J. & Torres, S. (2002). Numerical method for backward stochastic differential equations, *The Annals of Applied Probability* **12**(1), 302–316.
- [34] Ma, J., Protter, P. & Yong, J. (1994). Solving forward-backward stochastic differential equations explicitly - a four step scheme, *Probability Theory and Related Fields* **98**, 339–359.
- [35] Ma, J. & Yong, J. (1999). *Forward-backward Stochastic Differential Equations and their Applications*, Lecture Notes in Mathematics, Springer, Vol. 1702.
- [36] Ma, J. & Zhang, J. (2002). Representation theorems for backward SDEs, *The Annals of Applied Probability* **12**, 1390–1418.
- [37] Pardoux, E. & Peng, S. (1990). Adapted solutions of backward stochastic equations, *System and Control Letters* **14**, 55–61.
- [38] Pardoux, E. & Peng, S. (1992). *Backward Stochastic Differential Equations and Quasilinear Parabolic Partial Differential Equations*, Lecture Notes in CIS, Springer, Vol. 176, pp. 200–217.
- [39] Pardoux, E. & Tang, S. (1999). Forward-backward stochastic differential equations and quasilinear parabolic PDEs, *Probability Theory and Related Fields* **114**(2), 123–150.
- [40] Peng, S. (1990). A general stochastic maximum principle for optimal control problems, *SIAM Journal on Control and Optimization* **28**(4), 966–979.
- [41] Peng, S. (1992). A nonlinear Feynman-Kac formula and applications, in *Control Theory, Stochastic Analysis and Applications: Proceedings of the Symposium on System Sciences and Control Theory (Hangzhou, 1992)*, S.P. Shen & J.M. Yong, eds, World Scientific Publications, River Edge, NJ, pp. 173–184.
- [42] Peng, S. & Wu, Z. (1999). Fully coupled forward-backward stochastic differential equations and applications to optimal control, *SIAM Journal on Control and Optimization* **37**(3), 825–843.
- [43] Yong, J. (1997). Finding adapted solutions of forward-backward stochastic differential equations: method of

- continuation, *Probability Theory and Related Fields* **107**(4), 537–572.
- [44] Yong, J. & Zhou, X. (1999). *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer.
- [45] Zhang, J. (2004). A numerical scheme for BSDEs, *The Annals of Applied Probability* **14**(1), 459–488.
- [46] Zhang, J. (2006). The wellposedness of FBSDEs, *Discrete and Continuous Dynamical Systems-series B* **6**, 927–940.

Related Articles

**Backward Stochastic Differential Equations;
Backward Stochastic Differential Equations: Numerical Methods; Doob–Meyer Decomposition.**

JIANFENG ZHANG

Martingale Representation Theorem

The “martingale representation theorem” is one of the fundamental theorems of stochastic calculus. It was first noted by Itô [9] (see **Itô, Kiyosi (1915–2008)**) as an application of multiple Wiener–Itô integrals. It was later modified and extended to various forms by many authors, but the basic theme remains the same: a square-integrable (local) martingale with respect to the filtration generated by a Brownian motion can always be represented as an Itô integral with respect to *that* Brownian motion. An immediate consequence would then be that every square-integrable martingale with respect to a *Brownian filtration* must have continuous paths. The martingale representation theorem is particularly useful in fields such as nonlinear filtering and mathematical finance [12] (see **Second Fundamental Theorem of Asset Pricing**) and it is a fundamental building block of the theory of *backward stochastic differential equations* [17, 19] (see **Backward Stochastic Differential Equations**).

To state the martingale representation theorem more precisely, let us consider a probability space $(\Omega, \mathcal{F}, P)$, on which is defined a d -dimensional Brownian motion B . We denote the filtration generated by B as $\mathbf{F}^B \triangleq \{\mathcal{F}_t^B\}_{t \geq 0}$, where $\mathcal{F}_t^B = \sigma\{B_s : s \leq t\} \vee \mathcal{N}$, $t \geq 0$, and $\mathcal{N}$ is the set of all P -null sets in $\mathcal{F}$. It can be checked that the filtration $\mathbf{F}^B$ is *right continuous* (i.e., $\mathcal{F}_t = \mathcal{F}_{t+}^B \triangleq \bigcap_{\varepsilon > 0} \mathcal{F}_{t+\varepsilon}^B$, $t \geq 0$), and $\mathcal{F}_t^B$ contains all P -null sets of $\mathcal{F}$. In other words, $\mathbf{F}^B$ satisfies the so-called usual hypotheses [20] (see **Filtrations**). Let us denote $\mathcal{M}^2(\mathbf{F}^B)$ to be the set of all square-integrable $\mathbf{F}^B$ -martingales and $\mathcal{M}_c^2(\mathbf{F}^B)$ to be the subspace of $\mathcal{M}^2(\mathbf{F}^B)$ of all those martingales that have continuous paths. The most common martingale representation theorem is the following:

Theorem 1 *Let $M \in \mathcal{M}^2(\mathbf{F}^B)$. Then there exists a d -dimensional $\mathbf{F}^B$ -predictable process H with $E \int_0^T |H_s|^2 ds < \infty$ for all $T > 0$, such that*

$$M_t = M_0 + \int_0^t (H_s, dB_s)$$

$$= M_0 + \sum_{i=1}^d \int_0^t H_s^i dB_s^i \quad \forall t \geq 0 \quad (1)$$

Furthermore, the process H is unique modulo $dt \times dP$ -null sets. Consequently, it holds that $\mathcal{M}^2(\mathbf{F}^B) = \mathcal{M}_c^2(\mathbf{F}^B)$.

The proof of this theorem can be found in standard reference books in stochastic analysis, for example, Ikeda and Watanabe [8], Karatzas and Shreve [12], Liptser and Shiryaev [14], Protter [20], and Rogers and Williams [21], to mention a few. But the work of Dellacherie [1] is worth mentioning, since it is the basis for many other proofs in the literature.

Note that if ξ is an $\mathcal{F}_T^B$ -measurable random variable for some $T > 0$ with finite second moments, then $M_t \triangleq E\{\xi | \mathcal{F}_t^B\}$, $t \geq 0$, defines a square-integrable $\mathbf{F}^B$ -martingale. We therefore have the following corollary:

Corollary 1 *Assume that ξ is a $\mathcal{F}_T^B$ -measurable random variable for some $T > 0$, such that $E[|\xi|^2] < \infty$. Then there exists a d -dimensional $\mathbf{F}^B$ -predictable process H with $E \int_0^T |H_s|^2 ds < \infty$ such that*

$$\begin{aligned} \xi &= E[\xi] + \int_0^T (H_s, dB_s) \\ &= E[\xi] + \sum_{i=1}^d \int_0^T H_s^i dB_s^i, \quad P \text{ a.s.} \end{aligned} \quad (2)$$

Furthermore, the process H is unique modulo $dt \times dP$ -null sets.

We remark that in the above corollary, the process H , often referred to as the *martingale integrand* or *representation kernel* of the martingale M , could depend on the duration $T > 0$; therefore, a more precise notation would be $H = H^T$, if the time duration T has to be taken into consideration. But the uniqueness of the representation implies that the family $\{H^T\}$ is actually “consistent” in the sense that $H_t^{T_1} = H_t^{T_2}$, $dt \times dP$ a.e. on $[0, T_1] \times \Omega$, if $T_1 \leq T_2$.

The martingale representation theorem can be generalized to local martingales [12, 20, 21]:

Theorem 2 *Every $\mathbf{F}^B$ -local martingale is continuous and is the stochastic integral with respect to B of*

a predictable process H such that

$$P \left\{ \int_0^t |H_s|^2 ds < \infty : t \geq 0 \right\} = 1 \quad (3)$$

We note that there is a slight difference between Corollary 1 and Theorem 2, on the integrability of the integrand H . In fact, without the local martingale assumption the “local” square integrability such as equation (3) does not guarantee the uniqueness of the process H in Corollary 1. A very elegant result in this regard is attributed to Dudley [4], who proved that any almost surely finite $\mathcal{F}_T$ -measurable random variable ξ can be represented as a stochastic integral evaluated at T , and the “martingale integrand” satisfies only equation (3). However, such representation *does not* have uniqueness. This point was further investigated in [7]. In this study, the filtration is generated by a higher dimensional Brownian motion, of which B is only a part of the components. We also refer to [12] for the discussions on this issue.

Itô’s original martingale representation theorem has been extended to many other situations when the Brownian motion is replaced by certain semimartingales. In this section, we give a brief summary of these cases. For simplicity in what follows, we shall consider only martingales rather than local martingales. The versions for the latter are essentially identical, but with slightly relaxed integrability requirements on the representing integrands, as we saw in Theorem 2.

Representation under Non-Brownian Filtrations

We recall that one of the most important assumptions in the martingale representation theorems is that the filtration is generated by the Brownian motion (or “Brownian-filtration”). When this assumption is removed, the representation may still hold, but the form will change. There are different ways to adjust the result:

1. Fix the probability space, but change the form of representation (by adding an orthogonal martingale).
2. Fix the probability space, but use more information of the martingale to be represented.
3. Extend the probability space, but keep the form of the representation.

The generalization of type (1) essentially uses the idea of orthogonal decomposition of the Hilbert space. In fact, note that $\mathcal{M}^2(\mathbf{F})$ is a Hilbert space, and let $\mathcal{H}$ denote all $H \in \mathcal{M}^2(\mathbf{F})$ such that $H_t = \int_0^t \Phi_s dB_s$, $t \geq 0$ for some progressively measurable process $\Phi \in L^2([0, T] \times \Omega)$. Then $\mathcal{H}$ is a closed subspace of $\mathcal{M}^2(\mathbf{F})$; thus for any $M \in \mathcal{M}^2(\mathbf{F})$ the following decomposition holds:

$$\begin{aligned} M_t &= M_0 + H_t + N_t \\ &= M_0 + \int_0^t \Phi_s dB_s + N_t, \quad t \geq 0 \end{aligned} \quad (4)$$

where $N \in \mathcal{N}^\perp$, the subspace of $\mathcal{M}^2(\mathbf{F})$ consisting of all martingales that are “orthogonal” to $\mathcal{H}$. We refer to [12] and [20], for example, for detailed discussions for this type of representations. The generalizations of types (2) and (3) keep the original form of the representation. We now list two results adapted from Ikeda–Watanabe [8].

Theorem 3 *Let $M^i \in \mathcal{M}_c^2(\mathbf{F})$, $i = 1, 2, \dots, d$. Suppose that $\Phi^{i,j} \in \mathcal{L}^1(\mathbf{F})$ and $\Psi^{i,k} \in \mathcal{L}^2(\mathbf{F})$, $i, j, k = 1, 2, \dots, d$, exist such that for $i, j = 1, 2, \dots, d$,*

$$\begin{aligned} \langle M^i, M^j \rangle_t &= \int_0^t \Phi_s^{ij} ds \quad \text{and} \\ \Phi_s^{i,j} &= \sum_{k=1}^d \Psi_s^{ik} \Psi_s^{jk}, \quad P \text{ a.s.} \end{aligned} \quad (5)$$

and $\det(\Psi_s^{jk}) \neq 0$, a.s., for all $s \geq 0$. Then there exists a d -dimensional $\mathbf{F}$ -Brownian motion $B = \{B_t^1, \dots, B_t^d\} : t \geq 0\}$ such that

$$M_t^i = M_0^i + \sum_{k=1}^d \int_0^t \Psi_s^{ik} dB_s^k, \quad i = 1, 2, \dots, d \quad (6)$$

We remark that the assumption $\det(\Psi_s^{jk}) \neq 0$ in Theorem 3 is quite restrictive, which implies, among other things, that the representing Brownian motion has to have the same dimension as the given martingale (thus the representation kernel is “squared”). This restriction can be removed by allowing the probability space to be enlarged (or *extended*, see [8]).

Theorem 4 Let $M^i \in \mathcal{M}_c^2(\mathbf{F})$, $i = 1, 2, \dots, d$. Suppose that $\Phi^{i,j}, \Psi^{i,k} \in \mathcal{L}^0(\mathbf{F})$, $i, j = 1, 2, \dots, d$, $k = 1, 2, \dots, r$ exist such that for $i, j = 1, 2, \dots, d$ and $k = 1, 2, \dots, r$, $\int_0^t |\Phi_s^{ij}| ds < \infty$ and $\int_0^t |\Psi_s^{ik}|^2 ds < \infty$, $t \geq 0$, P a.s., and that

$$\langle M^i, M^j \rangle_t = \int_0^t \Phi_s^{ij} ds \quad \text{and} \quad \Phi_s^{i,j} = \sum_{k=1}^d \Psi_s^{ik} \Psi_s^{jk}, \quad P \text{ a.s.} \quad (7)$$

Then there exists an extension $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}; \tilde{\mathbf{F}})$ of $(\Omega, \mathcal{F}, P; \mathbf{F})$, and a d -dimensional $\tilde{\mathbf{F}}$ -Brownian motion $B = \{B_t^1, \dots, B_t^d : t \geq 0\}$ such that

$$M_t^i = M_0^i + \sum_{k=1}^d \int_0^t \Psi_s^{ik} dB_s^k, \quad i = 1, 2, \dots, d \quad (8)$$

Representation for Discontinuous Martingales

Up to this point, all the representable martingales are, in fact necessarily, continuous. This clearly excludes many important martingales, most notably the compensated Poisson processes. Thus another generalization of the martingale representation theorem is to replace the Brownian motion by Poisson random measure. We refer to Ikeda and Watanabe [8], for example, for the basic notions of Poisson point process and Poisson random measures.

Let p be a Poisson point process (see **Point Processes**) on some state space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, where $\mathcal{B}(\mathcal{X})$ stands for the Borel field of $\mathcal{X}$. For each $t > 0$ and $U \in \mathcal{B}(\mathcal{X})$, define the counting measure $N_p(t, U) \triangleq \sum_{s \leq t} \mathbf{1}_U(p(s))$. We assume that the point process p is of class (QL), that is, the compensator $\hat{N}_p(\cdot, U) = E[N_p(\cdot, U)]$ is continuous for each U ; and $\tilde{N}_p(t, U) = N_p(t, U) - \hat{N}_p(t, U)$ is a martingale. Similar to the Brownian case, we can define the filtration generated by p as $\mathcal{F}_t^p \triangleq \sigma\{N_p(s, U) : s \leq t, U \in \mathcal{B}(\mathcal{X})\}$ (or make it right continuous by defining $\tilde{\mathcal{F}}_t^p = \cap_{\varepsilon > 0} \mathcal{F}_{t+\varepsilon}^p$), and denote $\mathbf{F}^p = \{\mathcal{F}_t^p\}_{t \geq 0}$. We then have the following analog of Theorem 1.

Theorem 5 Let $M \in \mathcal{M}^2(\mathbf{F}^p)$. Then there exists an $\mathbf{F}^p$ -predictable random field $f : \Omega \times [0, \infty) \times$

$\mathcal{X} \mapsto \mathbb{R}$ satisfying $E \left\{ \int_0^t \int_{\mathcal{X}} |f(s, x, \cdot)|^2 \hat{N}_p(ds, dx) \right\} < \infty$, such that

$$M_t = M_0 + \int_0^{t+} \int_{\mathcal{X}} f(s, x, \cdot) \tilde{N}_p(ds, dx), \quad t \geq 0 \quad (9)$$

We should note that like Theorem 1, Theorem 5 also has generalizations that could be considered as counterparts of Theorems 3 and 4 [8]. It is worth noting that by combining Theorems 1 and 5, it is possible to obtain a martingale representation theorem that involves both Brownian motion and the Poisson random measure. Keeping the Lévy–Khintchine formula (see **Lévy Processes**) (or Lévy–Itô Theorem) in mind, we have the following representation theorem, which is a simplified version resulting from a much deeper and extensive exposition by Jacod and Shiryaev [10] (see also [13]). Let $\mathbf{F}$ be the filtration generated by a Lévy process with the Brownian component B and Poisson component N .

Theorem 6 Suppose that $M \in \mathcal{M}^2(\mathbf{F})$. Then there exist an $\mathbf{F}$ -adapted process H and a random field G satisfying $E \int_0^T |H_s|^2 ds < \infty$, $E \left\{ \int_0^t \int_{\mathbb{R} \setminus 0} |G(s, x)|^2 \tilde{N}(ds, dx) \right\} < \infty$, such that

$$M_t = M_0 + \int_0^t H_s dB_s + \int_0^t \int_{\mathbb{R} \setminus 0} G(s, x) \tilde{N}(ds, dx) \quad (10)$$

Moreover, the elements of the pair (H, G) are unique in their respective spaces.

In Theorem 6, the Brownian component and the Poisson component of the Lévy process have to be treated *separately*, and one *cannot* simply replace the Brownian motion in Theorem 1 by a Lévy process. In fact, the martingale representation for Lévy process is a much more subtle issue, and was recently studied by Nualart and Schoutens [18] via the chaotic representation using the so-called Teugels martingales. We refer also to Løkka [15] for a more recent development on this issue.

A natural question now is whether the martingale representation theorem can still hold (in the usual sense) for martingales with jumps. The answer to this question has an important implication in finance, since, as we shall see subsequently, this is the

same as asking whether a market could be complete when the dynamics of the underlying assets have jumps. It turns out that there indeed exists a class of martingales, known as the *normal martingales*, that are discontinuous in general but the martingale representation theorem holds. A square-integrable martingale M is called *normal* if $\langle M \rangle_t = t$ (cf. [2]). The class of normal martingale, in particular, includes those martingales that satisfy the so-called structure equation (cf. [5, 6]). Examples of normal martingales satisfying the structure equation include Brownian motion, compensated Poisson process, the Azéma martingale, and the “parabolic” martingale [20]. The martingale representation, or more precisely the Clark–Ocone formula, was proved in [16]. The application of such a representation in finance was first done by Dritschel and Protter [3] (see also [11]).

Relation with Hedging

The martingale representation theorem is the basis for the arguments leading to *market completeness*, a fundamental component in the “Second Fundamental Theorem” of mathematical finance (see **Second Fundamental Theorem of Asset Pricing**). Consider a market modeled by a probability space $(\Omega, \mathcal{F}, P, \mathbf{F})$, where $\mathbf{F}$ is the filtration generated by a Brownian motion that represents market randomness, and denote it by B . Assume that the market is arbitrage free; then there exists a risk neutral measure Q (see **Fundamental Theorem of Asset Pricing**), equivalent to P . The arbitrage price at time $t \in [0, T]$ for any contingent T -claim $\mathcal{X}$ is given by the *discounted present value formula*:

$$V_t = e^{-r(T-t)} E^Q[\mathcal{X} | \mathcal{F}_t], \quad t \in [0, T] \quad (11)$$

where r is the (constant) interest rate. If $\mathcal{X}$ is square integrable, then $M_t = e^{-rt} V_t$, $t \geq 0$, is a square-integrable $\mathbf{F}$ -martingale under Q . Applying the martingale representation theorem one has

$$M_t = M_0 + \int_0^t \phi_s dB_s, \quad t \in [0, T] \quad (12)$$

for some square-integrable, $\mathbf{F}$ -predictable process ϕ . Or equivalently, assuming that the volatility of the

market, denoted by σ , is positive, we can write

$$V_t = V_0 + \int_0^t r V_s ds + \int_0^t \pi_t \sigma_s dB_s, \quad t \in [0, T] \quad (13)$$

where $\pi_t = e^{rt} \phi_t \sigma_t^{-1}$, $t \geq 0$. The process π is then exactly the “hedging strategy” for the claim $\mathcal{X}$, that is, the amount of money one should invest in the stock, so that $V_T = \mathcal{X}$, almost surely.

The martingale representation theorem also plays an important role in portfolio optimization problems, especially in finding optimal strategies [12].

One of the abstract forms of the hedging problem described earlier is the so-called backward stochastic differential equation (BSDE), which is the problem of finding a pair of $\mathbf{F}$ -adapted processes (V, Z) so that the following *terminal value problems* for a stochastic differential equation similar to (13) holds:

$$\begin{aligned} dV_t &= f(t, V_t, Z_t) dt + Z_t dB_t, \quad t \in [0, T] \\ V_T &= \mathcal{X} \end{aligned} \quad (14)$$

See **Forward–Backward Stochastic Differential Equations (SDEs); Backward Stochastic Differential Equations**.

References

- [1] Dellacherie, C. (1974). *Intégrales Stochastiques par Rapport aux Processus de Wiener et de Poisson*, Séminaire de Probability (Univ. de Strasbourg) IV, Lecture Notes in Math, Springer-Verlag, Berlin, Vol. 124, 77–107.
- [2] Dellacherie, C., Maisonneuve, B. & Meyer, P.A. (1992). *Probabilités et Potentiel: Chapitres XVII à XXIV*, Hermann, Paris.
- [3] Dritschel, M. & Protter, P. (1999). Complete markets with discontinuous security price, *Finance and Stochastics* **3**(2), 203–214.
- [4] Dudley, R.M. (1977). Wiener functionals as Itô integrals, *Annals of Probability* **5**, 140–141.
- [5] Emery, M. (1989). *On the Azéma Martingales*, Séminaire de Probabilités XXIII, Lecture Notes in Mathematics, Vol. 1372, Springer Verlag, pp. 66–87.
- [6] Emery, M. (2006). Chaotic representation property of certain Azéma martingales, *Illinois Journal of Mathematics* **50**(2), 395–411.
- [7] Emery, M., Stricker, C. & Yan, J. (1983). Valuers prises par les martingales locales continues à un instant donné, *Annals of Probability* **11**, 635–641.

-
- [8] Ikeda, N. & Watanabe, S. (1981). *Stochastic Differential Equations and Diffusion Processes*, North-Holland.
 - [9] Itô, K. (1951). Multiple Wiener integral, *Journal of Mathematical Society of Japan* **3**, 157–169.
 - [10] Jacod, J. & Shiryaev, A.N. (1987). *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin.
 - [11] Jeanblanc, M. & Privault, N. (2002). A complete market model with Poisson and Brownian components, *Seminar on Stochastic Analysis, Random Fields and Applications*, Ascona; *Progress in Probability*, **52**, 189–204.
 - [12] Karatzas, I. & Shreve, S.E. (1987). *Brownian Motion and Stochastic Calculus*, Springer.
 - [13] Kunita, H. (2004). Representation of martingales with jumps and applications to mathematical finance, in *Stochastic Analysis and Related Topics in Kyoto, Advanced Studies in Pure Mathematics 41*, H. Kunita, S. Watanabe & Y. Takahashi eds, Mathematical Society of Japan, Tokyo, pp. 209–232.
 - [14] Liptser, R.S. & Shiryaev, A.N. (1977). *Statistics of Random Processes. Vol I: General Theory*, Springer-Verlag, New York.
 - [15] Løkka, A. (2004). Martingale representation of functionals of Lévy processes, *Stochastic Analysis and Applications* **22**(4), 867–892.
 - [16] Ma, J., Protter, P., & San Martin, J. (1998). Anticipating integrals for a class of martingales, *Bernoulli* **4**(1), 81–114.
 - [17] Ma, J. & Yong, J. (1999). *Forward-Backward Stochastic Differential Equations and Their Applications*, LNM 1702, Springer.
 - [18] Nualart, D. & Schoutens, W. (2000). Chaotic and predictable representations for Lévy processes, *Stochastic Processes and their Applications* **90**, 109–122.
 - [19] Pardoux, E. & Peng, S. (1990). Adapted solutions of backward stochastic equations, *System and Control Letters* **14**, 55–61.
 - [20] Protter, P. (1990). *Stochastic Integration and Stochastic Differential Equations*, Springer.
 - [21] Rogers, L.C.G. & Williams, D. (1987). *Diffusions, Markov Processes and Martingales, Vol. 2: Itô Calculus*, John Wiley & Sons.

Further Reading

- Dellacherie, C. & Meyer, P. (1978). *Probabilities and Potential*, North-Holland.
- Doob, J.L. (1984). *Classical Potential Theory and its Probabilistic Counterparts*, Springer.
- Revuz, D. & Yor, M. (1991, 1994). *Continuous Martingales and Brownian Motion*, Springer.

Related Articles

Backward Stochastic Differential Equations; Convex Duality; Complete Markets; Filtrations; Second Fundamental Theorem of Asset Pricing.

JIN MA

Backward Stochastic Differential Equations

Backward stochastic differential equations (BSDEs) occur in situations where the terminal (as opposed to the initial) condition of stochastic differential equations is a given random variable. Linear BSDEs were first introduced by Bismut (1976) as the adjoint equation associated with the stochastic version of the Pontryagin maximum principle in control theory. The general case of a nonlinear BSDE was first introduced by Peng and Pardoux [23] to give a Feynman–Kac representation of nonlinear parabolic partial differential equations (PDEs). The solution of a BSDE consists of a pair of adapted processes (Y, Z) satisfying

$$-dY_t = f(t, Y_t, Z_t)dt - Z'_t dW_t, \quad Y_T = \xi \quad (1)$$

where f is called the *driver* and ξ the *terminal condition*. This type of equation appears naturally in hedging problems. For example, in a complete market (see **Complete Markets**), the price process $(Y_t)_{0 \leq t \leq T}$ of a European contingent claim ξ with maturity T corresponds to the solution of a BSDE with a linear driver f and a terminal condition equal to ξ .

Reflected BSDEs were introduced by El Karoui *et al.* [6]. In the case of a reflected BSDE, the solution Y is constrained to be greater than a given process called the *obstacle*. A nondecreasing process K is introduced in the equation in order to push (upward) the solution so that the constraint is satisfied, and this push is minimal, that is, Y satisfies the following equation:

$$-dY_t = f(t, Y_t, Z_t)dt + dK_t - Z'_t dW_t, \quad Y_T = \xi \quad (2)$$

with $(Y_t - S_t) dK_t = 0$. One can show that the price of an American option (with eventually some non-linear constraints) is the solution of a reflected BSDE, where the obstacle is given by the payoff process.

Definition and Properties

We adopt the following notation: $\mathcal{IF} = \{\mathcal{F}_t, 0 \leq t \leq T\}$ is the natural filtration of an n -dimensional

Brownian motion W ; L^2 is the set of random variables ξ that are $\mathcal{F}_T$ -measurable and square-integrable; $\mathcal{IH}^2$ is the set of predictable processes ϕ such that $E \int_0^T |\phi_t|^2 dt < \infty$. In the following, the sign $'$ denotes transposition.

Let us consider the following BSDE (with dimension 1 to simplify the presentation):

$$-dY_t = f(t, Y_t, Z_t)dt - Z'_t dW_t, \quad Y_T = \xi \quad (3)$$

where $\xi \in L^2$ and f is a *driver*, that is, it satisfies the following assumptions: $f : \Omega \times [0, T] \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$ est $\mathcal{P} \otimes \mathcal{B} \otimes \mathcal{B}^n$ -measurable, $f(\cdot, 0, 0) \in \mathcal{IH}^2$ and f is uniformly Lipschitz with respect to y, z with constant $C > 0$. Such a pair (ξ, f) is called a pair of *standard parameters*. If the driver f does not depend on y and z , the solution Y of equation (3) is then given as

$$Y_t = E \left[\xi + \int_t^T f(s) ds / \mathcal{F}_t \right] \quad (4)$$

and the martingale representation theorem for Brownian motion ([16] Theorem 4.15) gives the existence of a unique process $Z \in \mathcal{IH}^2$ such that

$$E \left[\xi + \int_0^T f(s) ds / \mathcal{F}_t \right] = Y_0 + \int_0^t Z'_s dW_s \quad (5)$$

In 1990, Peng and Pardoux [23] stated the following theorem.

Theorem 1 *If $\xi \in L^2$ and if f is a driver, then there exists a unique pair of solutions $(Y, Z) \in \mathcal{IH}^2 \times \mathcal{IH}^2$ of equation (3).*

In [7], El Karoui *et al.* have given a short proof of this theorem based on *a priori estimations* of the solutions. More precisely, the proposition is given as follows:

Proposition 1 (A Priori Estimations). *Let f^1, ξ^1, f^2, ξ^2 be standard parameters. Let (Y^1, Z^1) be the solution associated with f^1, ξ^1 and (Y^2, Z^2) be the solution associated with f^2, ξ^2 . Let C be the Lipschitz constant of f^1 . Substitute $\delta Y_t = Y_t^1 - Y_t^2, \delta Z_t = Z_t^1 - Z_t^2$, and $\delta_2 f_t = f^1(t, Y_t^2, Z_t^2) - f^2(t, Y_t^2, Z_t^2)$. For (λ, μ, β) such that $\lambda^2 > C$ and β sufficiently*

large, that is, $\beta > C(2 + \lambda^2) + \mu^2$, the following estimations hold:

$$\|\delta Y\|_\beta^2 \leq T \left[e^{\beta T} E(|\delta Y_T|^2) + \frac{1}{\mu^2} \|\delta_2 f\|_\beta^2 \right] \quad (6)$$

$$\|\delta Z\|_\beta^2 \leq \frac{\lambda^2}{\lambda^2 - C} \left[e^{\beta T} E(|\delta Y_T|^2) + \frac{1}{\mu^2} \|\delta_2 f\|_\beta^2 \right] \quad (7)$$

where $\|\delta Y\|_\beta^2 = E \int_0^T e^{\beta t} |\delta Y_t|^2 dt$.

From these estimations, *uniqueness* and *existence* of a solution follow by using the fixed point theorem applied to the function $\Psi : \mathbb{H}_\beta^2 \otimes \mathbb{H}_\beta^2 \rightarrow \mathbb{H}_\beta^2 \otimes \mathbb{H}_\beta^2; (y, z) \mapsto (Y, Z)$, where (Y, Z) is the solution associated with the driver $f(t, y_t, z_t)$ and $\mathbb{H}_\beta^2$ denotes the space $\mathbb{H}^2$ endowed with norm $\|\cdot\|_\beta$. Indeed, by using the previous estimations, one can show that for sufficiently large β , the mapping Ψ is strictly contracting, which gives the existence of a unique fixed point, which is the solution of the BSDE.

In addition, from “*a priori* estimations” (Proposition 1), some *continuity* and *differentiability* of solutions of BSDEs (with respect to some parameter) can be derived ([7] section 2).

Furthermore, estimations (1) are also very useful to derive some results concerning approximation or discretization of BSDEs [14].

Recall the dependence of the solutions of BSDEs with respect to terminal time T and terminal condition ξ by the notation $(Y_t(T, \xi), Z_t(T, \xi))$. We have the following flow property.

Proposition 2 (Flow Property). *Let $(Y(T, \xi), Z(T, \xi))$ be the solution of a BSDE associated with the terminal time $T > 0$ and standard parameters (ξ, f) .*

For any stopping time $S \leq T$,

$$\begin{aligned} Y_t(T, \xi) &= Y_t(S, Y_S(T, \xi)), \\ Z_t(T, \xi) &= Z_t(S, Y_S(T, \xi)), \\ t &\in [0, S], \quad dP \otimes dt\text{-almost surely} \end{aligned} \quad (8)$$

Proof By conventional notation, we define the solution of the BSDE with terminal condition (T, ξ) for $t \geq T$ by $(Y_t = \xi, Z_t = 0)$. Thus, if $T' \geq T$, then $((Y_t, Z_t); t \leq T')$ is the unique solution of the BSDE with terminal time T' , coefficient $f(t, y, z)\mathbf{1}_{[t \leq T]}$, and terminal condition ξ .

Let $S \leq T$ be a stopping time, and denote by $Y_t(S, \xi')$ the solution of the BSDE with terminal time T , coefficient $f(t, y, z)\mathbf{1}_{[t \leq S]}$, and terminal condition ξ' ($\mathcal{F}_S$ -measurable). Both the processes $(Y_t(S, Y_S), Z_t(S, Y_S); t \in [0, T])$ and $(Y_{t \wedge S}(T, \xi), Z(T, \xi)\mathbf{1}_{[t \leq S]}; t \in [0, T])$ are solutions of the BSDE with terminal time T , coefficient $f(t, y, z)\mathbf{1}_{[t \leq S]}$, and terminal condition Y_S . By uniqueness, these processes are the same $dP \otimes dt$ -a.s.

The simplest case is that of a *linear* BSDE. Let (β, γ) be a bounded $(\mathbb{R}, \mathbb{R}^n)$ -valued predictable process and let $\varphi \in \mathbb{H}^2(\mathbb{R})$, $\xi \in \mathbb{L}^2(\mathbb{R})$. We consider the following BSDE:

$$\begin{aligned} -dY_t &= (\varphi_t + Y_t \beta_t + Z_t' \gamma_t) dt - Z_t' dW_t, \\ Y_T &= \xi \end{aligned} \quad (9)$$

By applying Itô's formula to $\Gamma_t Y_t$, it can easily be shown that the process $\Gamma_t Y_t + \int_0^t \Gamma_s \varphi_s ds$ is a local martingale and even a uniformly integrable martingale, which gives the following proposition.

Proposition 3 *The solution (Y, Z) of the linear BSDE (9) satisfies*

$$\Gamma_t Y_t = E \left[\xi \Gamma_T + \int_t^T \Gamma_s \varphi_s ds \mid \mathcal{F}_t \right] \quad (10)$$

where Γ is the adjoint process (corresponding to a change of numéraire or a deflator in finance) defined by $d\Gamma_t = \Gamma_t[\beta_t dt + \gamma_t^* dW_t]$, $\Gamma_0 = 1$.

Remark 1 First, it can be noted that if ξ and φ are positive, then the process Y is positive. Second, if in addition $Y_0 = 0$ a.s., then for any t , $Y_t = 0$ a.s. and $\varphi_t = 0$ $dt \otimes dP$ -a.s.

From the first point in this remark, one can derive the classical *comparison theorem*, which is a key property of BSDEs.

Theorem 2 (Comparison Theorem). *If f^1, ξ^1 and f^2, ξ^2 are standard parameters and if (Y^1, Z^1) (respectively (Y^2, Z^2)) is the solution associated with (f^1, ξ^1) (respectively (f^2, ξ^2)) satisfying*

1. $\xi^1 \geq \xi^2$ P -a.s.
2. $\delta_2 f_t = f^1(t, Y_t^2, Z_t^2) - f^2(t, Y_t^2, Z_t^2) \geq 0$ $dt \times dP$ -a.s.
3. $f^1(t, Y_t^2, Z_t^2) \in \mathbb{H}^2$.

Then, we have $Y^1 \geq Y^2$ P -a.s.

In addition, the comparison theorem is strict, that is, on the event $\{Y_t^1 = Y_t^2\}$, we have $\xi_1 = \xi_2$ a.s., $f^1(t, Y_t^1, Z_t^1) = f^2(t, Y_t^2, Z_t^2)$ $ds \times dP$ -a.s. and $Y_s^1 = Y_s^2$ a.s., $t \leq s \leq T$.

Idea of the proof. We denote by δY the spread between those two solutions: $\delta Y_t = Y_t^2 - Y_t^1$ and $\delta Z_t = Z_t^2 - Z_t^1$. The problem is to show that under the above assumptions, $\delta Y_t \geq 0$.

Now, the pair $(\delta Y, \delta Z)$ is the solution of the following LBSDE:

$$\begin{aligned} -d\delta Y_t &= \delta_y f^2(t) \delta Y_t + \delta_z f^2(t) \delta Z_t + \varphi_t dt \\ &\quad - \delta Z_t dW_t, \\ \delta Y_T &= \xi^2 - \xi^1 \end{aligned} \quad (11)$$

where $\delta_y f^2(t) = \frac{f^2(t, Y_t^2, Z_t^2) - f^2(t, Y_t^1, Z_t^1)}{Y_t^2 - Y_t^1}$ if $Y_t^2 - Y_t^1$ is not equal to 0, and 0 otherwise (and the same for $\delta_z f^2(t)$). Now, since the driver f^2 is supposed to be uniformly Lipschitz with respect to (y, z) , it follows that $\delta_y f^2(t)$ and $\delta_z f^2(t)$ are bounded. In addition, φ_t and δY_T are nonnegative. It follows from the first point of Remark (1) that the solution δY_t of the LBSDE (11) is nonnegative. In addition, the second point of Remark (1) gives the *strict comparison theorem*.

From this theorem, we then state a general principle for *minima* of BSDEs [7]: if a driver f can be written as an infimum of a family of drivers f^α and if a random variable ξ can be written as an infimum of random variables ξ^α , then the solution of the BSDE associated with f and ξ can be written as the infimum of the solutions of the BSDEs associated with f^α, ξ^α .

More precisely, we have the following proposition.

Proposition 4 (Minima of BSDEs). *Let $(f, f^\alpha; \alpha \in \mathcal{A})$ be a family of drivers and let $(\xi, \xi^\alpha; \alpha \in \mathcal{A})$ be a family of terminal conditions. Let (Y, Z) be the solution of the BSDE associated with (f, ξ) and let (Y^α, Z^α) be the solution of the BSDE associated with (f^α, ξ^α) . Suppose that there exists a parameter $\bar{\alpha}$ such that*

$$\begin{aligned} f(t, Y_t, Z_t) &= \operatorname{ess\,inf}_{\alpha} f^\alpha(t, Y_t, Z_t) \\ &= \bar{f}(t, Y_t, Z_t), \quad dt \otimes dP\text{-a.s.} \end{aligned} \quad (12)$$

$$\xi = \operatorname{ess\,inf}_{\alpha} \xi^\alpha = \bar{\xi}, \quad P\text{-a.s.} \quad (13)$$

Then,

$$Y_t = \operatorname{ess\,inf}_{\alpha} Y_t^\alpha = \bar{Y}_t, \quad 0 \leq t \leq T, \quad P\text{-a.s.} \quad (14)$$

Proof For each α , since $f(t, Y_t, Z_t) \leq f^\alpha(t, Y_t, Z_t)$ $dt \otimes dP$ -a.s. and $\xi \leq \xi^\alpha$, the comparison theorem gives that $Y_t \leq Y_t^\alpha$ $0 \leq t \leq T$, P -a.s. It follows that

$$Y_t \leq \operatorname{ess\,inf}_{\alpha} Y_t^\alpha, \quad 0 \leq t \leq T, \quad P\text{-a.s.} \quad (15)$$

Now, by assumption, it is clear that $Y_t = \bar{Y}_t$, $0 \leq t \leq T$, P -a.s., which gives that the inequality in (15) is an equality, which ends the proof.

Note also that from the strict comparison theorem, one can derive an *optimality criterium* [7]:

Proposition 5 *A parameter $\bar{\alpha}$ is 0-optimal (i.e., $\min_{\alpha} Y_0^\alpha = Y_0^{\bar{\alpha}}$) if and only if*

$$\begin{aligned} f(s, Y_s, Z_s) &= f^{\bar{\alpha}}(s, Y_s, Z_s) dP \otimes ds\text{-a.s.} \\ \xi &= \bar{\xi} \quad P\text{-a.s.} \end{aligned} \quad (16)$$

The flow property (Proposition 2) of the value function corresponds to the *dynamic programming principle* in stochastic control.

Indeed, using the same notation as in Proposition 2, for any stopping time $S \leq T$,

$$\begin{aligned} Y_t(T, \xi) &= \operatorname{ess\,inf}_{\alpha} Y_t^\alpha(S, Y_S(T, \xi)), \\ 0 \leq t \leq S, \quad P\text{-a.s.} \end{aligned} \quad (17)$$

From the principle on minima of BSDEs (Proposition 4), one can easily obtain some links between BSDEs and *stochastic control* (see, e.g. [10] Section 3 for a financial presentation or [26] for a more classical presentation in stochastic control).

Note, in particular, that if this principle on minima of BSDEs is formulated a bit differently, it can be seen as a *verification theorem* for some stochastic control problem written in terms of BSDEs. More precisely, let $(f^\alpha; \alpha \in \mathcal{A})$ be a family of drivers and let $(\xi^\alpha; \alpha \in \mathcal{A})$ be a family of terminal conditions. Let (Y^α, Z^α) be the solution of the BSDE associated with (f^α, ξ^α) . The value function is defined at time t as

$$\bar{Y}_t = \operatorname{ess\,inf}_{\alpha} Y_t^\alpha, \quad P\text{-a.s.} \quad (18)$$

If there exist standard parameters f and ξ and a parameter $\bar{\alpha}$ such that equation (12) holds, then the value function coincides with the solution of the BSDE associated with (f, ξ) . In other words, $\bar{Y}_t = Y_t$, $0 \leq t \leq T$, P -a.s., where (Y, Z) denotes the solution of the BSDE associated with (f, ξ) . It can be noted that this *verification theorem* generalizes the well-known Hamilton–Jacobi–Bellman–verification theorem, which holds in a Markovian framework.

Indeed, recall that in the *Markovian case*, that is, the case where the driver and the terminal condition are functions of a state process, Peng and Pardoux (1992) have given an interpretation of the solution of a BSDE in terms of a PDE [24]. More precisely, the state process $X^{t,x}$ is a diffusion of the following type:

$$dX_s = b(s, X_s)ds + \sigma(s, X_s)dW_s, \quad X_t = x \quad (19)$$

Then, let us consider $(Y^{t,x}, Z^{t,x})$ solution of the following BSDE:

$$\begin{aligned} -dY_s &= f(s, X_s^{t,x}, Y_s, Z_s)ds - Z_s' dW_s, \\ Y_T &= g(X_T^{t,x}) \end{aligned} \quad (20)$$

where b , σ , f , and g are deterministic functions. In this case, one can show that under quite weak conditions, the solution $(Y_s^{t,x}, Z_s^{t,x})$ depends only on time s and on the state process $X_s^{t,x}$ (see [7] Section 4). In addition, if f and g are uniformly continuous with respect to x and if u denotes the function such that $Y_t^{t,x} = u(t, x)$, one can show (see [24] or [10] p. 226 for a shorter proof) that u is a viscosity solution of the following PDE:

$$\begin{aligned} \partial_t u + \mathcal{L}u(t, x) + f(t, x, u(t, x), \partial_x u \sigma(t, x)) &= 0, \\ u(T, x) &= g(x) \end{aligned} \quad (21)$$

where $\mathcal{L}$ denotes the infinitesimal generator of X (see **Forward–Backward Stochastic Differential Equations (SDEs); Markov Processes**). There are some complementary results concerning the case of a *non-Brownian* filtration (see [1] or [7] Section 5). In addition, some properties of *differentiability* in Malliavin’s sense of the solution of a BSDE can be given [7, 24]. In particular, under some smoothness assumptions on f , the process Z_t corresponds to the Malliavin derivative of Y_t , that is,

$$D_t Y_t = Z_t, \quad dP \otimes dt\text{-a.s.} \quad (22)$$

Many tentatives have been made to relax the Lipschitz assumption on the driver f ; for instance, Lepeltier and San Martín [19] and have proved the existence of a solution for BSDEs with a driver f , which is only *continuous with linear growth* by an approximation method. Kobylanski [17] studied the case of *quadratic* BSDEs [20]. To give some intuition on quadratic BSDEs, let us consider the following simple example:

$$\begin{aligned} -dY_t &= \frac{Z_t^2}{2}dt - Z_t dW_t, \\ Y_T &= \xi \end{aligned} \quad (23)$$

Let us make the exponential change of variable $y_t = e^{Y_t}$. By applying Itô’s formula, we easily derive

$$\begin{aligned} dy_t &= e^{Y_t} Z_t dW_t, \\ y_T &= e^\xi \end{aligned} \quad (24)$$

and hence, if ξ is supposed to be bounded and $Z \in H^2$, we have $y_t = E[e^\xi / \mathcal{F}_t]$. Thus, for quadratic BSDEs, it seems quite natural to suppose that the terminal condition is bounded. More precisely, the following existence result holds [17].

Proposition 6 (Quadratic BSDEs). *If the terminal condition ξ is bounded and if the driver f is linear growth in y and quadratic in z , that is,*

$$|f(t, y, z)| \leq C(1 + |y| + |z|^2) \quad (25)$$

then there exists an adapted pair of processes (Y, Z) , which is the solution of the quadratic BSDE associated with f and ξ such that the process Y is bounded and $Z \in H^2$.

The idea is to make an exponential change of variable $y_t = e^{2CY_t}$ and to show the existence of a solution by an approximation method. More precisely, it is possible to show that there exists a nonincreasing sequence of Lipschitz drivers F^p , which converges to F (where F is the driver of the BSDE satisfied by y_t). Then, one can show that the (nonincreasing) sequence y^p of solutions of classical BSDEs associated with F^p converges to a solution y of the BSDE associated with the driver F and terminal condition $e^{2C\xi}$, which gives the desired result.

BSDE for a European Option

Consider a market model with a nonrisky asset, where price per unit $P_0(t)$ at time t satisfies

$$dP_0(t) = P_0(t)r(t)dt \quad (26)$$

and n risky assets, the price of the i th stock $P_i(t)$ is modeled by the linear stochastic differential equation

$$dP_i(t) = P_i(t) \left[b_i(t)dt + \sum_{j=1}^n \sigma_{i,j}(t)dW_t^j \right] \quad (27)$$

driven by a standard n -dimensional Wiener process $W = (W^1, \dots, W^n)'$, defined on a filtered probability space $(\Omega, \mathcal{F}, P)$. We assume the filtration $\mathcal{F}$ generated by the Brownian W is complete. The probability P corresponds to the objective probability measure. The coefficients $r, b_i, \sigma_{i,j}$ are $\mathcal{F}$ -predictable processes. We denote the vector $b := (b^1, \dots, b^n)'$ by b and the volatility matrix $\sigma := (\sigma_{i,j}, 1 \leq i \leq n, 1 \leq j \leq n)$ by σ . We will assume that the matrix σ_t has full rank for any $t \in [0, T]$. Let $\theta_t = (\theta_t^1, \dots, \theta_t^n)'$ be the classical risk-premium vector defined as

$$\theta_t = \sigma^{-1}(b_t - r_t \mathbf{1}) \text{ } P\text{-a.s.} \quad (28)$$

The coefficients σ, b, θ , and r are supposed to be bounded.

Let us consider a small investor, who can invest in the $n + 1$ basic securities. We denote by (X_t) the wealth process. At each time t , he/she chooses the amount $\pi_i(t)$ invested in the i th stock.

More precisely, a portfolio process is an adapted process $\pi = (\pi_1, \dots, \pi_n)'$ with $\int_0^T |\sigma_t' \pi_t|^2 dt < \infty$, P -a.s.

The strategy is supposed to be self-financing, that is, the wealth process satisfies the following dynamics:

$$dX_t^{x,\pi} = r_t X_t dt + \pi_t' \sigma_t (dW_t + \theta_t dt) \quad (29)$$

Generally, the initial wealth $x = X_0$ is taken as a primitive, and for an initial endowment and portfolio process (x, π) , there exists a unique wealth process X , which is the solution of the linear equation (29) with initial condition $X_0 = x$. Therefore, there exists a one-to-one correspondence between pairs (x, π) and trading strategies (X, π) .

Let T be a strictly positive real, which will be the terminal time of our problem. Let ξ be a European

contingent claim settled at time T , that is, an $\mathcal{F}_T$ -measurable square-integrable random variable (it can be thought of as a contract that pays the amount ξ at time T). By a direct application of BSDE results, we derive that there exists a unique P -square-integrable strategy (X, π) such that

$$\begin{aligned} dX_t &= r_t X_t dt + \pi_t' \sigma_t \theta_t dt + \pi_t' \sigma_t dW_t, \\ X_T &= \xi \end{aligned} \quad (30)$$

X_t is the price of claim ξ at time t and (X, π) is a hedging strategy for ξ .

In the case of constraints such as the case of a borrowing interest rate R_t greater than the bond rate r (see [10] p. 201 and 216 or [7]), the case of taxes [8], or the case of a large investor (whose strategy has an influence on prices, see [10] p. 216), the dynamics of the wealth-portfolio strategy is no longer linear. Generally, it can be written as follows:

$$-dX_t = b(t, X_t, \sigma_t' \pi_t) dt - \pi_t' \sigma_t dW_t \quad (31)$$

where b is a driver (the classical case corresponds to the case where $b(t, x, z) = -r_t x - z' \theta_t$).

Let ξ be a square-integrable European contingent claim. BSDE results give the existence and the uniqueness of a P -square-integrable strategy (X, π) such that

$$\begin{aligned} -dX_t &= b(t, X_t, \sigma_t' \pi_t) dt - \pi_t' \sigma_t dW_t, \\ X_T &= \xi \end{aligned} \quad (32)$$

As in the classical case, X_t is the price of the claim ξ at time t and (X, π) is a hedging strategy of ξ . Also note that, under some smoothness assumptions on the driver b , by equality (22), the hedging portfolio process (multiplied by the volatility) $\pi_t' \sigma_t$ corresponds to the Malliavin derivative $D_t X_t$ of the price process, that is,

$$D_t X_t = \sigma_t' \pi_t, \text{ } dP \otimes dt\text{-a.s.} \quad (33)$$

which generalizes (to the nonlinear case) the useful result stated by Karatzas and Ocone [21] in the linear case. Thus, we obtain a nonlinear *price system* (see [10] p. 209), that is, an application that, for each $\xi \in L^2(\mathcal{F}_T)$ and $T \geq 0$, associates an adapted process $(X_t^b(\xi, T))_{0 \leq t \leq T}$, where $X_t^b(\xi, T)$ denotes the solution of the BSDE associated with the driver b , terminal condition ξ , and terminal time T .

By the comparison theorem, this price system is *nondecreasing* with respect to ξ and satisfies the *no-arbitrage* property:

- A1.** If $\xi^1 \geq \xi^2$ and if $X_t^b(\xi^1, T) = X_t^b(\xi^2, T)$ on an event $A \in \mathcal{F}_t$, then $\xi^1 = \xi^2$ on A .

By the flow property of BSDEs (Proposition 2), it is also *consistent*: more precisely, if S is a stopping time (smaller than T), then for each time t smaller than S , the price associated with payoff ξ and maturity T coincides with the price associated with maturity S and payoff $X_S^b(\xi, T)$, that is,

- A2.** $\forall t \leq S, X_t^b(\xi, T) = X_t^b(X_S^b(\xi, T), S)$.

In addition, if $b(t, 0, 0) \geq 0$, then, by the comparison theorem, the price X^b is positive. At least, if b is sublinear with respect to (x, π) (which is generally the case), then, by the comparison theorem, the price system is *sublinear*. Also note that if $b(t, 0, 0) = 0$, then the price of a contingent claim $\xi = 0$ is equal to 0, that is, $X_t^b(0, T) = 0$ and moreover (see, e.g., [25]), the price system satisfies the *zero-one law* property, that is,

- A3.** $X_t(\mathbf{1}_A \xi, T) = \mathbf{1}_A X_t(\xi, T)$ a.s. for $t \leq T$, $A \in \mathcal{F}_t$, and $\xi \in L^2(\mathcal{F}_T)$.

Furthermore, if b does not depend on x , then the price system satisfies the *translation invariance property*:

- A4.** $X_t(\xi + \xi', T) = X_t(\xi, T) + \xi'$, for any $\xi \in L^2(\mathcal{F}_T)$ and $\xi' \in L^2(\mathcal{F}_t)$.

Intuitively, it can be interpreted as a market with interest rate r equal to zero.

In the case where the driver b is convex with respect to (x, π) (which is generally the case), we have a *variational formulation of the price* of a European contingent claim (see [7] or [10] Prop. 3.8 p. 215). Indeed, by classical properties of convex analysis, b can be written as the maximum of a family of affine functions. More precisely, we have

$$b(t, x, \pi) = \sup_{(\beta, \gamma) \in \mathcal{A}} \{b^{\beta, \gamma}(t, x, \pi)\} \quad (34)$$

where $b^{\beta, \gamma}(t, x, \pi) = B(t, \beta_t, \gamma_t) - \beta_t x - \gamma_t' \pi$, where $B(t, \cdot, \cdot)$ is the polar function of b with respect to x, π , that is,

$$B(\omega, t, \beta, \gamma) = \inf_{(x, \pi) \in \mathbb{R} \times \mathbb{R}^n} [b(\omega, t, x, \pi) + \beta_t(\omega)x + \gamma_t'(\omega)\pi] \quad (35)$$

$\mathcal{A}$ is a bounded set of pairs of adapted processes (β, γ) such that $E \int_0^T B(t, \beta_t, \gamma_t)^2 dt < +\infty$. BSDEs' properties give the following variational formulation:

$$X_t^b = \text{ess sup}_{(\beta, \gamma) \in \mathcal{A}} X_t^{\beta, \gamma} \quad (36)$$

where $X^{\beta, \gamma}$ is the solution of the linear BSDE associated with the driver $b^{\beta, \gamma}$ and terminal condition ξ . In other words, $X^{\beta, \gamma}$ is the classical linear price of ξ in a fictitious market with interest rate β and risk-premium γ . The function B can be interpreted as a *cost* function or a *penalty* function (which is equal to 0 in quite a few examples).

An interesting question that follows is “Under what conditions does a nonlinear price system have a BSDE representation?” In 2002, Coquet *et al.* [3] gave the first answer to this question.

Theorem 3 *Let $X(\cdot)$ be a price system, that is, an application that, for each $\xi \in L^2(\mathcal{F}_T)$ and $T \geq 0$, associates an adapted process $(X_t(\xi, T))_{0 \leq t \leq T}$ that is nondecreasing, which satisfies the no-arbitrage property (A1), time consistency (A2), zero-one law (A3), and translation invariance property (A4).*

Suppose that it satisfies the following assumption:

There exists some $\mu > 0$ such that

$X_0(\xi + \xi', T) - X_0(\xi, T) \leq Y_0^\mu(\xi', T)$, for any $\xi \in L^2(\mathcal{F}_T)$ and ξ' a positive random variable $\in L^2(\mathcal{F}_T)$, where $Y_t^\mu(\xi', T)$ is solution of the following BSDE:

$$-dY_t = \mu|Z_t|dt - Z_t dW_t, \quad Y_T = \xi' \quad (37)$$

Then the price system has a BSDE representation, that is, there exists a standard driver $b(t, z)$ that does not depend on x such that $b(t, 0) = 0$ and that is Lipschitz with respect to z with coefficient μ , such that $X(\xi, T)$ corresponds to the solution of the BSDE associated with the terminal time T , driver b , and terminal condition ξ , for any $\xi \in L^2(\mathcal{F}_T)$, $T \geq 0$, that is, $X(\xi, T) = X^b(\xi, T)$.

In this theorem, the existence of the coefficient μ might be interpreted in terms of risk aversion.

Many nonlinear BSDEs also appear in the case of an *incomplete* market (see **Complete Markets**). For example, the superreplication price of a European contingent claim can be obtained as the limit

of a nondecreasing sequence of penalized prices, which are solutions of nonlinear BSDEs [9, 10]. Another example is given by the pricing a European contingent claim *via* exponential utility maximization in an incomplete market. In this case, El Karoui and Rouge [11] have stated that the price of such an option is the solution of a *quadratic* BSDE. More precisely, let us consider a complete market (see **Complete Markets**) [11] that contains n securities, whose (invertible) volatility matrix is denoted by σ_t . Suppose that only the first j securities are available for hedging and their volatility matrix is denoted by σ_t^1 . The utility function is given by $u(x) = -e^{-\gamma x}$, where $\gamma (\geq 0)$ corresponds to the risk-aversion coefficient. Let ξ be a given contingent claim corresponding to an exercise time T ; in other words, ξ is a bounded $\mathcal{F}_T$ -measurable variable. Let $(X_t(\xi, T))$ (also denoted by (X_t)) be the forward price process defined *via* the exponential utility function as in [11]. By Theorem 5.1 in [11], there exists $Z \in H^2(\mathbb{R}^n)$ such that the pair (X, Z) is solution of the quadratic BSDE:

$$\begin{aligned} -dX_t = & \left\{ -(\eta_t + \sigma_t^{-1} v_t^0) \cdot Z_t + \frac{\gamma}{2} |\Pi(Z_t)|^2 \right\} \\ & \times dt - Z_t dW_t, \quad X_T = \xi \end{aligned} \quad (38)$$

where η is the classical relative risk process, v^0 is a given process [11], and $\Pi(z)$ denotes the orthogonal projection of z onto the kernel of σ_t^1 .

Dynamic Risk Measures

In the same way as in the previous section, some dynamic measures of risk can be induced quite simply by BSDEs (note that time-consistent dynamic risk-measures are otherwise very difficult to deal with).

More precisely, let b be a standard driver. We define a *dynamic risk-measure* ρ^b as follows: for each $T \geq 0$ and $\xi \in L^2(\mathcal{F}_T)$, we set

$$\rho_t^b(\xi, T) = X_t^b(-\xi, T) \quad (39)$$

where $(X_t^b(-\xi, T))$ denotes the solution of the BSDE associated with the terminal condition $-\xi$, terminal time T , and driver $b(t, \omega, x, z)$ [25]. Also note that $\rho_t^b(\xi, T) = -X_t^{\bar{b}}(\xi, T)$, where $\bar{b}(t, x, z) = -b(t, -x, -z)$.

Then, by the results of the previous section, the dynamic risk measure ρ^b is *nonincreasing* and satisfies the *no-arbitrage* property (A1). In addition, the risk measure ρ^b is also *consistent*.

If b is superadditive with respect to (x, z) , then the dynamic risk-measure ρ^b is *subadditive*, that is,

$$\text{For any } T \geq 0, \xi, \xi' \in L^2(\mathcal{F}_T), \rho_t^b(\xi + \xi', T) \leq \rho_t^b(\xi, T) + \rho_t^b(\xi', T).$$

If $b(t, 0, 0) = 0$, then ρ^b satisfies *zero-one law* (A3).

In addition, if b does not depend on x , then the measure of risk satisfies the *translation invariance property* (A4).

In addition, if b is positively homogeneous with respect to (x, z) , then the risk measure ρ^b is *positively homogeneous* with respect to ξ , that is, $\rho^b(\lambda\xi, T) = \lambda\rho^b(\xi, T)$, for each real $\lambda \geq 0$, $T \geq 0$, and $\xi \in L^2(\mathcal{F}_T)$.

If b is convex (respectively, concave) with respect to (x, z) , then ρ^b is concave (respectively, convex) with respect to ξ . Furthermore, if b is concave (respectively, convex), we have a variational formulation of the risk measure ρ^b (similar to the one obtained for nonlinear price systems). Note that in the case where b does not depend on x , this dual formulation corresponds to a famous theorem for convex and translation-invariant risk measures [12] and the polar function B corresponds to the penalty function.

Clearly, Theorem 3 can be written in terms of risk measures. Thus, it gives the following interesting result.

Proposition 7 *Let ρ be a dynamic risk measure, that is, an application that, for each $\xi \in L^2(\mathcal{F}_T)$ and $T \geq 0$, associates an adapted process $(\rho_t(\xi, T))_{\{0 \leq t \leq T\}}$. Suppose that ρ is nonincreasing and satisfies assumptions (A1)–(A4) and that there exists some $\mu > 0$ such that $\rho_0(\xi + \xi', T) - \rho_0(\xi, T) \geq -Y_0^\mu(\xi', T)$, for any $\xi \in L^2(\mathcal{F}_T)$ and ξ' a positive random variable $\in L^2(\mathcal{F}_T)$, where $Y_t^\mu(\xi', T)$ is solution of BSDE (37). Then, ρ can be represented by a backward equation, that is, there exists a standard driver $b(t, z)$, which is Lipschitz with respect to z with coefficient μ , such that $\rho = \rho^b$ a.s.*

Relation with Recursive Utility

Another example of BSDEs in finance is given by *recursive utilities* introduced by Duffie and Epstein [5]. Such a utility function associated with

a consumption rate $(c_t, 0 \leq t \leq T)$ corresponds to the solution of BSDE (3) with terminal condition ξ , which can be interpreted as a terminal reward (which can be a function of terminal wealth) and a driver $f(t, c_t, y)$ depending on the consumption rate c_t . The case of a standard utility function corresponds to a linear driver f of the form $f(t, c, y) = u(c) - \beta_t y$, where u is a nondecreasing and concave deterministic function and β corresponds to the discounted rate.

Note that by BSDE results, we may consider a driver f that depends on the variability process Z_t [7]. The generalized recursive utility is then the solution of the BSDE associated with ξ and $f(t, c_t, y, z)$. The standard utility function can be generalized to the following model first introduced by Chen and Epstein [2]:

$$f(t, c, y, z) = u(c) - \beta_t y - K \cdot |z| \quad (40)$$

where $K = (K_1, \dots, K_n)$ and $|z| = (|z^1|, \dots, |z^n|)$. The constants K_i can be interpreted as risk-aversion coefficients (or ambiguity-aversion coefficients).

By the flow property of BSDEs, recursive utility is *consistent*. In addition, by the comparison theorem, if f is concave with respect to (c, y, z) (respectively, nondecreasing with respect to c), then recursive utility is *concave* (respectively, nondecreasing) with respect to c .

In the case where the driver f is concave, we have a *variational formulation* of recursive utility (first stated in [7]) similar to the one obtained for nonlinear convex price systems (see the previous section). Let $F(t, c_t, \cdot, \cdot)$ be the polar function of f with respect to y, z and let $\mathcal{A}(c)$ be the (bounded) set of pairs of adapted processes (β, γ) such that $E \int_0^T F(t, c_t, \beta_t, \gamma_t)^2 dt < +\infty$. Properties on optimization of BSDEs lead us to derive the following *variational formulation*:

$$Y_t = \text{ess} \inf_{(\beta, \gamma) \in \mathcal{A}} Y_t^{\beta, \gamma} \quad (41)$$

where $Y^{\beta, \gamma}$ is the solution of the linear BSDE associated with the driver $f^{\beta, \gamma}(t, c, x, \pi) := F(t, c, \beta_t, \gamma_t) + \beta_t y + \gamma_t' z$ and the terminal condition ξ . Note that $Y^{\beta, \gamma}$ corresponds to a standard utility function evaluated under a discounted rate $-\beta$ and under a probability Q^γ with density with respect to P given by $Z^\gamma(T) = \exp\left(-\int_0^T \gamma_s' dW_s - \frac{1}{2} \int_0^T |\gamma_s|^2 ds\right)$. Indeed,

we have

$$Y_t^{\beta, \gamma} = E_{Q^\gamma} \left[\int_t^T e^{\int_t^s \beta_u du} F(s, c_s, \beta_s, \gamma_s) ds + e^{\int_t^T \beta_u du} Y \middle| \mathcal{F}_t \right] \quad (42)$$

El Karoui *et al.* [8] considered the optimization problem of a recursive utility with nonlinear constraints on the wealth. By using BSDE techniques, the authors state a maximum principle that gives a necessary and sufficient condition of optimality. The variational formulation can also lead to transform the initial problem into a max-min problem, which can be written as a min-max problem under some assumptions.

Reflected BSDEs

Reflected BSDEs have been introduced by El Karoui *et al.* [6]. For a reflected BSDE, the solution is constrained to be greater than a given process called the *obstacle*.

Let S^2 be the set of predictable processes ϕ such that $E(\sup_t |\phi_t|^2) < +\infty$. We are given a couple of standard parameters, that is, a standard driver $f(t, y, z)$ and a process $\{\xi_t, 0 \leq t \leq T\}$ called the *obstacle*, which is supposed to be continuous on $[0, T]$, adapted, belonging to S^2 and satisfying $\lim_{t \rightarrow T} \xi_t \leq \xi_T$.

A solution of the reflected BSDE associated with f and ξ corresponds to a triplet $(Y, Z, K) \in S^2 \times \mathcal{H}^2 \times S^2$ such that

$$\begin{aligned} -dY_t &= f(t, Y_t, Z_t)dt + dK_t - Z_t' dW_t, \\ Y_T &= \xi_T \end{aligned} \quad (43)$$

with $Y_t \geq \xi_t, 0 \leq t \leq T$ and where K is nondecreasing, continuous, adapted process equal to 0 at time 0 such that $\int_0^T (Y_s - \xi_s) dK_s = 0$. The process K can be interpreted as the minimal push, which allows the solution to stay above the obstacle.

We first give a characterization of the solution (first stated by El Karoui and Quenez [10]). For each $t \in [0, T]$, let us denote the set of stopping times by T_t τ such that $\tau \in [t, T]$ a.s.

For each $\tau \in T_t$, we denote by $(X_s(\tau, \xi_\tau), \pi_s(\tau, \xi_\tau), t \leq s \leq \tau)$ the (unique) solution of the

BSDE associated with the terminal time τ , terminal condition ξ_τ , and coefficient f . We easily derive the following property.

Proposition 8 (Characterization). *Suppose that (Y, Z, K) is solution of the reflected BSDE (43). Then, for each $t \in [0, T]$,*

$$Y_t = X_t(D_t, \xi_{D_t}) = \operatorname{ess\,sup}_{\tau \in T_t} X_t(\tau, \xi_\tau) \quad (44)$$

where $D_t = \inf\{u \geq t; Y_u = \xi_u\}$.

Proof By using the fact that $Y_{D_t} = \xi_{D_t}$ and since the process K is constant on $[t, D_t]$, we easily derive that $(Y_s, t \leq s \leq D_t)$ is the solution of the BSDE associated with the terminal time D_t , terminal condition ξ_{D_t} , and coefficient f , that is,

$$Y_t = X_t(D_t, \xi_{D_t}) \quad (45)$$

It remains now to show that $Y_t \geq X_t(\tau, \xi_\tau)$, for each $\tau \in T_t$.

Fix $\tau \in T_t$. On the interval $[t, \tau]$, the pair (Y_s, Z_s) satisfies

$$\begin{aligned} -dY_s &= f(s, Y_s, Z_s)ds + dK_s - Z_s dW_s, \\ Y_\tau &= Y_t \end{aligned} \quad (46)$$

In other words, the pair $(Y_s, Z_s, t \leq s \leq D_t)$ is the solution of BSDE associated with the terminal time τ , terminal condition Y_τ , and coefficient

$$f(s, y, z) + dK_s$$

Since $f(s, y, z) + dK_s \geq f(s, y, z)$ and since $Y_\tau \geq \xi_\tau$, the comparison theorem for BSDEs gives

$$Y_t \geq X_t(\tau, \xi_\tau) \quad (47)$$

and the proof is complete.

Proposition 8 gives the uniqueness of the solution:

Corollary 1 (Uniqueness). *There exists a unique solution of reflected BSDE(43).*

In addition, from Proposition 8 and the comparison theorem for classical BSDEs, we quite naturally derive the following comparison theorem for RBSDEs (see [6] or [18] for a shorter proof).

Proposition 9 (Comparison). *Let ξ^1, ξ^2 be two obstacle processes and let f^1, f^2 be two coefficients.*

Let (Y^1, Z^1, K^1) (respectively, (Y^2, Z^2, K^2)) be a solution of the reflected BSDE (43) for (ξ^1, f^1) (respectively, for (ξ^2, f^2)) and assume that

- $\xi^1 \leq \xi^2$ a.s.
 - $f^1(t, y, z) \leq f^2(t, y, z), \quad t \in [0, T], (y, z) \in \mathbb{R} \times \mathbb{R}^d$.
- Then, $Y_t^1 \leq Y_t^2 \quad \forall t \in [0, T]$ a.s.*

As in the case of classical BSDEs, some *a priori estimations* similar to equations (6) and (7) can be given [6]. From these estimations, we can derive the existence of a solution, that is, the following theorem.

Theorem 4 *There exists a unique solution (Y, Z, K) of RBSDE (43).*

Sketch of the proof. The arguments are the same as in the classical case. The only problem is to show the existence of a solution in the case where the driver f does not depend on y, z . However, this problem is already solved by optimal stopping time theory. Indeed, recall that by Theorem (4), we have Y that is a solution of the RBSDE associated with the driver $f(t)$ and obstacle ξ ; then,

$$\begin{aligned} Y_t &= \operatorname{ess\,sup}_{\tau \in T_t} X(\tau, \xi_\tau) \\ &= \operatorname{ess\,sup}_{\tau \in T_t} E \left[\int_t^\tau f(s) ds + \xi_\tau \middle| \mathcal{F}_t \right] \end{aligned} \quad (48)$$

Thus, to show the existence of a solution, a natural candidate is the process

$$\bar{Y}_t = \operatorname{ess\,sup}_{\tau \in T_t} E \left[\int_t^\tau f(s) ds + \xi_\tau \middle| \mathcal{F}_t \right] \quad (49)$$

Then, by using classical results of the Snell envelope theory, we derive that there exist a nondecreasing continuous process K and an adapted process Z such that $(\bar{Y}, Z, K)$ is the solution of the RBSDE associated with f and ξ .

Remark 2 The existence of a solution of the reflected BSDE can also be derived by an approximation method *via penalization* [6]. Indeed, one can show that the sequence of penalized processes $(Y^n, n \in \mathbb{N})$, defined as the solutions of classical

BSDEs

$$\begin{aligned} -dY_t^n &= f(t, Y_t^n, Z_t^n)dt \\ &\quad + n(Y_t^n - S_t)^- dt - Z_t^n dW_t, \quad Y_T^n = \xi \end{aligned} \quad (50)$$

is nondecreasing (by the comparison theorem) and that it converges a.s. to the solution Y of the reflected BSDE.

In the *Markovian case* [6], that is, in the case where the driver and the obstacle are functions of a state process, we can give an interpretation of the solution of the reflected BSDE in terms of an obstacle problem. More precisely, the framework is the same as in the case of a Markovian BSDE. The state process $X^{t,x}$ follows the dynamics (19). Let $(Y^{t,x}, Z^{t,x}, K^{t,x})$ be the solution of the reflected BSDE:

$$\begin{aligned} -dY_s &= f(s, X_s^{t,x}, Y_s, Z_s)ds + dK_s - Z_s' dW_s, \\ Y_T &= g(X_T^{t,x}) \end{aligned} \quad (51)$$

with $Y_s \geq \xi_s := h(s, X_s^{t,x})$, $t \leq s \leq T$. Moreover, we assume that $h(T, x) \leq g(x)$ for $x \in \mathbb{R}^d$. The functions f, h are deterministic and satisfy

$$h(t, x) \leq K(1 + |x|^p), \quad t \in [0, T], \quad x \in \mathbb{R}^d \quad (52)$$

In this case, if u denotes the function such that $Y_t^{t,x} = u(t, x)$, we have the following theorem.

Theorem 5 *Suppose that the coefficients f, b, σ , and h are jointly continuous with respect to t and x . Then, the function $u(t, x)$ is a viscosity solution of the following obstacle problem:*

$$\begin{aligned} \min((u - h)(t, x), -\partial_t u - \mathcal{L}u - f(t, x, u(t, x), \\ \partial_x u \sigma(t, x))) = 0, \quad u(T, x) = g(x) \end{aligned} \quad (53)$$

Idea of the proof. A first proof [6] can be given by using the approximation of the solution Y of the RBSDE by the increasing sequence Y^n of penalized solutions of BSDEs (50). By the previous results on classical BSDEs in the Markovian case, we know that $Y_t^{n,t,x} = u_n(t, x)$ where u_n is the unique viscosity

solution of a parabolic PDE. Thus, we have that $u_n(t, x) \uparrow u(t, x)$ as $n \rightarrow \infty$ and by using classical techniques of the theory of viscosity solutions, it is possible to show that $u(t, x)$ is a viscosity solution of the obstacle problem (53).

Another proof can be given by directly showing that u is a viscosity solution of the obstacle problem [18].

Under quite standard assumptions on the coefficients, there exists a unique viscosity solution (see **Monotone Schemes**) of the obstacle problem (53) [6]. Generalizations of the previous results have been done on reflected BSDEs. Cvitanic and Karatzas [4] have studied reflected BSDEs with two obstacles and their links with stochastic games. Hamadène *et al.* [15] have studied reflected BSDEs with two obstacles with continuous coefficients. Gegout-Petit and Pardoux [13] have studied reflected BSDEs in a convex domain, Ouknine [22] has studied reflected BSDEs with jumps, and finally Kobylanski *et al.* [18] have studied reflected quadratic RBSDEs.

Reflected BSDEs and Pricing of an American Option under Constraints

In this section, we see how these results can be applied to the problem of evaluation of an American option (see, e.g., [10] Section 5.4). The framework is the one that is described in the previous section (a complete market with nonlinear constraints such as a large investor).

Recall that an American option consists, at time t , in the selection of a stopping time $\nu \geq t$ and (once this exercise time is chosen) of a payoff ξ_ν , where $(\xi_t, 0 \leq t \leq T)$ is a continuous adapted process on $[0, T[$ with $\lim_{t \rightarrow T} \xi_t \leq \xi_T$.

Let ν be a fixed stopping time. Then, from the results on classical BSDEs, there exists a unique pair of square-integrable adapted processes $(X(\nu, \xi_\nu), \pi(\nu, \xi_\nu))$ denoted also by (X^ν, π^ν) , satisfying

$$\begin{aligned} -dX_t^\nu &= b(t, X_t^\nu, \pi_t^\nu)dt - (\pi_t^\nu)' dW_t, \quad X_T^\nu = \xi \end{aligned} \quad (54)$$

(To simplify the presentation, σ_t is assumed to be equal to the identity). $X(\nu, \xi_\nu)$ corresponds to the price of a European option of exercise time ν and payoff ξ_ν .

The price of the American option is then given by a right continuous left limited (RCLL) process Y , satisfying for each t ,

$$Y_t = \operatorname{ess\,sup}_{v \in T_t} X_t(v, \xi_v), \quad P\text{-p.s.} \quad (55)$$

By the previous results, the price $(Y_t, 0 \leq t \leq T)$ corresponds to the solution of a reflected BSDE associated with the coefficient b and obstacle ξ . In other words, there exists a process $\pi \in \mathbb{H}^2$ and K an increasing continuous process such that

$$\begin{aligned} -dY_t &= b(t, Y_t, \pi_t)dt + dK_t - \pi_t' dW_t, \\ Y_T &= \xi_T \end{aligned} \quad (56)$$

with $Y \geq \xi$ and $\int_0^T (Y_t - \xi_t) dK_t = 0$. In addition, the stopping time $D_t = \inf\{s \geq t / Y_s = \xi_s\}$ is optimal, that is,

$$Y_t = \operatorname{ess\,sup}_{v \in T_t} X(v, \xi_v) = X_t(D_t, \xi_{D_t}) \quad (57)$$

Moreover, by the minimality property of the increasing process K , the process Y corresponds to the surreplication price of the option, that is, the smallest price that allows the surreplication of the payoff.

One can also easily state that the price system $\xi \mapsto Y(\xi)$ is nondecreasing and sublinear if b is sublinear with respect to x, π . Note (see [10] p. 239) that the nonarbitrage property holds only in a weak sense: more precisely, let ξ and ξ' be two payoffs and let Y and Y' their associated prices. If $\xi \geq \xi'$ and also $Y_0 = Y'_0$, then $D_0 \leq D'_0$, the payoffs are equal at time D'_0 , and the prices are equal until D'_0 .

In the previous section, we have seen how, in the case where the driver b is convex, one can obtain a variational formulation of the price of a European option. Similarly, one can show that the price of an American option is equal to the *value function of a mixed control problem* [10].

References

- [1] Buckdahn, R. (1993). *Backward Stochastic Differential Equations Driven by a Martingale*. Preprint.
- [2] Chen, Z. & Epstein, L. (1998). *Ambiguity, Risk and Asset Returns in Continuous Time*, working paper 1998, University of Rochester.
- [3] Coquet, F., Hu, Y., Mémin, J. & Peng, S. (2002). Filtration-consistent nonlinear expectations and related g -expectations, *Probability Theory and Related Fields* **123**, 1–27.
- [4] Cvitanic, J. & Karatzas, I. (1996). Backward stochastic differential equations with reflection and Dynkin games, *Annals of Probability* **4**, 2024–2056.
- [5] Duffie, D. & Epstein, L. (1992). Stochastic differential utility, *Econometrica* **60**, 353–394.
- [6] El Karoui, N., Kapoudjian, C., Pardoux, E., Peng, S. & Quenez, M.C. (1997). Reflected solutions of Backward SDE's and related obstacle problems for PDE's, *The Annals of Probability* **25**(2), 702–737.
- [7] El Karoui, N., Peng, S. & Quenez, M.C. (1997). Backward stochastic differential equations in finance, *Mathematical Finance* **7**(1), 1–71.
- [8] El Karoui, N., Peng, S. & Quenez, M.C. (2001). A dynamic maximum principle for the optimization of recursive utilities under constraints, *Annals of Applied Probability* **11**(3), 664–693.
- [9] El Karoui, N. & Quenez, M.C. (1995). Dynamic programming and pricing of a contingent claim in an incomplete market, *SIAM Journal on Control and optimization* **33**(1), 29–66.
- [10] El Karoui, N. & Quenez, M.C. (1996). Non-linear pricing theory and backward stochastic differential equations, in *Financial Mathematics*, Lectures Notes in Mathematics, Bressanone 1656, W.J. Runggaldieredssnm, ed., collection, Springer.
- [11] El Karoui, N. & Rouge, R. (2000). Contingent claim pricing via utility maximization, *Mathematical Finance* **10**(2), 259–276.
- [12] Föllmer, H. & Shied, A. (2004). *Stochastic Finance: An introduction in Discrete Time*, Walter de Gruyter, Berlin.
- [13] Gegout-Petit, A. & Pardoux, E. (1996). Equations différentielles stochastiques rétrogrades réfléchies dans un convexe, *Stochastics and Stochastic Reports* **57**, 111–128.
- [14] Gobet, E. & Labart, C. (2007). Error expansion for the discretization of Backward Stochastic Differential Equations, *Stochastic Processes and their Applications* **10**(2), 259–276.
- [15] Hamadane, S., Lepeltier, J.P. & Matoussi, A. (1997). Double barrier reflected backward SDE's with continuous coefficient, in *Backward Stochastic Differential Equations*, Collection Pitman Research Notes in Mathematics Series 364, N. El Karoui & L. Mazliak, eds, Longman.
- [16] Karatzas, I. & Shreve, S. (1991). *Brownian Motion and Stochastic Calculus*, Springer Verlag.
- [17] Kobylanski, M. (2000). Backward stochastic differential equations and partial differential equations with quadratic growth, *The Annals of Probability* **28**, 558–602.
- [18] Kobylanski, M., Lepeltier, J.P., Quenez, M.C. & Torres, S. (2002). Reflected BSDE with super-linear quadratic coefficient, *Probability and Mathematical Statistics* **22**, Fasc.1, 51–83.

- [19] Lepeltier, J.P. & San Martí, J. (1997). Backward stochastic differential equations with continuous coefficients, *Statistics and Probability Letters* **32**, 425–430.
- [20] Lepeltier, J.P. & San Martín, J. (1998). Existence for BSDE with superlinear-quadratic coefficient, *Stochastic and Stochastic Reports* **63**, 227–240.
- [21] Ocone, D. & Karatzas, I. (1991). A generalized Clark representation formula with application to optimal portfolios, *Stochastics and Stochastic Reports* **34**, 187–220.
- [22] Ouknine, Y. (1998). Reflected backward stochastic differential equation with jumps, *Stochastics and Stochastic Reports* **65**, 111–125.
- [23] Pardoux, P. & Peng, S. (1990). Adapted solution of backward stochastic differential equation, *Systems and Control Letters* **14**, 55–61.
- [24] Pardoux, P. & Peng, S. (1992). Backward stochastic differential equations and Quasilinear parabolic partial differential equations, *Lecture Notes in CIS* **176**, 200–217.
- [25] Peng, S. (2004). *Nonlinear Expectations, Nonlinear Evaluations and Risk Measures*, Lecture Notes in Math., 1856, Springer, Berlin, pp. 165–253.
- [26] Quenez, M.C. (1997). “*Stochastic Control and BSDE’s*”, N. El Karoui & L. Mazliak, eds, Collection Pitman Research Notes in Mathematics Series 364, Longman.

Related Articles

Backward Stochastic Differential Equations; Numerical Methods; Convex Risk Measures; Forward–Backward Stochastic Differential Equations (SDEs); Markov Processes; Martingale Representation Theorem; Mean–Variance Hedging; Recursive Preferences; Stochastic Control; Stochastic Integrals; Superhedging.

MARIE-CLAIRE QUENEZ

Backward Stochastic Differential Equations: Numerical Methods

Nonlinear backward stochastic differential equations (BSDEs) were introduced in 1990 by Pardoux and Peng [34]. The interest in BSDEs comes from their connections with partial differential equations (PDEs) [14, 38]; stochastic control (see **Stochastic Control**); and mathematical finance (see [16, 17], among others). In particular, as shown in [15], BSDEs are a useful tool in the pricing and hedging of European options. In a complete market, the price process Y of ξ is a solution of a BSDE. BSDEs are also useful in quadratic hedging problems in incomplete markets (see **Mean-Variance Hedging**).

The result that there exist unique BSDE equations under the assumption that the generator is locally Lipschitz can be found in [19]. A similar result was obtained in the case when the coefficient is continuous with linear growth [24]. The same authors, Lepeltier and San Martín [23], generalized these results under the assumption that the coefficients have a superlinear quadratic growth. Other extensions of existence and uniqueness of BSDE are dealt with in [20, 25, 30]. Stability of solutions for BSDE have been studied, for example, in [1], where the authors analyze stability under disturbances in the filtration. In [6], the authors show the existence and uniqueness of the solution and the link with integral-PDEs (see **Partial Integro-differential Equations (PIDEs)**). An existence theorem for BSDEs with jumps is presented in [25, 36]. The authors state a theorem for Lipschitz generators proved by fixed point techniques [37].

Since BSDE solutions are explicit in only a few cases, it is natural to search for numerical methods approximating the unique solution of such equations and to know the associated type of convergence. Some methods of approximation have been developed.

A four-step algorithm is proposed in [27] to solve equations of forward-backward type, relating the type of approximation to PDEs theory. On the other hand, in [3], a method of random discretization in time is used where the convergence of the method for the solution (Y, Z) needs regularity

assumptions only, but for simulation studies multiple approximations are needed. See also [10, 13, 28] for forward-backward systems of SDE (FBSDE) solutions, [18] for a regression-based Monte Carlo method, [39] for approximating solutions of BSDEs, and [35] for Monte Carlo valuation of American Options.

On the other hand, in [2, 9, 11, 26] the authors replace Brownian motion by simple random walks in order to define numerical approximations for BSDEs. This technique simplifies the computation of conditional expectations involved at each time step.

A quantization (see **Quantization Methods**) technique was suggested in [4, 5] for the resolution of reflected backward stochastic differential equations (RBSDEs) when the generator f does not depend on the control variable z . This method is based on the approximation of continuous time processes on a finite grid, and requires a further estimation of the transition probabilities on the grid.

In [8], the authors propose a discrete-time approximation for approximations of RBSDEs. The L^p norm of the error is shown to be of the order of the time step. On the other hand, a numerical approximation for a class of RBSDEs based on numerical approximations for BSDE and approximations given in [29], can be found in [31, 33].

Recently, work on numerical schemes for jumps is given in [22] and is based on the approximation for the Brownian motion and a Poisson process by two simple random walks. Finally, for decoupled FBSDEs with jumps a numerical scheme is proposed in [7].

Let $\Omega = \mathcal{C}([0, 1], \mathbb{R}^d)$ and consider the canonical Wiener space $(\Omega, \mathcal{F}, \mathbb{P}, \mathcal{F}_t)$, in which $B_t(\omega) = \omega(t)$ is a standard d -dimensional Brownian motion. We consider the following BSDE:

$$Y_t = \xi + \int_t^T f(s, Y_s, Z_s) ds - \int_t^T Z_s dB_s \quad (1)$$

where ξ is a $\mathcal{F}_T$ -measurable square integrable random variable and f is Lipschitz continuous in the space variable with Lipschitz constant L . The solution of equation (1) is a pair of adapted processes (Y, Z) , which satisfies the equation.

Numerical Methods for BSDEs

One approach for a numerical scheme for solving BSDEs is based upon a discretization of the equation

(1) by replacing B with a simple random walk. To be more precise, let us consider the symmetric random walk W^n :

$$W_t^n := \frac{1}{\sqrt{n}} \sum_{k=0}^{c_n(t)} \xi_k^n, \quad 0 \leq t \leq T \quad (2)$$

where $\{\xi_k^n\}_{1 \leq k \leq n}$ is an i.i.d. Bernoulli symmetric sequence. We define $\mathcal{G}_k^n := \sigma(\xi_1^n, \dots, \xi_k^n)$. Throughout this section $c_n(t) = \lfloor nt \rfloor / n$, and ξ^n denotes a square integrable random variable, measurable w.r.t. $\mathcal{G}_n^n$ that should converge to ξ . We assume that W^n and B are defined in the same probability space.

In [26], the authors consider the case when the generator depends only on the variable Y , which makes the analysis simpler. In this situation, the BSDE (1) is given by

$$Y_t = \xi + \int_t^T f(Y_s) ds - \int_t^T Z_s dB_s \quad (3)$$

whose solution is given by

$$Y_t = \mathbb{E} \left(\xi + \int_t^T f(Y_s) ds \middle| \mathcal{F}_t \right) \quad (4)$$

which can be discretized in time with step-size $h = T/n$ by solving a discrete BSDE given by

$$Y_{t_i}^n = \xi^n + \frac{1}{n} \sum_{j=i}^n f(Y_{t_j}^n) - \sum_{j=i}^{n-1} Z_{t_j}^n \Delta W_{t_{j+1}}^n \quad (5)$$

This equation has a unique solution (Y_t^n, Z_t^n) since the martingale W^n has the predictable representation property. It can be checked that solving this equation is equivalent to finding a solution to the following implicit iteration problem:

$$Y_{t_i}^n = \mathbb{E} \left\{ Y_{t_{i+1}}^n + \frac{1}{n} f(Y_{t_i}^n) \middle| \mathcal{G}_i^n \right\} \quad (6)$$

which, due to the adaptedness condition, is equivalent to

$$Y_{t_i}^n - \frac{1}{n} f(Y_{t_i}^n) = \mathbb{E} \left\{ Y_{t_{i+1}}^n \middle| \mathcal{G}_i^n \right\} \quad (7)$$

Furthermore, once $Y_{t_{i+1}}^n$ is determined, $Y_{t_i}^n$ is solved via equation (7) by a fixed point technique:

$$\begin{cases} X^0 = \mathbb{E} \left\{ Y_{t_{i+1}}^n \middle| \mathcal{G}_i^n \right\} \\ X^1 = X^0 + \frac{1}{n} f(X^0) \end{cases} \quad (8)$$

It is standard to show that if f is uniformly Lipschitz in the spatial variable x with Lipschitz constant L (we also assume that f is bounded by R), then the iterations of this procedure will converge to the true solution of equation (7) at a geometric rate L/n . Therefore, in the case where n is large enough, one iteration would already give us the error estimate: $|Y_{t_i}^n - X^1| \leq \frac{LR}{n^2}$, producing a good approximate solution of equation (7). Consequently, the explicit numerical scheme is given by

$$\begin{cases} \hat{Y}_T^n = \xi^n, \hat{Z}_T^n = 0 \\ X_{t_i}^n = \mathbb{E} \left\{ \hat{Y}_{t_{i+1}}^n \middle| \mathcal{G}_i^n \right\} \\ \hat{Y}_{t_i}^n = X_{t_i}^n + \frac{1}{n} f(X_{t_i}^n) \\ \hat{Z}_{t_i}^n = \mathbb{E} \left\{ \left[\hat{Y}_{t_{i+1}}^n + \frac{1}{n} f(\hat{Y}_{t_i}^n) - \hat{Y}_{t_i}^n \right] (\Delta W_{t_{i+1}}^n)^{-1} \middle| \mathcal{G}_i^n \right\} \end{cases} \quad (9)$$

The convergence of $\hat{Y}^n$ to Y is proved in the sense of the Skorohod topology in [9, 26]. In [11], the convergence of the sequence Y^n is established using the tool of convergence of filtrations. See also [3] for the case where f depends on both variables y and z .

Application to European Options

In the Black–Scholes model (see **Black–Scholes Formula**)

$$dS_t = \mu S_t dt + \sigma S_t dB_t \quad (10)$$

which is the continuous version of

$$\frac{S_{t+\Delta t} - S_t}{S_t} \approx \mu \Delta t + \sigma \Delta B_t \quad (11)$$

where the relative return has linear growth plus a random perturbation. σ is called the *volatility* and it is a measure of uncertainty. In this particular case, S has an explicit solution given by the Doleans–Dade exponential

$$S_t = S_0 e^{(\mu - \frac{1}{2}\sigma^2 t) + \sigma B_t} \quad (12)$$

We assume the existence of a riskless asset whose evolution is given by $\beta_t = \beta_0 e^{rt}$, where r is a constant interest rate. Then β satisfies the ODE:

$$\beta_t = \beta_0 + r \int_0^t \beta_s ds \quad (13)$$

A portfolio is a pair of adapted processes (a_t, b_t) that represent the amount of investment in both assets at time t (both can be positive or negative). The wealth process is then given by

$$Y_t = a_t S_t + b_t \beta_t \quad (14)$$

We assume Y is self-financing:

$$dY_t = a_t dS_t + b_t d\beta_t \quad (15)$$

A call option gives the holder the right to buy an agreed quantity of a particular commodity S at a certain time (the expiration date, T) for a certain price (the strike price K). The holder has to pay a fee (called a premium q) for this right. If the option can be exercised only at T , the option is called *European*. If it can be exercised at any time before T , it is called *American*. The main question is, what is the right price for an option? Mathematically, q is determined by the existence of a replication strategy with the initial value q and final value $(S_T - K)^+$; that is, find (a_t, b_t) such that

$$Y_t = a_t S_t + b_t \beta_t \quad Y_T = (S_T - K)^+ \quad Y_0 = q \quad (16)$$

We look for a solution to this problem of the form $Y_t = w(t, S_t)$ with $w(T, x) = (x - K)^+$. Using Itô's formula, we get

$$\begin{aligned} Y_t &= Y_0 + \int_0^t \frac{\partial w}{\partial x} dS_s + \int_0^t \frac{\partial^2 w}{\partial x^2} d[S, S]_s \\ &+ \int_0^t \frac{\partial w}{\partial t} ds = Y_0 + \int_0^t \frac{\partial w}{\partial x} \{ \mu S_s ds + \sigma S_s dB_s \} \\ &+ \int_0^t \frac{1}{2} \frac{\partial^2 w}{\partial x^2} \sigma^2 S_s^2 ds \\ &+ \int_0^t \frac{\partial w}{\partial t} ds = Y_0 + \int_0^t \frac{\partial w}{\partial x} \sigma S_s dB_s \\ &+ \int_0^t \left(\frac{1}{2} \frac{\partial^2 w}{\partial x^2} \sigma^2 S_s^2 + \mu S_s \frac{\partial w}{\partial x} + \frac{\partial w}{\partial t} \right) ds \end{aligned} \quad (17)$$

Using the self-financing property, we obtain

$$\begin{aligned} Y_t &= Y_0 + \int_0^t a_s dS_s + \int_0^t b_s d\beta_s = Y_0 + \int_0^t a_s \{ \mu S_s ds \\ &+ \sigma S_s dB_s \} + \int_0^t b_s d\beta_s = Y_0 + \int_0^t a_s \sigma S_s dB_s \end{aligned}$$

$$+ \int_0^t (r b_s \beta_s + a_s \mu S_s) ds \quad (18)$$

Using the uniqueness in the predictable representation property for Brownian motion (see **Martingale Representation Theorem**), we obtain that

$$\begin{aligned} a_s \sigma S_s &= \sigma S_s \frac{\partial w}{\partial x} \\ r b_s \beta_s + a_s \mu S_s &= \frac{1}{2} \sigma^2 S_s^2 \frac{\partial^2 w}{\partial x^2} + \mu S_s \frac{\partial w}{\partial x} + \frac{\partial w}{\partial t} \\ a_s &= \frac{\partial w}{\partial x}(s, S_s) \\ b_s &= \frac{Y_s - a_s S_s}{\beta_s} \end{aligned} \quad (19)$$

Since $r \frac{(Y_s - a_s S_s)}{\beta_s} \beta_s + a_s \mu S_s = \frac{1}{2} \sigma^2 S_s^2 \frac{\partial^2 w}{\partial x^2} + \mu S_s \frac{\partial w}{\partial x} + \frac{\partial w}{\partial t}$, the equation for w is

$$\begin{aligned} r \frac{\partial w}{\partial t} + \frac{1}{2} \sigma^2 x^2 \frac{\partial^2 w}{\partial x^2} &= -r x \frac{\partial w}{\partial x} + r w \\ w(T, x) &= (x - K)^+ \end{aligned} \quad (20)$$

The solution of this PDE is related to a BSDE, which we deduce now. Let us start again from the self-financing assumption

$$\begin{aligned} (S_T - K)^+ &= Y_T = Y_t + \int_t^T \frac{\partial w}{\partial x} dS_s \\ &+ \int_t^T r \left(Y_s - S_s \frac{\partial w}{\partial x} \right) ds = Y_t \\ &+ \int_t^T \sigma S_s \frac{\partial w}{\partial x} dB_s \\ &+ \int_t^T \left(r Y_s + (\mu - r) S_s \frac{\partial w}{\partial x} \right) ds \end{aligned} \quad (21)$$

from which we deduce

$$Y_t = \xi + \int_t^T (\alpha Z_s - r Y_s) ds - \int_t^T Z_s dB_s \quad (22)$$

with $\alpha = \frac{r-\mu}{\sigma}$, $\xi = (S_0 e^{(\mu - \frac{1}{2}\sigma^2 T) + \sigma B_T} - K)^+$, and $Z_s = \sigma S_s \frac{\partial w}{\partial x}$. In this case, we have an explicit solution for w given by

$$Y_0 = S_0 \Phi(g(T, S_0)) - K e^{-rT} \Phi(h(T, S_0)) \quad (23)$$

$$w(t, x) = x \Phi(g(T-t, x)) - K e^{-r(T-t)} \Phi(h(T-t, x)) \quad (24)$$

where $g(t, x) = \frac{\ln(x/K) + (r+1/2\sigma^2)t}{\sigma\sqrt{t}}$, $h(t, x) = g(t, x) - \sigma\sqrt{t}$ and $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy$ is the standard normal distribution. In general, for example, when σ may depend on time and (S_t) , we obtain a BSDE for (Y_t) coupled with a forward equation for (S_t) , that can be solved numerically.

Numerical Methods for RBSDEs

In this section, we are interested in the numerical approximation of BSDEs with reflection (in short, RBSDEs). We present here the case of one lower barrier, which we assume is an Itô process (a sum of a Brownian martingale and a continuous finite variation process).

$$\begin{aligned} Y_t &= \xi + \int_t^T f(s, Y_s, Z_s) ds \\ &\quad - \int_t^T Z_s dB_s + K_T - K_t \quad 0 \leq t \leq T \quad (25) \\ Y_t &\geq L_t, \quad 0 \leq t \leq T, \quad \text{and} \quad \int_0^T (Y_t - L_t) dK_t = 0 \quad (26) \end{aligned}$$

where, as before, f is the generator, ξ is the terminal condition, and $L = (L_t)$ is the reflecting barrier. Under the Lipschitz assumption of f (see [14] and for generalizations see [12, 21, 32]) there is a unique solution (Y, Z, K) of adapted processes, with the condition that K is increasing and minimal in the sense that it is supported at the times Y touches the boundary.

The numerical scheme for RBSDEs that we present here is based on a penalization of equation

(26) [14] coupled with a use of the standard Euler scheme. The penalization equation is given by

$$\begin{aligned} Y_t^\varepsilon &= \xi + \int_t^1 f(s, Y_s^\varepsilon, Z_s^\varepsilon) ds \\ &\quad - \int_t^1 Z_s^\varepsilon dB_s + \frac{1}{\varepsilon} \int_t^1 (L_s - Y_s^\varepsilon)^+ ds \end{aligned} \quad (27)$$

In this framework, we define

$$K_t^\varepsilon := \frac{1}{\varepsilon} \int_0^t (L_s - Y_s^\varepsilon)^+ ds, \quad 0 \leq t \leq 1 \quad (28)$$

where ε is the penalization parameter. In order to have an explicit iteration, we include an extra Picard iteration, and the numerical procedure is then

$$\begin{aligned} Y_{t_i}^{\varepsilon, p+1, n} &= Y_{t_{i+1}}^{\varepsilon, p+1, n} + \frac{1}{n} f(t_i, Y_{t_i}^{\varepsilon, p, n}, Z_{t_i}^{\varepsilon, p, n}) \\ &\quad + \frac{1}{n\varepsilon} (L_{t_i} - Y_{t_i}^{\varepsilon, p, n})^+ - \frac{1}{\sqrt{n}} Z_{t_i}^{\varepsilon, p+1, n} \zeta_{i+1} \end{aligned} \quad (29)$$

$$\begin{aligned} K_{t_{i+1}}^{\varepsilon, p+1, n} - K_{t_i}^{\varepsilon, p+1, n} &:= \frac{1}{n\varepsilon} \left(S - \check{Y}_{t_i}^{\varepsilon, p+1, n} \right)^+ \\ \text{for } i &\in \{n-1, \dots, 0\} \end{aligned} \quad (30)$$

Theorem 1 *Under the assumptions*

- A1.** f is Lipschitz continuous and bounded;
- A2.** L is assumed to be an Itô process;
- A3.** $\lim_{n \rightarrow +\infty} \mathbb{E} \left[\sup_{s \in [0, T]} |\mathbb{E}[\xi | \mathcal{F}_s] - \mathbb{E}[\xi^n | \mathcal{G}_{c_n(s)}^n]| \right] = 0$

the triplet $(\xi^n, Y^{\varepsilon, p, n}, Z^{\varepsilon, p, n}, K^{\varepsilon, p, n})$ converges in the Skorohod topology toward the solution (ξ, Y, Z, K) of the RBSDE (26) (the order is first $p \rightarrow \infty$, then $n \rightarrow \infty$ and finally $\varepsilon \rightarrow 0$).

A Procedure Based on Ma and Zhang's Method

We now introduce a numerical scheme based on a suggestion given in [29]. The new ingredient is to use a standard BSDE with no reflection and then

impose in the final condition of every step of the discretization that the solution must be above the barrier. Schematically we have

- $Y_1^n := \xi^n$
- for $i = n, n-1, \dots, 1$ let $(\tilde{Y}^n, Z^n)$ be the solution of the BSDE:

$$\tilde{Y}_{t_{i+1}}^n = Y_{t_i}^n + \frac{1}{n} f(s, \tilde{Y}_s^n, Z_s^n) - Z_s^n (W_{t_{i+1}}^n - W_{t_i}^n) \quad (31)$$

- define $Y_{t_{i+1}}^n = \tilde{Y}_{t_{i+1}}^n \vee L_{t_{i+1}}$
- let $K_0^n = 0$ and define $K_{t_i}^n := \sum_{j=1}^i (Y_{t_j}^n - \tilde{Y}_{t_j}^n)$

Clearly K^n is predictable and we have

$$Y_{t_{i-1}}^n = Y_{t_i}^n + \int_{t_{i-1}}^{t_i} f(s, \tilde{Y}_s^n, Z_s^n) ds - \int_{t_{i-1}}^{t_i} Z_s^n dW_s^n + K_{t_i}^n - K_{t_{i-1}}^n \quad (32)$$

Theorem 2 Under the assumptions **A1**, **A2** of Theorem 1 and

$$\lim_{n \rightarrow +\infty} \mathbb{E} \left[\sup_{s \in [0, T]} |\mathbb{E}[\xi | \mathcal{F}_s] - \mathbb{E}[\xi^n | \mathcal{G}_{c_n(s)}^n]| \right]^2 = 0 \quad (33)$$

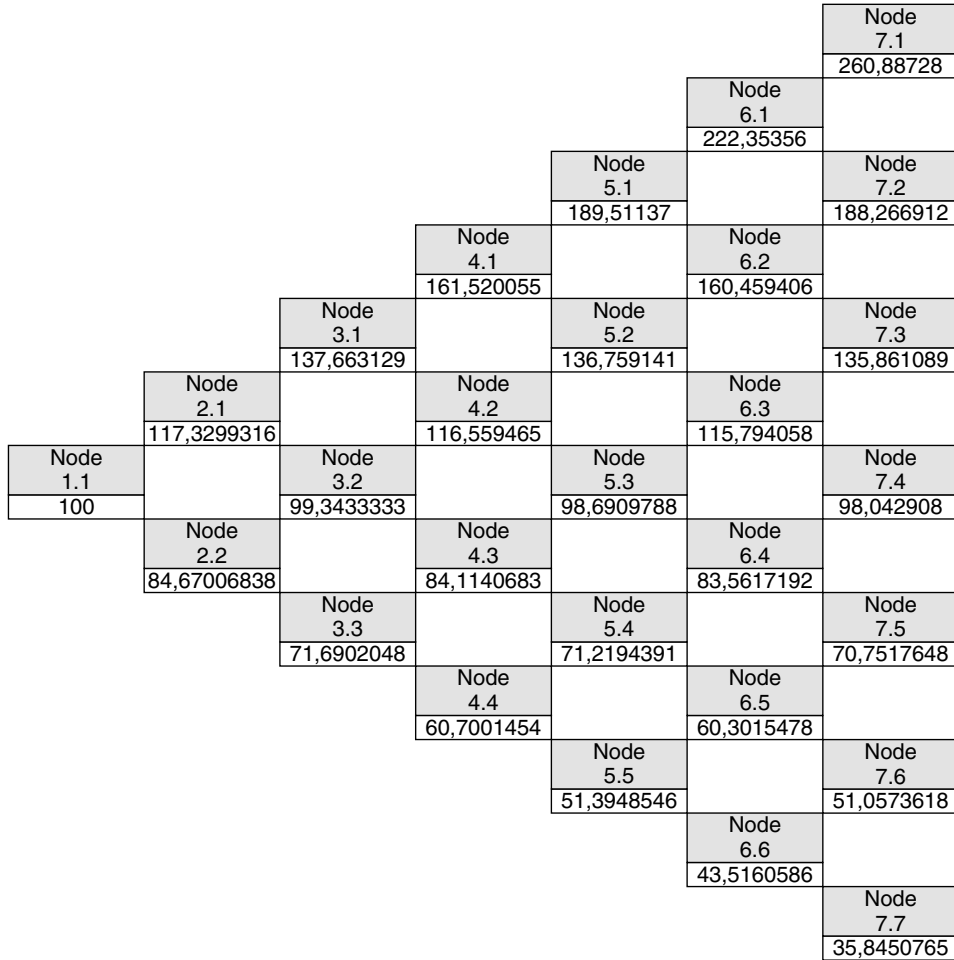


Figure 1 Binomial tree for six time steps, $r = 0.06$, $\sigma = 0.4$, and $T = 0.5$

we have

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[\sup_{0 \leq i \leq n} |Y_{t_i} - Y_{t_i}^n|^2 + \int_0^1 |Z_t - Z_t^n|^2 dt \right] = 0 \quad (34)$$

Application to American Options

An American option (see **American Options**) is one that can be exercised at any time between the purchase date and the expiration date T , which we assume is nonrandom and for the sake of simplicity we take $T = 1$. This situation is more general than the European-style option, which can only be exercised on the date of expiration. Since an American option provides an investor with a greater degree of flexibility, the premium for this option should be higher than the premium for a European-style option.

We consider a financial market described by a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_{0 \leq t \leq T}, \mathbb{P})$. As above, we consider the following adapted processes: the price of the risk asset $S = (S_t)_{0 \leq t \leq T}$ and the wealth process $Y = (Y_t)_{0 \leq t \leq T}$. We assume that the rate interest r is constant. The aim is to obtain Y_0 , the value of the American Option.

We assume that there exists a risk-neutral measure (see **Equivalent Martingale Measures**) allowing one to compute prices of all contingent claims as the expected value of their discounted cash flows. The equation that describes the evolution of Y is given by a linear reflected BSDE coupled with the forward equation for S .

$$Y_t = (K - S_t)^+ - \int_t^1 (rY_s + (\mu - r)Z_s) ds + K_1 - K_t - \int_t^1 Z_s dB_s \quad (35)$$

$$S_t = S_0 + \int_0^t \mu S_s ds + \int_0^t \sigma S_s dB_s \quad (36)$$

The increasing process K keeps the process Y above the barrier $L_t = (S_t - K)^+$ (for a call option) in a minimal way, that is, $Y_t \geq L_t$, $dK_t \geq 0$, and

$$\int_0^1 (Y_t - L_t) dK_t = 0 \quad (37)$$

Table 1 Numerical scheme for an American option with 18 steps, $K = 100$, $r = 0.06$, $\sigma = 0.4$, and $T = 0.5$, and different values of S_0

n	$S_0 = 80$	$S_0 = 100$	$S_0 = 120$
1	20	11.2773	4.1187
2	22.1952	10.0171	3.8841
3	21.8707	10.7979	3.1489
4	22.8245	10.1496	3.9042
$\vdots$	$\vdots$	$\vdots$	$\vdots$
15	22.6775	10.8116	3.7119
16	22.6068	10.6171	3.6070
17	22.7144	10.7798	3.6811
18	22.6271	10.6125	3.6364
Real values	21.6059	9.9458	4.0611

The exercise random time is given by the following stopping time $\tau = \inf\{t : Y_t - L_t < 0\}$ that represents the exit time from the market for the investor. As usual, we take $\tau = 1$ if Y never touches the boundary L . At τ the investor will buy the stock if $\tau < 1$, otherwise he/she does not exercise the option. In this problem, we are interested in finding Y_t , Z_t , and τ .

In Table 1 and Figure 1, we summarize the results of a simulation for the American option.

Acknowledgments

Jaime San Martín's research is supported by Nucleus Millennium Information and Randomness P04-069-F and BASAL project. Soledad Torres' research is supported by PBCT-ACT 13 Stochastic Analysis Laboratory, Chile.

References

- [1] Antonelli, F. (1996). Stability of backward stochastic differential equations, *Stochastic Processes and Their Applications* **62**(1), 103–114.
- [2] Antonelli, F. & Kohatsu-Higa, A. (2000). Filtration stability of backward SDE's, *Stochastic Analysis and Applications* **18**(1), 11–37.
- [3] Bally, V. (1997). *Approximation Scheme for Solutions of BSDE. Backward Stochastic Differential Equations*. (Paris, 1995–1996), Pitman Research Notes Mathematics Series, Longman, Harlow, Vol. 364, pp. 177–191.
- [4] Bally, V. & Pagès, G. (2003). A quantization algorithm for solving multi-dimensional discrete-time optimal stopping problems, *Bernoulli* **9**(6), 1003–1049.
- [5] Bally, V., Pagès, G. & Printems, J. (2001). *A Stochastic Quantization Method for Nonlinear Problems*. Monte

- Carlo and Probabilistic Methods for Partial Differential Equations* (Monte Carlo, 2000). Monte Carlo Methods and Applications 7 (no. 1–2), pp. 21–33.
- [6] Barles, G., Buckdahn, R. & Pardoux, E. (1997). BSDEs and integral-partial differential equations, *Stochastics and Stochastics Reports* **60**(1–2), 57–83.
- [7] Bouchard, B. & Elie, R. (2005). Discrete time approximation of decoupled forward-backward SDE with jumps. *Stochastic Processes and Their Applications* **118**(1), 53–75.
- [8] Bouchard, B. & Touzi, N. (2004). Discrete-time approximation and Monte-Carlo simulation of backward stochastic differential equations, *Stochastic Processes and Their Applications* **111**(2), 175–206.
- [9] Briand, P., Delyon, B. & Mémin, J. (2001). Donsker-Type theorem for BSDEs, *Electronic Communications in Probability* **6**, 1–14.
- [10] Chevance, D. (1997). *Numerical Methods for Backward Stochastic Differential Equations. Numerical Methods in Finance*, Publications of the Newton Institute, Cambridge University Press, Cambridge, pp. 232–244.
- [11] Coquet, F., Mémin, J. & Slomiński, L. (2001). *On Weak Convergence of Filtrations*, Séminaire de Probabilités, XXXV, Lecture Notes in Mathematics, Springer, Berlin, Vol. 1755, pp. 306–328.
- [12] Cvitanic, J. & Karatzas, I. (1996). Backward stochastic differential equations with reflections and Dynkin games, *Annals of Probability* **24**, 2024–2056.
- [13] Douglas, J., Ma, J. & Protter, P. (1996). Numerical methods for forward-backward stochastic differential equations, *Annals of Applied Probability* **6**(3), 940–968.
- [14] El Karoui, N., Kapoudjian, C., Pardoux, E. & Quenez, M.C. (1997). Reflected solutions of backward SDE's, and related obstacle problems for PDE's, *Annals of Probability* **25**(2), 702–737.
- [15] El Karoui, N., Peng, S. & Quenez, M.C. (1997). Backward stochastic differential equations in finance, *Mathematical Finance* **7**, 1–71.
- [16] El Karoui, N. & Quenez, M.C. (1997). *Imperfect Markets and Backward Stochastic Differential Equation. Numerical Methods in Finance*, publications of the Newton Institute, Cambridge University Press, Cambridge, pp. 181–214.
- [17] El Karoui, N. & Rouge, R. (2000). Contingent claim pricing via utility maximization, *Mathematical Finance* **10**(2), 259–276.
- [18] Gobet, E., Lemor, J.-P. & Warin, X. (2005). A regression-based Monte Carlo method to solve backward stochastic differential equations, *Annals of Applied Probability* **15**(3), 2172–2202.
- [19] Hamadene, S. (1996). Équations différentielles stochastiques rétrogrades: les cas localement Lipschitziens, *Annales de l'institut Henri Poincaré (B) Probabilités et Statistiques* **32**(5), 645–659.
- [20] Kobylanski, M. (2000). Backward stochastic differential equations and partial differential equations with quadratic growth, *Annals of Probability* **28**, 558–602.
- [21] Kobylanski, M., Lepeltier, J.P., Quenez, M.C. & Torres, S. (2002). Reflected BSDE with Superlinear quadratic coefficient, *Probability and Mathematical Statistics* **22**(Fasc. 1), 51–83.
- [22] Lejay, A., Mordecki, E. & Torres, S. (2008). Numerical method for backward stochastic differential equations with jumps. Submitted, preprint inria-00357992.
- [23] Lepeltier, J.P. & San Martín, J. (1997). Backward stochastic differential equations with continuous coefficient, *Statistics and Probability Letters* **32**(4), 425–430.
- [24] Lepeltier, J.P. & San Martín, J. (1998). Existence for BSDE with superlinear-quadratic coefficients, *Stochastics Stochastics Reports* **63**, 227–240.
- [25] Li, X. & Tang, S. (1994). Necessary condition for optimal control of stochastic systems with random jumps, *SIAM Journal on Control and Optimization* **33**(5), 1447–1475.
- [26] Ma, J., Protter, P., San Martín, J. & Torres, S. (2002). Numerical method for backward stochastic differential equations, *Annals of Applied Probability* **12**, 302–316.
- [27] Ma, J., Protter, P. & Yong, J. (1994). Solving forward-backward stochastic differential equations explicitly a four step scheme, *Probability Theory and Related Fields* **98**(3), 339–359.
- [28] Ma, J. & Yong, J. (1999). *Forward-Backward Stochastic Differential Equations and their Applications*. Lecture notes in Mathematics, Springer Verlag, Berlin, p. 1702.
- [29] Ma, J. & Zhang, L. (2005). Representations and regularities for solutions to bsde's with reflections, *Stochastic Processes and their Applications* **115**, 539–569.
- [30] Mao, X.R. (1995). Adapted Solutions of BSDE with Non-Lipschitz coefficients, *Stochastic Processes and their Applications* **58**, 281–292.
- [31] Martínez, M., San Martín, J. & Torres, S. *Numerical method for Reflected Backward Stochastic Differential Equations*. Submitted.
- [32] Matoussi, A. (1997). Reflected solutions of backward stochastic differential equations with continuous coefficient, *Statistics and Probability Letters* **34**, 347–354.
- [33] Mémin, J., Peng, S. & Xu, M. (2008). Convergence of solutions of discrete reflected backward SDE's and simulations, *Acta Mathematicae Applicatae Sinica* **24**(1), 1–18.
- [34] Pardoux, P. & Peng, S. (1990). Adapted solution of backward stochastic differential equation, *Systems and Control Letters* **14**, 55–61.
- [35] Rogers, L.C.G. (2002). Monte Carlo valuation of American options, *Mathematical Finance* **12**(3), 271–286.
- [36] Situ, R. (1997). On solution of backward stochastic differential equations with jumps, *Stochastic Processes and their Applications* **66**(2), 209–236.
- [37] Situ, R. & Yin, J. (2003). On solutions of forward-backward stochastic differential equations with Poisson jumps, *Stochastic Analysis and Applications* **21**(6), 1419–1448.

- [38] Sow, A.B. & Pardoux, E. (2004). Probabilistic interpretation of a system of quasilinear parabolic PDEs, *Stochastics and Stochastics Reports* **76**(5), 429–477.
- [39] Zhang, J. (2004). A numerical scheme for BSDEs, *Annals of Applied Probability* **14**(1), 459–488.

Related Articles

American Options; Backward Stochastic Differential Equations; Forward–Backward Stochastic

Differential Equations (SDEs); Markov Processes; Martingales; Martingale Representation Theorem; Mean–Variance Hedging; Partial Differential Equations; Partial Integro-differential Equations (PIDEs); Quantization Methods; Stochastic Control.

JAIME SAN MARTÍN & SOLEDAD TORRES

Stochastic Exponential

Let X be a semimartingale with $X_0 = 0$. Then there exists a unique semimartingale Z that satisfies the equation

$$Z_t = 1 + \int_0^t Z_{s-} dX_s \quad (1)$$

It is called the *stochastic exponential* of X and is denoted by $\mathcal{E}(X)$. Sometimes the stochastic exponential is also called the *Doléans exponential*, after the French mathematician Catherine Doléans-Dade. Note that Z_- denotes the left-limit process, so that the integrand in the stochastic integral is predictable.

We first give some examples as follows:

1. If B is a Brownian motion, then an application of Itô's formula reveals that

$$\mathcal{E}(B)_t = \exp\left(B_t - \frac{1}{2}t\right) \quad (2)$$

2. Likewise, the stochastic exponential for a compensated Poisson process $N - \lambda t$ is given as

$$\begin{aligned} \mathcal{E}(N - \lambda t)_t &= \exp\left(-\frac{1}{2}\lambda t\right) \times 2^{N_t} \\ &= \exp\left(\ln(2)N_t - \frac{1}{2}\lambda t\right) \end{aligned} \quad (3)$$

3. The classical Samuelson model for the evolution of stock prices is also given as a stochastic exponential. The price process S is modeled here as the solution of the stochastic differential equation

$$\frac{dS_t}{S_t} = \sigma dB_t + \mu dt \quad (4)$$

Here, we consider the constant trend coefficient μ , the volatility σ , and a Brownian motion B . The solution to this equation is

$$\begin{aligned} S_t &= \mathcal{E}(\sigma B_t + \mu t) \\ &= \exp\left(\sigma B_t + \left(\mu - \frac{1}{2}\sigma^2\right)t\right) \end{aligned} \quad (5)$$

For a general semimartingale X as above, the expression for the stochastic exponential is

$$\begin{aligned} Z_t &= \exp\left(X_t - \frac{1}{2}[X]_t\right) \prod_{0 < s \leq t} (1 + \Delta X_s) \\ &\quad \times \exp\left(-\Delta X_s + \frac{1}{2}(\Delta X_s)^2\right) \end{aligned} \quad (6)$$

where the possibly infinite product converges. Here $[X]$ denotes the *quadratic variation* process of X .

In case X is a local martingale vanishing at zero with $\Delta X > -1$, then $\mathcal{E}(X)$ is a strictly positive local martingale. This property renders the stochastic exponential very useful as a model for asset prices in case the price process is directly modeled under a martingale measure, that is, in the risk neutral world. However, considering some Lévy-process X , many authors prefer to model the price process as $\exp(X)$ rather than $\mathcal{E}(X)$ since this form is better suited for applying Laplace transform methods. In fact, the two representations are equivalent because starting with a model of the form $\exp(X)$, one can always find a Lévy-process $\hat{X}$ such that $\exp(X) = \mathcal{E}(\hat{X})$ and *vice versa* (in case the stochastic exponential is positive). The detailed calculations involving characteristic triplets can be found in Goll and Kallsen [3].

Finally, for any two semimartingales X, Y we have the formula

$$\mathcal{E}(X)\mathcal{E}(Y) = \mathcal{E}(X + Y + [X, Y]) \quad (7)$$

which generalizes the multiplicative property of the usual exponential function.

Martingale Property

The most crucial issue from the point of mathematical finance is that, given X is a local martingale, the stochastic exponential $\mathcal{E}(X)$ may fail to be a martingale. Let us give an illustration of this phenomenon.

We assume that the price process of a risky asset evolves as the stochastic exponential $Z_t = \exp(B_t - \frac{1}{2}t)$ where B is a standard Brownian motion starting in zero. Since one-dimensional Brownian motion is almost-surely recurrent, and therefore gets negative for arbitrarily large times, zero must be an accumulation point of Z . As Z can be written as a stochastic integral of B , it is a local martingale, and hence a supermartingale by Fatou's lemma

because it is bounded from below. We conclude by the supermartingale convergence theorem that Z converges (necessarily to zero). This shows that

$$\lim_{t \rightarrow \infty} Z_t = 0 \quad P - \text{a.s.} \quad (8)$$

Holding one stock of the asset with price process Z therefore amounts to following a suicide strategy, since one starts with an initial capital of one and ends up with no money at all at time infinity. The mathematical explanation for this phenomenon is that Z is not a martingale on the closed interval $[0, \infty]$, or equivalently, the family $\{Z_t, t \in \mathbb{R}_+\}$ is not uniformly integrable.

What is more, one of the main applications of stochastic exponentials is that they are intricately related to measure changes since they qualify as candidates for density processes (see Girsanov's theorem). Let us fix a filtered probability space $(\Omega, \mathcal{F}_\infty, (\mathcal{F}_t), P)$. In case the stochastic exponential is positive, we may define a new measure Q on $\mathcal{F}_\infty$ via

$$\frac{dQ}{dP} = Z_\infty \quad (9)$$

If Z is a uniformly integrable martingale, then Q is a probability measure since $E[Z_\infty] = Z_0 = 1$. On the other hand, if Z is a strict local martingale, hence a strict supermartingale, then we get $Q(\Omega) = E[Z_\infty] < 1$. It is therefore of paramount interest to have criteria at hand for stochastic exponentials to be true martingales. We first focus on the continuous case.

Theorem 1 (Kazamaki's Criterion). *Let M be a continuous local martingale. Suppose*

$$\sup_T E \left[\exp \left(\frac{1}{2} M_T \right) \right] < \infty \quad (10)$$

where the supremum is taken over all bounded stopping times T . Then $\mathcal{E}(M)$ is a uniformly integrable martingale.

A slightly weaker result, which, however, is often easier to apply, is given by the following criterion.

Theorem 2 (Novikov's Criterion). *Let M be a continuous local martingale. Suppose*

$$E \left[\exp \left(\frac{1}{2} [M]_\infty \right) \right] < \infty \quad (11)$$

Then $\mathcal{E}(M)$ is a uniformly integrable martingale.

Nevertheless, these results are still not applicable in many practically important situations, for example, if one wants to construct martingale measures in stochastic volatility models driven by Brownian motions. In that case, the following result taken from Liptser and Shiryaev [8] often turns out to be useful.

Theorem 3 *Let T be a finite time horizon, ϑ a predictable process with*

$$P \left(\int_0^T \vartheta_s^2 ds < \infty \right) = 1 \quad (12)$$

and B a Brownian motion. Provided that there is $\varepsilon > 0$ such that

$$\sup_{0 \leq t \leq T} E \left[\exp \left(\varepsilon \vartheta_t^2 \right) \right] < \infty \quad P - \text{a.s.} \quad (13)$$

then the stochastic exponential $\mathcal{E}(\int \vartheta dB)$ is a martingale on $[0, T]$.

Let us now turn to the discontinuous case. A generalization of Novikov's criterion has been obtained by Lepingle and Mémín [7] where more results in this direction can be found.

Theorem 4 *Let M be a locally bounded local P -martingale with $\Delta M > -1$. If*

$$E \left[\exp \left(\frac{1}{2} [M^c]_\infty \right) \prod_t (1 + \Delta M_t) \times \exp \left(-\frac{\Delta M_t}{1 + \Delta M_t} \right) \right] < \infty \quad (14)$$

then $\mathcal{E}(M)$ is a uniformly integrable martingale. Here M^c denotes the continuous local martingale part of M .

The situation is particularly transparent for Lévy processes; see Cont and Tankov [1].

Theorem 5 *If M is both a Lévy process and a local martingale, then its stochastic exponential $\mathcal{E}(M)$ (given that it is positive) is already a martingale.*

Alternative conditions for ensuring that stochastic exponentials are martingales in case of Brownian motion driven stochastic volatility models have been

provided in Hobson [4] as well as in Wong and Heyde [9]. Moreover, Kallsen and Shiryaev [6] give results generalizing and complementing the criterions in Lepingle and Mémin [7]. In case of local martingales of stochastic exponential form $\mathcal{E}(X)$, where X denotes one component of a multivariate affine process, Kallsen and Muhle-Garbe [5] give sufficient conditions for M to be a true martingale. Finally, there are important links between stochastic exponentials of BMO -martingales, reverse Hölder inequalities, and weighted norm inequalities (i.e., inequalities generalizing martingale inequalities to certain semimartingales); compare Doléans-Dade and Meyer [2].

References

- [1] Cont, R. & Tankov P. (2003). *Financial Modelling with Jump Processes*, Chapman & Hall/CRC Press, Boca Raton.
- [2] Doléans-Dade, C. & Meyer, P.A. (1979). Inégalités de normes avec poids, *Séminaire de Probabilités de Strasbourg* **13**, 313–331.
- [3] Goll, T. & Kallsen, J. (2000). Optimal portfolio with logarithmic utility, *Stochastic Processes and their Applications* **89**, 91–98.
- [4] Hobson, D. (2004). Stochastic volatility models, correlation and the q -optimal measure, *Mathematical Finance* **14**, 537–556.
- [5] Kallsen, J. & Muhle-Garbe, J. (2007). *Exponentially Affine Martingales and Affine Measure Changes*, preprint, TU München.
- [6] Kallsen, J. & Shiryaev, A.N. (2002). The cumulant process and Esscher's change of measure, *Finance and Stochastics* **6**, 397–428.
- [7] Lepingle, D. & Mémin, J. (1978). Sur l'intégrabilité uniforme des martingales exponentielles, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **42**, 175–203.
- [8] Liptser, R. & Shiryaev, A.N. (1977). *Statistics of Random Processes I*, Springer, Berlin.
- [9] Wong, B. & Heyde, C.C. (2004). On the martingale property of stochastic exponentials, *Journal of Probability and its Applications* **41**, 654–664.

THORSTEN RHEINLÄNDER

Martingales

The word *martingale* originated from Middle French. It means a device for steadying a horse's head or checking its upward movement. In eighteenth-century France, *martingale* also referred to a class of betting strategies in which a player increases the stake usually by doubling each time a bet is lost. The word "martingale", which appeared in the official dictionary of the Academy in 1762 (in the sense of a strategy) means "a strategy that consists in betting all that you have lost". See [7] for more about the origin of martingales. The simplest version of the *martingale betting strategies* was designed to beat a fair game in which the gambler wins his stake if a coin comes up heads and loses it if the coin comes up tails. The strategy had the gambler keep doubling his bet until the first head eventually occurs. At this point, the gambler stops the game and recovers all previous losses, besides winning a profit equal to the original stake. Logically, if a gambler is able to follow this "doubling strategy" (in French, it is still referred to as *la martingale*), he would win sooner or later. But in reality, the exponential growth of the bets would bankrupt the gambler quickly. It is Doob's optional stopping theorem (the cornerstone of martingale theory) that shows the impossibility of successful betting strategies.

In probability theory, a *martingale* is a stochastic process (a collection of random variables) such that the conditional expectation of an observation at some future time t , given all the observations up to some earlier time $s < t$, is equal to the observation at that earlier time s . The name "martingale" was introduced by Jean Ville (1910–1989) as a synonym of "gambling system" in his book on "collectif" in the Borel collection, 1938. However, the concept of martingale was created and investigated as early as in 1934 by Paul Pierre Lévy (1886–1971), and a lot of the original development of the theory was done by Joseph Leo Doob (1910–2004). At present, the martingale theory is one of the central themes of modern probability. It plays a very important role in the study of stochastic processes. In practice, a martingale is a model of a fair game. In financial markets, a fair game means that there is no arbitrage. Mathematical finance builds the bridge that connects no-arbitrage arguments and martingale theory. The fundamental theorem (principle) of asset pricing states, roughly

speaking, that a mathematical model for stochastic asset prices X is free of arbitrage if and only if X is a martingale under an equivalent probability measure. The fair price of a *contingent claim* associated with those assets X is the expectation of its payoff under the martingale equivalent measure (risk neutral measure).

Martingale theory is a vast field of study, and this article only gives an introduction to the theory and describes its use in finance. For a complete description, readers should consult texts such as [4, 13] and [6].

Discrete-time Martingales

A (finite or infinite) sequence of random variables $X = \{X_n | n = 0, 1, 2, \dots\}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called a discrete-time *martingale* (respectively, *submartingale*, *supermartingale*) if for all $n = 0, 1, 2, \dots$, $\mathbb{E}[X_n] < \infty$ and

$$\mathbb{E}[X_{n+1} | X_0, X_1, \dots, X_n] = X_n$$

(respectively $\geq X_n$, $\leq X_n$) (1)

By the tower property of conditional expectations, equation (1) is equivalent to

$$\mathbb{E}[X_n | X_0, X_1, \dots, X_k] = X_k$$

(respectively $\geq X_k$, $\leq X_k$), for any $k \leq n$ (2)

Obviously, X is a submartingale if and only if $-X$ is a supermartingale. Every martingale is also a submartingale and a supermartingale; conversely, any stochastic process that is both a submartingale and a supermartingale is a martingale. The expectation $\mathbb{E}[X_n]$ of a martingale X at time n , is a constant for all n . This is one of the reasons that in a fair game, the asset of a player is supposed to be a martingale. For a supermartingale X , $\mathbb{E}[X_n]$ is a nonincreasing function of n , whereas for a submartingale X , $\mathbb{E}[X_n]$ is a nondecreasing function of n . Here is a mnemonic for remembering which is which: "*Life is a supermartingale; as time advances, expectation decreases.*" The conditional expectation of X_n in equation (2) should be evaluated on the basis

of *all* information available up to time k , which can be summarized by a σ -algebra $\mathcal{F}_k$,

$$\mathcal{F}_k = \{\text{all events occurring at times } i = 0, 1, 2, \dots, k\} \quad (3)$$

A sequence of increasing σ -algebras $\{\mathcal{F}_n | n = 0, 1, 2, \dots\}$, that is, $\mathcal{F}_k \subseteq \mathcal{F}_n \subseteq \mathcal{F}$ for $k \leq n$, is called a *filtration*, denoted by $\mathbb{F}$. When $\mathcal{F}_n$ is the smallest σ -algebra containing all the information of X up to time n , $\mathcal{F}_n$ is called the σ -algebra generated by $X_0, X_1, \dots, X_n$, denoted by $\sigma\{X_0, X_1, \dots, X_n\}$, and $\mathbb{F}$ is called the *natural filtration* of X . For another sequence of random variables $\{Y_k | k = 0, 1, \dots\}$, let $\mathcal{F}_k = \sigma\{Y_0, Y_1, \dots, Y_k\}$, then $\mathbb{E}[X_n | Y_0, Y_1, \dots, Y_k] = \mathbb{E}[X_n | \mathcal{F}_k]$.

A sequence of random variables $X = \{X_n | n = 0, 1, 2, \dots\}$ on the filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ is said to be *adapted* if X_n is $\mathcal{F}_n$ -measurable for each n , which means that given $\mathcal{F}_n$, there is no randomness in X_n . An adapted X is called a discrete-time *martingale* (respectively *submartingale*, *supermartingale*) with respect to the filtration $\mathbb{F}$, if for each n , $\mathbb{E}[|X_n|] < \infty$, and

$$\mathbb{E}[X_n | \mathcal{F}_k] = X_k \quad (\text{respectively } \geq X_k, \leq X_k), \quad \text{for any } k \leq n \quad (4)$$

Example 1 (Closed Martingales). Let Z be a random variable with $\mathbb{E}[Z] < \infty$, then for any filtration $\mathbb{F} = (\mathcal{F}_n)$, $X_n = \mathbb{E}[Z | \mathcal{F}_n]$ is a martingale (also called a *martingale closed by Z*). Conversely, for any martingale X on a finite probability space, there exists a random variable Z such that $X_n = \mathbb{E}[Z | \mathcal{F}_n]$.

Example 2 (Partial Sums of i.i.d. Random Variables). Let $Z_1, Z_2, \dots$ be a sequence of independent, identically distributed (i.i.d.) random variables such that $\mathbb{E}[Z_n] = \mu$, and $\mathbb{E}[Z_n^2] = \sigma^2 < \infty$, and that the moment generating function $\phi(\theta) = \mathbb{E}[e^{\theta Z_1}]$ exists for some $\theta > 0$. Let S_n be the partial sum, $S_n = Z_1 + \dots + Z_n$, also called a *random walk*. Let $\mathcal{F}_n = \sigma\{Z_1, \dots, Z_n\}$. Then

$$S_n - n\mu, \quad (S_n - n\mu)^2 - n\sigma^2, \quad \frac{\theta^{S_n}}{[\phi(\theta)]^n} \quad (5)$$

are all martingales. If $\mathbb{P}(Z_k = +1) = p$, $\mathbb{P}(Z_k = -1) = q = 1 - p$, then S_n is called a *simple random*

walk and $(q/p)^{S_n}$ is a martingale since $\phi(p/q) = 1$; in particular, when $p = q = 1/2$, S_n is called a *simple symmetric random walk*. If Z_k has the Bernoulli distribution, $\mathbb{P}(Z_k = +1) = p$, $\mathbb{P}(Z_k = 0) = q = 1 - p$, then S_n has the binomial distribution (n, p) , and $(q/p)^{2S_n - n}$ is a martingale since $\phi([q/p]^2) = q/p$.

Example 3 (Polya's Urn). An urn initially contains r red and b blue marbles. One is chosen randomly. Then it is put back together with another one of the same color. Let X_n be the number of red marbles in the urn after n iterations of this procedure, and let $Y_n = X_n/(n + r + b)$. Then the sequence Y_n is a martingale.

Example 4 (A Convex Function of Martingales). By Jensen's inequality, a convex function of a martingale is a submartingale. Similarly, a convex and nondecreasing function of a submartingale is also submartingale. Examples of convex functions are $\max(x - k, 0)$ for constant k , $|x|^p$ for $p \geq 1$ and $e^{\theta x}$ for constant θ .

Example 5 (Martingale Transforms). Let X be a martingale with respect to the filtration $\mathbb{F}$ and H be a *predictable process* with respect to $\mathbb{F}$, that is, H_n is $\mathcal{F}_{n-1}$ -measurable for $n \geq 1$, where $\mathcal{F}_0 = \{\emptyset, \Omega\}$. A martingale transform of X by H is defined by

$$(H \cdot X)_n = H_0 X_0 + \sum_{i=1}^n H_i (X_i - X_{i-1}) \quad (6)$$

where the expression $H \cdot X$ is the discrete analog of the *stochastic integral* $\int H dX$. If $\mathbb{E}|(H \cdot X)_n| < \infty$ for $n \geq 1$, then $(H \cdot X)_n$ is a martingale with respect to $\mathbb{F}$. The interpretation is that in a fair game X , if we choose our bet at each stage on the basis of the prior history, that is, the bet H_n for the n th gamble only depends on $\{X_0, X_1, \dots, X_{n-1}\}$, then the game will continue to be fair. If X_n is the asset price at time n and H_n is the number of shares of the assets held by the investor during the time period from time n until time $n + 1$, more precisely, for the time interval $[n, n + 1)$, then $(H \cdot X)_n$ is the total gain (or loss) up to time n (the value of the portfolio at time n with the trading strategy H).

A random variable T taking values in $\{0, 1, 2, \dots, \infty\}$ is a *stopping time* T with respect to a filtration $\mathbb{F} = \{\mathcal{F}_n | n = 0, 1, 2, \dots\}$, if for each n , the

event $\{T = n\}$ is $\mathcal{F}_n$ -measurable, or equivalently, the event $\{T \leq n\}$ is $\mathcal{F}_n$ -measurable. If S and T are stopping times, then $S + T$, $S \vee T = \max(S, T)$, and $S \wedge T = \min(S, T)$ are all stopping times. Particularly, $T \wedge n$ is a bounded stopping time for any fixed time n . $X_n^T =: X_{T \wedge n}$ is said to be the *process X stopped at T* , since on the event $\{\omega | T(\omega) = k\}$, $X_n^T = X_k$ for $n = k, k + 1, \dots$.

Doob's Optional Stopping Theorem

Let X be a martingale and T be a bounded stopping time with respect to the same filtration $\mathbb{F}$, then $\mathbb{E}[X_T] = \mathbb{E}[X_0]$. Conversely, for an adapted process X , if $\mathbb{E}[|X_T|] < \infty$ and $\mathbb{E}[X_T] = \mathbb{E}[X_0]$ hold for all bounded stopping time T , then X is a martingale. This theorem says roughly that stopping a martingale at a stopping time T does not alter its expectation, provided that the decision when to stop is based only on information available up to time T . The theorem also shows that a martingale stopped at a stopping time is still a martingale, and there is no way to be sure to win in a fair game if the stopping time is bounded.

Continuous-time Martingales

A continuous-time stochastic process X on filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ is a collection of random variables $X = \{X_t : 0 \leq t \leq \infty\}$, where X_t is a random variable observed at time t , and the filtration $\mathbb{F} = \{\mathcal{F}_t : 0 \leq t \leq \infty\}$, which is a family of increasing σ -algebras, $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$ for $s \leq t$. A process X is said to be *adapted* if X_t is $\mathcal{F}_t$ measurable for each t . A random variable T taking values in $[0, \infty]$ is called a *stopping time*, if the event $\{T \leq t\}$ is $\mathcal{F}_t$ measurable for each t . The *stopping time σ -algebra $\mathcal{F}_T$* is defined to be $\mathcal{F}_T = \{A \in \mathcal{F} | A \cap \{T \leq t\} \in \mathcal{F}_t, \text{ all } t \geq 0\}$, which represents the information up to the stopping time T .

A real-valued, adapted process X is called a continuous-time *martingale* (respectively *supermartingale*, *submartingale*) with respect to the filtration $\mathbb{F}$ if

$$1. \quad \mathbb{E}[|X_t|] < \infty, \quad \text{for } t > 0 \quad (7)$$

$$2. \quad \mathbb{E}[X_t | \mathcal{F}_s] = X_s \text{ (respectively } \leq X_s, \geq X_s), \\ \text{a.s. for any } 0 \leq s \leq t \quad (8)$$

Continuous-time martingales have the same properties as discrete-time martingales. For example, *Doob's optional stopping theorem* says that for a martingale X_t with right continuous paths, which is closed in $\mathbf{L}^1$ by a random variable X_∞ , we have

$$\mathbb{E}[X_T | \mathcal{F}_S] = X_S \quad \text{a.s. for any two stopping times} \\ 0 \leq S \leq T \quad (9)$$

The most important continuous-time martingale is Brownian motion, which was named for the Scottish botanist Robert Brown, who, in 1827, observed ceaseless and irregular movement of pollen grains suspended in water. It was studied by Albert Einstein in 1905 at the level of modern physics. Its mathematical model was first rigorously constructed in 1923 by Norbert Wiener. Brownian motion is also called a *Wiener process*. The Wiener process gave rise to the study of continuous-time martingales, and has been an example that helps mathematicians to understand stochastic calculus and diffusion processes.

It was Louis Bachelier (1870–1946), now recognized as the founder of mathematical finance (see [9]), who first, in 1900, used Brownian motion B to model short-term stock prices S_t at a time t in financial markets, that is, $S_t = S_0 + \sigma B_t$, where $\sigma > 0$ is a constant. Now we can see that if Brownian motion B is defined on $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, then the price process S is a martingale under the probability measure $\mathbb{P}$.

In 1965, the American economist Paul Samuelson rediscovered Bachelier's ideas and proposed the *geometric Brownian motion* $S_0 \exp\{(\mu - (\sigma^2/2))t + \sigma B_t\}$ as a model for long-term stock prices S_t . That is, S_t follows the stochastic differential equation (SDE): $dS_t = \mu S_t dt + \sigma S_t dB_t$. From this simple structure, we get the famous Black–Scholes option price formulas for European calls and puts. This SDE is now called the Black–Scholes equation (model). Contrary to Bachelier's setting, the price process S is not a martingale under $\mathbb{P}$. However, by Girsanov's theorem, there is a unique probability measure $\mathbb{Q}$, which is equivalent to $\mathbb{P}$, such that the discounted stock price $e^{-rt} S_t$ is a martingale under $\mathbb{Q}$ for $0 \leq t \leq T$, where r is the *riskless rate of interest*, and $T > 0$ is a fixed constant.

The reality is not as simple as the above linear SDE. A simple generalization is $dS_t = \mu(t, S_t) dt + \sigma(t, S_t) dB_t$. If one believes that risky asset prices

have jumps, an appropriate model might be

$$dS_t = \mu(t, S_t) dt + \sigma(t, S_t) dB_t + J(t, S_t) dN_t \quad (10)$$

where N is a Poisson process with intensity λ , $J(t, S_t)$ refers to the jump size, and N indicates when the jumps occur. Since N is a counting (pure jump) process with independent and stationary increments, both $N_t - \lambda t$ and $(N_t - \lambda t)^2 - \lambda t$ are martingales. For a more general model, we could replace N by a Lévy process that includes the Brownian motion and Poisson process as special cases.

Under these general mathematical models, it becomes hard to turn the fundamental principle of asset pricing into a precise mathematical theorem: the absence of arbitrage possibilities for a stochastic process S , a *semimartingale* defined on $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, is equivalent to the existence of an equivalent measure $\mathbb{Q}$, under which S is a *local martingale*, sometimes, a *sigma martingale*. See [2] or [3].

Local Martingales and Finite Variation Processes

There are two types of processes with only jump discontinuities. A process is said to be *càdlàg* if it almost surely (a.s.) has sample paths that are right continuous, with left limits. A process is said to be *càglàd* if it almost surely has sample paths that are left continuous, with right limits. The words *càdlàg* and *càglàd* are acronyms from the French for *continu à droite, limites à gauche*, and *continu à gauche, limites à droite*, respectively. Let

$$\begin{aligned} \mathbb{D} &= \text{the space of adapted processes} \\ &\quad \text{with càdlàg paths} \\ \mathbb{L} &= \text{the space of adapted processes} \\ &\quad \text{with càglàd paths} \end{aligned} \quad (11)$$

An adapted, càdlàg process A is called a *finite variation* (FV) process if $\sup \sum_{i=1}^N |A_{t_i} - A_{t_{i-1}}|$ is bounded almost surely for each constant $t > 0$, where the supremum is taken over the set of all partitions $0 = t_0 \leq t_1 \leq \dots \leq t_N = t$. An FV process is a difference of two increasing processes. Although the Brownian motion B has continuous paths, it has

paths of *infinite variation* on $[0, t]$, which prevents us from defining the stochastic integral $\int H dB$ as a *Riemann–Stieltjes integral*, path by path.

An adapted, càdlàg process M is called a *local martingale* with respect to a filtration $\mathbb{F}$ if there exists a sequence of increasing stopping time T_n with $\lim_{n \rightarrow \infty} T_n = \infty$ almost surely, such that for each n , $M_{t \wedge T_n}$ is a martingale. A similar concept is that a function is *locally bounded*: for example, $1/t$ is not bounded over $(0, 1]$, but it is bounded on the interval $[1/n, 1]$ for any integer n . A process moving very rapidly though with continuous paths, or jumping unboundedly and frequently, might not be a martingale. However, we could modify it to be a martingale by stopping it properly, that is, it is a martingale up to a stopping time, but may not be a martingale for all time.

The class of local martingales includes martingales as special cases. For example, if for every $t > 0$, $\mathbb{E}\{\sup_{s \leq t} |M_s|\} < \infty$, then M is a martingale; if for all $t > 0$, $\mathbb{E}\{[M, M]_t\} < \infty$, then M is a martingale, and $\mathbb{E}\{M_t^2\} = \mathbb{E}\{[M, M]_t\}$. Conversely, if M is a martingale with $\mathbb{E}\{M_t^2\} < \infty$ for all $t > 0$, then $\mathbb{E}\{[M, M]_t\} < \infty$ for all $t > 0$. For the definition of quadratic variation $[M, M]_t$, see equation (14) in the next section.

Not all local martingales are martingales. Here is a typical *example* of a local martingale, but not a martingale. Lots of continuous-time martingales, supermartingales, and submartingales can be constructed from Brownian motion, since it has independent and stationary increments and it can be approximated by a random walk. For example, let B be a standard Brownian motion in $\mathbb{R}^3$ with $B_0 = x \neq 0$. Let $u(y) = \|y\|^{-1}$, be a *superharmonic* function on $\mathbb{R}^3$. $M_t = u(B_t)$ is a positive supermartingale. Since $\lim_{t \rightarrow \infty} \sqrt{t} \mathbb{E}\{M_t\} = \sqrt{\pi}$ and $\mathbb{E}\{M_0\} = u(x)$, M does not have constant expectations and it cannot be a martingale. M is known as the *inverse Bessel Process*. For each n , we define a stopping time $T_n = \inf\{t > 0 : \|B_t\| \leq 1/n\}$. Since the function u is *harmonic* outside of the ball of radius $1/n$ centered at the origin, the process $\{M_{t \wedge T_n} : t \geq 0\}$ is a martingale for each n . Therefore, M is a local martingale.

Semimartingales and Stochastic Integrals

Today stocks and bonds are traded globally almost 24 hours a day, and online trading happens every second.

When trading takes place almost continuously, it is simpler to use a continuous-time stochastic processes to model the price X . The value of the portfolio at time t with the continuous-time trading strategy H becomes the limit of sums as shown in the martingale transform $(H \cdot X)_n$ in equation (6), that is, the stochastic integral $\int_0^t H_s dX_s$. Stochastic calculus is more complicated than regular calculus because X can have paths of infinite variation, especially when X has unbounded jumps, for example, when X is Brownian motion, a continuous-time martingale, or a local martingale. For stochastic integration theory, see **Stochastic Integrals** or consult [8, 11] and [12], and other texts.

Let $0 = T_1 \leq \dots \leq T_{n+1} < \infty$ be a sequence of stopping times and $H_i \in \mathcal{F}_{T_i}$ with $|H_i| < \infty$. A process H with a representation

$$H_t = H_0 \mathbf{1}_{\{0\}}(t) + \sum_{i=1}^n H_i \mathbf{1}_{(T_i, T_{i+1}]}(t) \quad (12)$$

is called a *simple predictable process*. A collection of simple predictable processes is denoted by $\mathbf{S}$. For a process $X \in \mathbb{D}$ and $H \in \mathbf{S}$ having the representation (12), we define a linear mapping as the martingale transforms in equation (6) in the discrete-time case

$$(H \cdot X)_t = H_0 X_0 + \sum_{i=1}^n H_i (X_{t \wedge T_{i+1}} - X_{t \wedge T_i}) \quad (13)$$

If for any $H \in \mathbf{S}$ and each $t \geq 0$, the sequence of random variables $(H^n \cdot X)_t$ converges to $(H \cdot X)_t$ in probability, whenever $H^n \in \mathbf{S}$ converges to H uniformly, then X is called a *semimartingale*. For example, an FV process, a local martingale with continuous paths, and a Lévy process are all semimartingales.

Since the space $\mathbf{S}$ is dense in $\mathbb{L}$, for any $H \in \mathbb{L}$, there exists $H_n \in \mathbf{S}$ such that H_n converges to H . For a semimartingale X and a process $H \in \mathbb{L}$, the *stochastic integral* $\int H dX$, also denoted by $(H \cdot X)$, is defined by $\lim_{n \rightarrow \infty} (H^n \cdot X)$. For any $H \in \mathbb{L}$, $H \cdot X$ is a semimartingale, it is an FV process if X is, and it is a local martingale if X is. But $H \cdot X$ may not be a martingale even if X is. $H \cdot X$ is a martingale if X is a local martingale and $\mathbb{E}\{\int_0^t H_s^2 d[X, X]_s\} < \infty$ for each $t > 0$.

For a semimartingale X , its *quadratic variation* $[X, X]$ is defined by

$$[X, X]_t = X_t^2 - 2 \int_0^t X_{s-} dX_s \quad (14)$$

where X_{s-} denotes the left limit at s . Let $[X, X]^c$ denote the path-by-path continuous part of $[X, X]$, and $\Delta X_s = X_s - X_{s-}$ be the jump of X at s , then $[X, X]_t = [X, X]_t^c + \sum_{0 \leq s \leq t} (\Delta X_s)^2$. For an FV process X , $[X, X]_t = \sum_{0 \leq s \leq t} (\Delta X_s)^2$. In particular, if X is an FV process with continuous paths, then $[X, X]_t = X_0^2$ for all $t \geq 0$. For a continuous local martingale X , then $X^2 - [X, X]_t$ is a continuous local martingale. Moreover, if $[X, X]_t = X_0^2$ for all t , then $X_t = X_0$ for all t ; in other words, if an FV process is also a continuous local martingale, then it is a constant process.

Lévy's Characterization of Brownian Motion

A process X is a standard Brownian motion if and only if it is a continuous local martingale with $[X, X]_t = t$.

The theory of stochastic integration for integrands in $\mathbb{L}$ is sufficient to establish Itô's formula, the Girsanov–Meyer theorem, and to study SDEs. For example, the *stochastic exponential* of a semimartingale X with $X_0 = 0$, written $\mathcal{E}(X)$, is the unique semimartingale Z that is a solution of the linear SDE: $Z_t = 1 + \int_0^t Z_{s-} dX_s$. When X is a continuous local martingale, so is $\mathcal{E}(X)_t = \exp\{X_t - \frac{1}{2}[X, X]_t\}$. Furthermore, if *Kazamaki's Criterion* $\sup_T \mathbb{E}\{\exp(\frac{1}{2}X_T)\} < \infty$ holds, where the supremum is taken over all bounded stopping times, or if *Novikov's Criterion* $\mathbb{E}\{\exp(\frac{1}{2}[X, X]_\infty)\} < \infty$ holds (stronger but easier to check in practice), then $\mathcal{E}(X)$ is a martingale. See [10] for more on these conditions. When X is Brownian motion, $\mathcal{E}(X) = \exp\{X_t - \frac{1}{2}t\}$ is referred to as *geometric Brownian motion*.

The space of integrands $\mathbb{L}$ is not general enough to have *local times* and martingale representation theory, which is essential for hedging in finance. On the basis of the *Bichteler–Dellacherie theorem*, X is a semimartingale if and only if $X = M + A$, where M is a local martingale and A is an FV process, we can extend the stochastic integration from $\mathbb{L}$ to the space $\mathcal{P}$ of predictable processes, which are measurable with respect to $\sigma\{H : H \in \mathbb{L}\}$. For a semimartingale

X , if a predictable H is X integrable, that is, we can define the stochastic integral $H \cdot X$, then we write $H \in L(X)$ (see chapter 4 of [8]). If $H \in \mathcal{P}$ is locally bounded then $H \in L(X)$ and $H \cdot X$ is a local martingale if X is. However, if $H \in \mathcal{P}$ is not locally bounded or $H \notin \mathbb{L}$, then $H \cdot X$ may not be a local martingale even if X is an $\mathbf{L}^2$ martingale. For such an example due to M. Émery, see pp 152 of [5] or pp 176 of [8]. If X is a local martingale and $H \in L(X)$, then $H \cdot X$ is a sigma martingale.

Sigma Martingales

The concept of a sigma martingale was introduced by Chou [1] and further analyzed by Émery [5]. It has seen a revival in popularity owing to Delbaen and Schachermayer [2]; see [8] for a more detailed treatment. Sigma martingales relate to martingales analogously as sigma-finite measures relate to finite measures. A sigma martingale, which may not be a local martingale, has the essential features of a martingale.

A semimartingale X is called a *sigma martingale* if there exists a martingale M and a nonnegative $H \in \mathcal{P}$ such that $X = H \cdot M$, or, equivalently, if there exists a nonnegative $H \in \mathcal{P}$ such that $H \cdot X$ is a martingale.

A local martingale is a sigma martingale, but a sigma martingale with large jumps might fail to be a local martingale. If X is a sigma martingale and if either $\sup_{s \leq t} |X_s|$ or $\sup_{s \leq t} |\Delta X_s|$ is *locally integrable* (for example, X has continuous paths or bounded jumps), then X is a local martingale. If X is a sigma martingale and $H \in L(X)$, then $H \cdot X$ is always a sigma martingale.

The concept of a sigma martingale is new in the context of mathematical finance. It was introduced to deal with possibly unbounded jumps of the asset price process X . When we consider the process X with jumps, it is often convenient to assume the jumps to be unbounded, for example, the Lévy processes and the family of ARCH, GARCH processes. If the conditional distribution of jumps is Gaussian, then the process is not locally bounded. In that case, the concept of a sigma martingale is unavoidable. On the other hand, if we are only interested in how to price and hedge some contingent claims, not the underlying assets X , then it might not be necessary to require the asset price X to be a (local) martingale

and it suffices to require $H \cdot X$ to be a martingale for some H , that is, X is a sigma martingale. Moreover, nonnegative sigma martingales are local martingales, so in particular for stock prices, we do need to consider sigma martingales.

Finally, we cite two *fundamental theorems of asset pricing* from chapters 8 and 14 of [3] to see why we need sigma martingales in mathematical finance.

Theorem 1 *Let the discounted price process S be a locally bounded semimartingale defined on $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$. Then there exists a probability measure $\mathbb{Q}$ (equivalent to $\mathbb{P}$) under which S is a local martingale, if and only if S satisfies the condition of no free lunch with vanishing risk (NFLVR).*

Here the concept of NFLVR is a mild strengthening of the concept of no arbitrage, which is introduced by Delbaen and Schachermayer in [2].

Theorem 2 *If we assume that S is a nonlocally bounded semimartingale, then we have a general theorem by replacing the term “local martingale” by the term “sigma martingale” in Theorem 1 above. However if $S \geq 0$, then “local martingale” suffices, because sigma martingales bounded below are a priori local martingales.*

Conclusion

A local martingale is a martingale up to a sequence of stopping times that goes to ∞ , while a sigma martingale is a countable sum (a mixture) of martingales.

References

- [1] Chou, C.S. (1977). Caractérisation d’une classe de semimartingales, *Séminaire de Probabilités XIII*, LNM, Vol. 721, Springer, pp. 250–252.
- [2] Delbaen, F. & Schachermayer, W. (1998). *The Fundamental Theorem of Asset Pricing for Unbounded Stochastic Processes*, *Mathematische Annalen*, Vol. 312, Springer, pp. 215–250.
- [3] Delbaen, F. & Schachermayer, W. (2006). *The Mathematics of Arbitrage*, *Springer Finance Series*, Springer-Verlag, New York.
- [4] Dellacherie, C. & Meyer, P.A. (1982). *Probabilities and Potential*, Vol. 29 of *North-Holland Mathematics Studies*, North-Holland, Amsterdam.
- [5] Émery, M. (1980). Compensation de processus à variation finie non localement intégrables., *Séminaire de Probabilités XIV*, LNM, Vol. 784, Springer, pp. 152–160.

-
- [6] Ethier, S. & Kurtz, T.G. (1986). *Markov Processes: Characterization and Convergence*, Wiley, New York.
- [7] Mansuy, R. (2005). Histoire de martingales, *Mathématiques et Sciences Humaines/Mathematical Social Sciences* **169**(1), 105–113.
- [8] Protter, P. (2003). *Stochastic Integration and Differential Equations, Applications of Mathematics*, 2nd Edition, Springer, Vol. 21.
- [9] Protter, P. (2007). *Louis Bachelier's Theory of Speculation: The Origins of Modern Finance*, M. Davis & A. Etheridge, eds, a book review in the *Bulletin of the American Mathematical Society*, Vol. 45, No. 4, pp. 657–660.
- [10] Protter, P. & Shimbo, K. (2006). *No Arbitrage and General Semimartingales*. To appear in the Festschrift.
- [11] Revuz, D. & Yor, M. (1991). *Continuous Martingales and Brownian motion, Grundlehren der Mathematischen Wissenschaften*, 3rd Edition, Springer, Vol. 293.
- [12] Rogers, L.C.G. & Williams, D. (2000). *Diffusions, Markov Processes and Martingales*, Vols 1 and 2, Cambridge University Press.
- [13] Williams, D. (1991). *Probability with Martingales*, Cambridge University Press.

Related Articles

Equivalent Martingale Measures; Fundamental Theorem of Asset Pricing; Markov Processes; Martingale Representation Theorem.

LIQING YAN

Itô's Formula

For a function depending on space and time parameters, rules of differentiation are well known. For a function depending on space and time parameters and also on a randomness parameter, Itô's formulas provide rules of differentiation. These rules of differentiation are based on the complementary notion of stochastic integration (see **Stochastic Integrals**). More precisely, given a probability space $(\Omega, \mathbb{P}, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0})$, Itô's formulas deal with $(F(X_t); t \geq 0)$, where F is a deterministic function defined on $\mathbb{R}$ and $(X_t)_{t \geq 0}$ is a random process such that integration of locally bounded predictable processes is possible with respect to $(X_t)_{t \geq 0}$ and satisfies a property equivalent to the Lebesgue dominated convergence theorem. This means that $(X_t)_{t \geq 0}$ is a semimartingale and therefore has a finite quadratic variation process $([X]_t, t \geq 0)$ (see **Stochastic Integrals**) defined as

$$[X]_t = \lim_{n \rightarrow \infty} \sum \left(X_{s_{i+1}^n} - X_{s_i^n} \right)^2 \text{ in probability,} \\ \text{uniformly on time intervals} \quad (1)$$

where $(s_i^n)_{1 \leq i \leq n}$ is a subdivision of $[0, t]$ whose mesh converges to 0 as n tends to ∞ .

We will see that Itô's formulas also provide information on the stochastic structure of the process $(F(X_t), t \geq 0)$. We first introduce the formula established by Itô in 1951. Consider a process $(X_t)_{t \geq 0}$ of the form

$$X_t = \int_0^t H_s dB_s + \int_0^t G_s ds \quad (2)$$

where $(B_s)_{s \geq 0}$ is a real-valued Brownian motion, and $(H_s)_{s \geq 0}$ and $(G_s)_{s \geq 0}$ are locally bounded predictable processes. Then for every C^2 -function F from $\mathbb{R}$ to $\mathbb{R}$, we have

$$F(X_t) = F(X_0) + \int_0^t F'(X_s) H_s dB_s \\ + \int_0^t F'(X_s) G_s ds + \frac{1}{2} \int_0^t H_s^2 F''(X_s) ds \quad (3)$$

The process defined in formula (2) is an example of continuous semimartingale. Here is the classical Itô formula for a general semimartingale $(X_s)_{s \geq 0}$ (e.g., [7, 9]) and F in $\mathcal{C}^2$

$$F(X_t) = F(X_0) + \int_0^t F'(X_{s-}) dX_s \\ + \frac{1}{2} \int_0^t F''(X_s) d[X]_s^c \\ + \sum_{0 \leq s \leq t} \left\{ F(X_s) - F(X_{s-}) - F'(X_{s-}) \Delta X_s \right\} \quad (4)$$

where $[X]^c$ is the continuous part of $[X]$. For continuous semimartingales, formula (4) becomes

$$F(X_t) = F(X_0) + \int_0^t F'(X_s) dX_s \\ + \frac{1}{2} \int_0^t F''(X_s) d[X]_s \quad (5)$$

In the special case when $(X_t)_{t \geq 0}$ is a real Brownian motion, then $[X]_t = t$.

The multidimensional version of formula (4) gives the expansion of $F(X_t^{(1)}, X_t^{(2)}, \dots, X_t^{(d)})$ for F a real-valued function of $\mathcal{C}^2(\mathbb{R}^d)$ and d semimartingales $X^{(1)}, X^{(2)}, \dots, X^{(d)}$. We set $X = (X^{(1)}, X^{(2)}, \dots, X^{(d)})$:

$$F(X_t) = F(X_0) + \sum_{i=1}^d \int_0^t \frac{\partial F}{\partial x_i}(X_{s-}) dX_s^{(i)} \\ + \frac{1}{2} \sum_{1 \leq i, j \leq d} \int_0^t \frac{\partial^2 F}{\partial x_i \partial x_j}(X_{s-}) d[X^{(i)}, X^{(j)}]_s^c \\ + \sum_{0 \leq s \leq t} \left\{ F(X_s) - F(X_{s-}) \right. \\ \left. - \sum_{i=1}^d \frac{\partial F}{\partial x_i}(X_{s-}) \Delta X_s^{(i)} \right\} \quad (6)$$

Note the Itô formula corresponding to the case of the couple of semimartingales $(X_t, t)_{t \geq 0}$ with X

continuous and F in $\mathcal{C}^2(\mathbb{R}^2)$

$$\begin{aligned} F(X_t, t) = & F(X_0, 0) + \int_0^t \frac{\partial F}{\partial x}(X_s, s) dX_s \\ & + \int_0^t \frac{\partial F}{\partial t}(X_s, s) ds \\ & + \frac{1}{2} \int_0^t \frac{\partial^2 F}{\partial x^2}(X_s, s) d[X]_s \end{aligned} \quad (7)$$

Each of the above Itô formulas gives a decomposition of the process $(F(X_t), t \geq 0)$ that can be reduced to the sum of a local martingale and an adapted bounded variation process. This shows that $F(X)$ is a semimartingale. In practical situations, the considered function F might not be a $\mathcal{C}^2$ -function and the process $F(X)$ might not be a semimartingale. Hence, many authors have written extensions of the above formulas enlightening this $\mathcal{C}^2$ -condition. Some of them use the notion of local times (see **Local Times**) whose definition can actually be set by the following first extension of the Itô formula.

For F real-valued convex function and X semimartingale, $F(X)$ is a semimartingale too and

$$F(X_t) = F(X_0) + \int_0^t F'(X_{s-}) dX_s + A_t \quad (8)$$

where F' is the left derivative of F and $(A_t, t \geq 0)$ is an adapted, right continuous increasing process such that $\Delta A_s = F(X_s) - F(X_{s-}) - F'(X_{s-})\Delta X_s$.

Choosing $F(x) = |x - a|$, one obtains the existence of an increasing process $(L_t^a, t \geq 0)$ such that

$$\begin{aligned} |X_t - a| = & |X_0 - a| + \int_0^t \operatorname{sgn}(X_{s-} - a) dX_s + L_t^a \\ & + \sum_{0 < s \leq t} \{|X_s - a| - |X_{s-} - a| \\ & - \operatorname{sgn}(X_{s-} - a)\Delta X_s\} \end{aligned} \quad (9)$$

The process L^a is called the *local time process* of X at a (see **Local Times** for alternative definition and basic properties). Note that L^a is continuous in t .

Coming back to formula (8), denote by μ the second derivative of F in the generalized function sense; then the Meyer–Itô formula goes further by giving the expression of the bounded variation

process A :

$$\begin{aligned} F(X_t) = & F(X_0) + \int_0^t F'(X_{s-}) dX_s \\ & + \sum_{0 < s \leq t} \{F(X_s) - F(X_{s-}) - F'(X_{s-})\Delta X_s\} \\ & + \frac{1}{2} \int_{\mathbb{R}} L_t^x \mu(dx) \end{aligned} \quad (10)$$

The Meyer–Itô formula is also obviously available for functions F , which are difference of two convex functions.

For the semimartingales X , such that for every $t > 0$: $\sum_{0 < s \leq t} |\Delta X_s| < \infty$ a.s., Bouleau and Yor extended the Meyer–Itô formula to functions F , admitting a Radon–Nikodym derivative with respect to the Lebesgue measure. Indeed, the Bouleau–Yor formula [2] states in that case

$$\begin{aligned} F(X_t) = & F(X_0) + \int_0^t F'(X_{s-}) dX_s \\ & + \sum_{0 < s \leq t} \{F(X_s) - F(X_{s-}) - F'(X_{s-})\Delta X_s\} \\ & - \frac{1}{2} \int_{\mathbb{R}} F'(x) d_x L_t^x \end{aligned} \quad (11)$$

Note that the Bouleau–Yor formula requires the construction of a stochastic integration of deterministic functions with respect to the process $(L_t^x, x \in \mathbb{R})$, although this last process might not be a semimartingale. Besides, this formula shows that the process $(F(X_t), t \geq 0)$ might not be a semimartingale but a Dirichlet process (i.e., the sum of a local martingale and a 0-quadratic variation process).

In the special case of a real Brownian motion $(B_t, t \geq 0)$, Föllmer, Protter, and Shirayev formula offers an extension of the Bouleau–Yor formula to space–time functions G defined on $\mathbb{R} \times \mathbb{R}_+$ admitting a Radon–Nikodym derivative with respect to the space parameter $\partial G / \partial x$ with some continuity properties (see [6], for the detailed assumptions)

$$\begin{aligned} G(B_t, t) = & G(B_0, 0) + \int_0^t G(B_s, ds) \\ & + \int_0^t \frac{\partial G}{\partial x}(B_s, s) dB_s + \frac{1}{2} \left[\frac{\partial G}{\partial x}(B, \cdot, \cdot), B \right]_t \end{aligned} \quad (12)$$

with $\int_0^t G(B_s, ds) = \lim_{n \rightarrow \infty} \sum_{i=1}^n (G(B_{s_{i+1}^n}, s_{i+1}^n) - G(B_{s_i^n}, s_i^n))$ in probability, where $(s_i^n)_{1 \leq i \leq n}$ is a subdivision of $[0, t]$ whose mesh converges to 0 as n tends to ∞ (Reference 5 contains a similar result and Reference 1 extends it to nondegenerate diffusions).

Another way to extend the Bouleau–Yor formula, in the case of a real Brownian motion, consists in the construction of the stochastic integration of locally bounded deterministic space–time functions $f(x, t)$ with respect to the local time process $(L_t^x, x \in \mathbb{R}, t \geq 0)$ of B . That way one obtains, for the functions G admitting locally bounded first-order derivatives, Eisenbaum's formula [3]:

$$\begin{aligned} G(B_t, t) &= G(B_0, 0) + \int_0^t \frac{\partial G}{\partial x}(B_s, s) dB_s \\ &\quad + \int_0^t \frac{\partial G}{\partial t}(B_s, s) ds \\ &\quad - \frac{1}{2} \int_0^t \int_{\mathbb{R}} \frac{\partial G}{\partial x}(x, s) dL_s^x \end{aligned} \quad (13)$$

The comparison of formula (13) with formulas (12) and (7) provides some rules of integration with respect to the local time process of B such as

- for f continuous function on $\mathbb{R} \times \mathbb{R}_+$

$$\int_0^t \int_{\mathbb{R}} f(x, s) dL_s^x = -[f(B, \cdot), B]_t \quad (14)$$

- for f locally bounded function on $\mathbb{R} \times \mathbb{R}_+$ admitting a locally bounded Radon–Nikodym derivative $\partial f / \partial x$

$$\int_0^t \int_{\mathbb{R}} f(x, s) dL_s^x = - \int_0^t \frac{\partial f}{\partial x}(X_s, s) ds \quad (15)$$

See [2] for an extension of formula (13) to Lévy processes.

We now mention the special case of a space–time function $G(x, s)$ defined as follows:

$$G(x, s) = G_1(x, s)1_{\{x > b(s)\}} + G_2(x, s)1_{\{x \leq b(s)\}} \quad (16)$$

where $(b(s), s \geq 0)$ is a continuous curve and G_1 and G_2 are $\mathcal{C}^2$ -functions that coincide on $x = b(s)$

(but not their derivatives). This case is treated in [8] for X continuous semimartingale and in [4] for X Lévy process such that $\sum_{0 \leq s \leq t} |\Delta X_s| < \infty$ a.s. Both use the notion of local time of X along the curve b denoted $(L_s^{b(\cdot)}, s \geq 0)$, defined as

$$L_t^{b(\cdot)} = \lim_{\epsilon \rightarrow 0} \frac{1}{2\epsilon} \int_0^t 1_{(|X_s - b(s)| < \epsilon)} d[X]_s^c$$

uniformly on compacts in L^1 (17)

When b is equal to the constant a , $L^{b(\cdot)}$ coincides with the local time at the value a . These formulas have the following form:

$$\begin{aligned} G(X_t, t) &= G(X_0, 0) + \int_0^t \frac{\partial G}{\partial x}(X_{s-}, s) dX_s \\ &\quad + \int_0^t \frac{\partial G_1}{\partial t}(X_s, s) 1_{(X_s < b(s))} ds \\ &\quad + \int_0^t \frac{\partial G_2}{\partial t}(X_s, s) 1_{(X_s \geq b(s))} ds \\ &\quad + \frac{1}{2} \int_0^t \left(\frac{\partial^2 G_1}{\partial x^2}(X_s, s) 1_{(x < b(s))} \right. \\ &\quad \left. + \frac{\partial^2 G_2}{\partial x^2}(X_s, s) 1_{(x \geq b(s))} \right) d[X]_s^c \\ &\quad + \frac{1}{2} \int_0^t \left(\frac{\partial G_2}{\partial x} - \frac{\partial G_1}{\partial x} \right) (b(s), s) d_s L_s^{b(\cdot)} \\ &\quad + \sum_{0 < s \leq t} \left\{ G(X_s, s) - G(X_{s-}, s) \right. \\ &\quad \left. - \frac{\partial G}{\partial x}(X_{s-}, s) \Delta X_s \right\} \end{aligned} \quad (18)$$

Note that $\partial G / \partial x$ exists as a Radon–Nikodym derivative and is equal to $(\partial G_1 / \partial x)(x, s) 1_{(x < b(s))} + (\partial G_2 / \partial x)(x, s) 1_{(x \geq b(s))}$. The formula (18) is helpful in free-boundary problems of optimal stopping. Other illustrations of formula (13) are given in [4] for multidimensional Lévy processes.

References

- [1] Bardina X. & Jolis M. (1997). An extension of Itô's formula for elliptic diffusion processes, *Stochastic Processes and their Applications* **69**, 83–109.

- [2] Bouleau N. & Yor M. (1981). Sur la variation quadratique des temps locaux de certaines semimartingales, *Comptes Rendus de l'Académie des Sciences* **292**, 491–494.
- [3] Eisenbaum N. (2000). Integration with respect to local time, *Potential Analysis* **13**, 303–328.
- [4] Eisenbaum N. (2006). Local time-space stochastic calculus for Lévy processes, *Stochastic Processes and their Applications* **116**(5), 757–778.
- [5] Errami M., Russo F. & Vallois P. (2002). Itô formula for $C^{1,\lambda}$ -functions of a càdlàg process, *Probability Theory and Related Fields* **122**, 191–221.
- [6] Föllmer H., Protter P. & Shiriyayev A.N. (1995). Quadratic covariation and an extension of Itô's formula, *Bernoulli* **1**(1/2), 149–169.
- [7] Jacod J. & Shiriyayev A.N. (2003). *Limit Theorems for Stochastic Processes*, 2nd Edition, Springer.
- [8] Peskir G. (2005). A change-of-variable formula with local time on curves, *Journal of Theoretical Probability* **18**, 499–535.
- [9] Protter, P. (2004). *Stochastic Integration and Differential Equations*, 2nd Edition, Springer.

Related Articles

Lévy Processes; Local Times; Stochastic Integrals.

NATHALIE EISENBAUM

Lévy Copulas

Lévy copulas characterize the dependence among components of multidimensional Lévy processes. They are similar to copulas of probability distributions but are defined at the level of Lévy measures. Lévy copulas separate the dependence structure of a Lévy measure from the one-dimensional marginal measures meaning that any d -dimensional Lévy measure can be constructed from a set of one-dimensional margins and a Lévy copula. This suggests the construction of parametric multidimensional Lévy models by combining arbitrary one-dimensional Lévy processes with a Lévy copula from a parametric family. The Lévy copulas were introduced in [4] for spectrally one-sided Lévy processes and in [6, 7] in the general case. Subsequent theoretical developments include Barndorff-Nielsen and Lindner [1], who discuss further interpretations of Lévy copulas and various transformations of these objects. Farkas *et al.* [5] develop deterministic numerical methods for option pricing in models based on Lévy copulas, and the simulation algorithms for multidimensional Lévy processes based on their Lévy copulas are discussed in [4, 7].

In finance, Lévy copulas are useful to model joint moves of several assets in various settings including portfolio risk management, option pricing [8], insurance [3], and operational risk modeling [2].

Lévy Measures and Tail Integrals

A Lévy process on $\mathbb{R}^d$ is described by its characteristic triplet (A, ν, γ) , where A is a positive semidefinite $d \times d$ matrix, $\gamma \in \mathbb{R}^d$, and ν is a positive Radon measure on $\mathbb{R}^d \setminus \{0\}$, satisfying $\int_{\mathbb{R}^d \setminus \{0\}} (\|x\|^2 \wedge 1) \nu(dx) < \infty$ and called the *Lévy measure of X* . The matrix A is the covariance matrix of the continuous martingale (Brownian motion) part of X , and ν describes the independent jump part. It makes sense, therefore, to describe the dependence structure of the jump part of X with a suitable notion of copula at the level of the Lévy measure.

In the same way that the distribution of a random vector can be represented by its distribution function, the Lévy measure of a Lévy process will be represented by its tail integral. If we are only interested in,

say, positive jumps, the definition of the tail integral is simple: given a $\mathbb{R}^d$ -valued Lévy process with Lévy measure ν supported by $[0, \infty)^d$, the tail integral of ν is the function $U : (0, \infty)^d \rightarrow [0, \infty)$ defined by

$$U(x_1, \dots, x_d) = \nu((x_1, \infty) \times \dots \times (x_d, \infty)) \quad (1)$$

In the general case, care must be taken to avoid the possible singularity of ν near zero: so the tail integral is a function $U : (\mathbb{R} \setminus \{0\})^d \rightarrow \mathbb{R}$ defined by

$$U(x_1, \dots, x_d) := \prod_{i=1}^d \text{sgn}(x_i) \nu \left(\prod_{j=1}^d \mathcal{I}(x_j) \right) \quad (2)$$

where $\mathcal{I} := (x, \infty)$ if $x > 0$ and $\mathcal{I}(x) := (-\infty, x]$ if $x < 0$.

Given an $\mathbb{R}^d$ -valued Lévy process X and a nonempty set of indices $I \subset \{1, \dots, d\}$, the I margin of X is the Lévy process of lower dimension that contains only those components of X whose indices are in I : $X^I := (X^i)_{i \in I}$. The I -marginal tail integral U^I of X is then simply the tail integral of the process X^I .

Lévy Copulas: The General Case

Central to the theory of Lévy copulas are the notions of a d -increasing function and the margins of a d -increasing function. Intuitively speaking, a function F is d -increasing if dF is a positive measure on $\mathbb{R}^d$ in the sense of Lebesgue–Stieltjes integration. Similarly, the margin F^I is defined so that the measure $d(F^I)$ induced by F^I coincides with the I margin of the measure dF . Let us now turn to precise definitions.

We set $\overline{\mathbb{R}} := (-\infty, \infty]$ and for $a, b \in \overline{\mathbb{R}}^d$, we write $a \leq b$ if $a_k \leq b_k$, $k = 1, \dots, d$. In this case, $(a, b]$ denotes the interval

$$(a, b] := (a_1, b_1] \times \dots \times (a_d, b_d] \quad (3)$$

For a function $F : \overline{\mathbb{R}}^d \rightarrow \overline{\mathbb{R}}$, the F -volume of $(a, b]$ is defined by

$$V_F((a, b]) := \sum_{u \in \{a_1, b_1\} \times \dots \times \{a_d, b_d\}} (-1)^{N(u)} F(u) \quad (4)$$

where $N(u) := \#\{k : u_k = a_k\}$. In particular, $V_F((a, b]) = F(b) - F(a)$ for $d = 1$ and $V_F((a, b]) = F(b_1, b_2) + F(a_1, a_2) - F(a_1, b_2) - F(b_1, a_2)$ for

$d = 2$. If $F(u) = \prod_{i=1}^d u_i$, the F volume of any interval is equal to its Lebesgue measure.

A function $F : \overline{\mathbb{R}}^d \rightarrow \overline{\mathbb{R}}$ is called d increasing if $V_F((a, b]) \geq 0$ for all $a \leq b$. The distribution function of a random vector is one example of a d -increasing function. The tail integral U was defined in such way that $(-1)^d U$ is d increasing in every orthant (but not on the entire space).

Let $F : \overline{\mathbb{R}}^d \rightarrow \overline{\mathbb{R}}$ be a d -increasing function such that $F(u_1, \dots, u_d) = 0$ if $u_i = 0$ for at least one i . For an index set I , the I margin of F is the function $F^I : \overline{\mathbb{R}}^{|I|} \rightarrow \overline{\mathbb{R}}$, defined by

$$F^I((u_i)_{i \in I}) := \lim_{a \rightarrow \infty} \sum_{(u_i)_{i \in I^c} \in \{-a, \infty\}^{|I^c|}} \times F(u_1, \dots, u_d) \prod_{i \in I^c} \text{sgn } u_i \quad (5)$$

where $I^c := \{1, \dots, d\} \setminus I$. In particular, we have $F^{(1)}(u) = F(u, \infty) - \lim_{a \rightarrow -\infty} F(u, a)$ for $d = 2$. To understand the reasoning leading to the above definition of margins, note that any positive measure μ on $\overline{\mathbb{R}}^d$ naturally induces an increasing function F via

$$F(u_1, \dots, u_d) := \mu\left((u_1 \wedge 0, u_1 \vee 0] \times \dots \times (u_d \wedge 0, u_d \vee 0]\right) \prod_{i=1}^d \text{sgn } u_i \quad (6)$$

for $u_1, \dots, u_d \in \overline{\mathbb{R}}$. The margins of μ are usually defined by

$$\mu^I(A) = \mu\left(\{u \in \overline{\mathbb{R}}^d : (u_i)_{i \in I} \in A\}\right), \quad A \subset \overline{\mathbb{R}}^{|I|} \quad (7)$$

It is now easy to see that the margins of F are induced by the margins of μ in the sense of equation (6).

A function $F : \overline{\mathbb{R}}^d \rightarrow \overline{\mathbb{R}}$ is called *Lévy copula* if it satisfies the following four conditions (the first one is just a nontriviality requirement):

1. $F(u_1, \dots, u_d) \neq \infty$ for $(u_1, \dots, u_d) \neq (\infty, \dots, \infty)$;
2. $F(u_1, \dots, u_d) = 0$ if $u_i = 0$ for at least one $i \in \{1, \dots, d\}$;
3. F is d -increasing; and
4. $F^{(i)}(u) = u$ for any $i \in \{1, \dots, d\}$, $u \in \mathbb{R}$.

Lévy Copulas: The Spectrally One-sided Case

If X has only positive jumps in each component, or if we are only interested in the positive jumps of X , only the values $F(u_1, \dots, u_d)$ for $u_1, \dots, u_d \geq 0$ are relevant. We can then set $F(u_1, \dots, u_d) = 0$ if $u_i < 0$ for at least one i , which greatly simplifies the definition of the margins:

$$F^I((u_i)_{i \in I}) = F(u_1, \dots, u_d)|_{u_j = +\infty, j \notin I} \quad (8)$$

Taking the margins now amounts to replacing the variable that is being integrated out with infinity—exactly the same procedure as for probability distribution functions. Restricting a Lévy copula to $[0, \infty]^d$ in such way, we obtain a Lévy copula for spectrally positive Lévy processes, or, for short, a *positive Lévy copula*.

Sklar's Theorem for Lévy Processes

The following theorem [4, 7] characterizes the dependence structure of Lévy processes in terms of Lévy copulas:

Theorem 1

1. Let $X = (X^1, \dots, X^d)$ be a $\mathbb{R}^d$ -valued Lévy process. Then there exists a Lévy copula F such that the tail integrals of X satisfy

$$U^I((x_i)_{i \in I}) = F^I((U_i(x_i))_{i \in I}) \quad (9)$$

for any nonempty index set $I \subset \{1, \dots, d\}$ and any $(x_i)_{i \in I} \in (\mathbb{R} \setminus \{0\})^{|I|}$. The Lévy copula F is unique on $\prod_{i=1}^d \overline{\text{Ran } U_i}$.

2. Let F be a d -dimensional Lévy copula and $U_i, i = 1, \dots, d$, tail integrals of real-valued Lévy processes. Then there exists a $\mathbb{R}^d$ -valued Lévy process X whose components have tail integrals $U_1, \dots, U_d$ and whose marginal tail integrals satisfy equation (9) for any nonempty $I \subset \{1, \dots, d\}$ and any $(x_i)_{i \in I} \in (\mathbb{R} \setminus \{0\})^{|I|}$. The Lévy measure ν of X is uniquely determined by F and $U_i, i = 1, \dots, d$.

In particular, applying the above theorem with $I = \{1, \dots, d\}$, we obtain the usual formula

$$U(x_1, \dots, x_d) = F(U_1(x_1), \dots, U_d(x_d)) \quad (10)$$

If the one-dimensional marginal Lévy measures are infinite and have no atoms, $\text{Ran } U_i = (-\infty, 0) \cup (0, \infty)$ for any i and one can compute F directly via

$$F(u_1, \dots, u_d) = U(U_1^{-1}(u_1), \dots, U_d^{-1}(u_d)) \quad (11)$$

Examples and Parametric Families

The components of a pure-jump Lévy process are independent if and only if they never jump together, that is, if the Lévy measure is supported by the coordinate axes. This leads to a characterization of Lévy processes with independent components in terms of their Lévy copulas: the components $X^1, \dots, X^d$ of a $\mathbb{R}^d$ -valued Lévy process X are independent if and only if their Brownian motion parts are independent and if X has a Lévy copula of the form

$$F_{\perp}(x_1, \dots, x_d) := \sum_{i=1}^d x_i \prod_{j \neq i} 1_{\{\infty\}}(x_j) \quad (12)$$

The Lévy copula of independence is thus different from the copula of independent random variables $C_{\perp}(u_1, \dots, u_d) = u_1 \dots u_d$, which emphasizes the fact that the two notions are far from being the same and the “copula” intuition cannot always be applied to Lévy copulas.

The complete dependence copula, on the other hand, turns out to have a similar form to the classical case. Recall that a subset S of $\mathbb{R}^d$ is called *ordered* if, for any two vectors $u, v \in S$, either $u_k \leq v_k$, $k = 1, \dots, d$ or $u_k \geq v_k$, $k = 1, \dots, d$. Similarly, S is called *strictly ordered* if, for any two different vectors $u, v \in S$, either $u_k < v_k$, $k = 1, \dots, d$ or $u_k > v_k$, $k = 1, \dots, d$. Furthermore, set

$$K := \{x \in \mathbb{R}^d : \text{sgn } x_1 = \dots = \text{sgn } x_d\} \quad (13)$$

The jumps of an $\mathbb{R}^d$ -valued Lévy process X are said to be *completely dependent* or *comonotonic* if there exists a strictly ordered subset $S \subset K$ such that $\Delta X_t := X_t - X_{t-} \in S$, $t \in \mathbb{R}_+$ (except for some null set of paths). The condition $\Delta X_t \in K$ means that if the components of a Lévy process are comonotonic, they always jump in the same direction. A $\mathbb{R}^d$ -valued Lévy process whose Lévy measure is supported by an ordered set $S \subset K$ is described by the *complete*

dependence Lévy copula given by

$$F_{\parallel}(x) := \min(|x_1|, \dots, |x_d|) 1_K(x) \prod_{i=1}^d \text{sgn } x_i \quad (14)$$

Conversely, if $F_{\parallel}$ is a Lévy copula of X , then the Lévy measure of X is supported by an ordered subset of K . If, in addition, the tail integrals U_i of X^i are continuous and satisfy $\lim_{x \rightarrow 0} U_i(x) = \infty$, $i = 1, \dots, d$, then $F_{\parallel}$ is the unique Lévy copula of X and the jumps of X are completely dependent. For positive Lévy copulas, expression (14) simplifies to

$$F_{\parallel}(x_1, \dots, x_d) := \min(x_1, \dots, x_d) \quad (15)$$

that is, we recover the expression of the complete dependence copula of random variables (but the two functions are defined on different domains!).

One simple and convenient parametric family of positive Lévy copulas is similar to the Clayton family of copulas; it is therefore called the *Clayton–Lévy copula*:

$$F(u_1, \dots, u_d) = \left(\sum_{i=1}^d u_i^{-\theta} \right)^{-1/\theta}, \quad u_1, \dots, u_d \geq 0 \quad (16)$$

The reader can easily check that this copula converges to the complete dependence copula $F_{\parallel}$ as $\theta \rightarrow \infty$ and to the independence copula $F_{\perp}$ as $\theta \rightarrow 0$. This construction can be generalized to a Lévy copula on $\mathbb{R}^d$:

$$F(u_1, \dots, u_d) = 2^{2-d} \left(\sum_{i=1}^d |u_i|^{-\theta} \right)^{-1/\theta} \times (\eta 1_{\{u_1 \dots u_d \geq 0\}} - (1 - \eta) 1_{\{u_1 \dots u_d < 0\}}) \quad (17)$$

defines a two-parameter family of Lévy copulas. The role of the parameters is easiest to analyze in the case $d = 2$, when equation (17) becomes

$$F(u, v) = (|u|^{-\theta} + |v|^{-\theta})^{-1/\theta} \times (\eta 1_{\{uv \geq 0\}} - (1 - \eta) 1_{\{uv < 0\}}) \quad (18)$$

From this equation, it is readily seen that the parameter η determines the dependence of the *sign* of jumps: when $\eta = 1$, the two components always jump in the

same direction, and when $\eta = 0$, positive jumps in one component are accompanied by negative jumps in the other and *vice versa*. The parameter θ is responsible for the dependence of absolute values of jumps in different components.

Figure 1 shows the scatter plots of weekly returns in an exponential Lévy model with variance gamma (see **Variance-gamma Model**) margins and the dependence pattern given by the Lévy copula (18) with two different sets of dependence parameters,

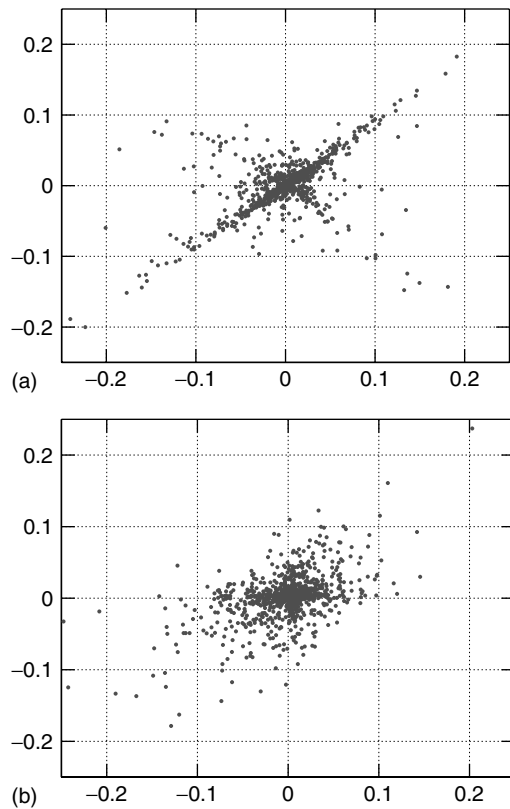


Figure 1 Scatter plots of returns in a two-dimensional variance gamma model with correlation $\rho = 50\%$ and different tail dependence. (a) Strong tail dependence ($\eta = 0.75$ and $\theta = 10$) and (b) weak tail dependence ($\eta = 0.99$ and $\theta = 0.61$)

both of which lead to a correlation of 50% but have different tail dependence patterns. It is clear that when a precise description of tail events such as simultaneous large jumps is necessary, Lévy copulas offer more freedom in modeling dependence than traditional correlation-based approaches. A natural application of Lévy copulas arises in the context of multidimensional gap options [8] that are exotic products whose payoff depends on the total number of sharp downside moves in a basket of assets.

References

- [1] Barndorff-Nielsen, O.E. & Lindner, A.M. (2007). Lévy copulas: dynamics and transforms of upilon type, *Scandinavian Journal of Statistics* **34**, 298–316.
- [2] Böcker, K. & Klüppelberg, C. (2007). Multivariate operational risk: dependence modelling with Lévy copulas, *ERM Symposium Online Monograph*, Society of Actuaries, and Joint Risk Management, section newsletter.
- [3] Bregman, Y. & Klüppelberg, C. (2005). Ruin estimation in multivariate models with Clayton dependence structure, *Scandinavian Actuarial Journal* **November**(6), 462–480.
- [4] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, Chapman & Hall/CRC Press.
- [5] Farkas, W., Reich, N. & Schwab, C. (2007). Anisotropic stable Lévy copula processes-analytical and numerical aspects, *Mathematical Models and Methods in Applied Sciences* **17**, 1405–1443.
- [6] Kallsen, J. & Tankov, P. (2006). Characterization of dependence of multidimensional Lévy processes using Lévy copulas, *Journal of Multivariate Analysis* **97**, 1551–1572.
- [7] Tankov, P. (2004). *Lévy Processes in Finance: Inverse Problems and Dependence Modelling*, PhD thesis, Ecole Polytechnique, France.
- [8] Tankov, P. (2008). *Pricing and Hedging Gap Risk*, preprint, available at <http://papers.ssrn.com>.

Related Articles

Copulas; Estimation; Exponential Lévy Models; Lévy Processes; Multivariate Distributions; Operational Risk.

PETER TANKOV

Convex Duality

Convex duality refers to a general principle that allows us to associate with an original minimization program (the *primal problem*) a class of concave maximization concave programs (the *dual problem*), which, under some conditions, are equivalent to the primal. The unifying principles underlying these methods can be traced back to the basic duality that exists between a convex set of points in the plane and the set of supporting lines (hyperplanes). Duality tools can be applied to nonconvex programs too, but are most effective for convex problems.

Convex optimization problems naturally arise in many areas of finance; we mention just few of them (see the list of the related entries at the end of this article): maximization of expected utility in complete or incomplete markets, mean–variance portfolio selection and CAPM, utility indifference pricing, selection of the minimal entropy martingale measure, and model calibration. This short and nonexhaustive list should give a hint of the scope of convex duality methods in financial applications.

Consider the following *primal* minimization (convex) problem:

$$\begin{aligned} \text{(P)} : \quad & \min f(v) \\ \text{subject to} \quad & v \in A \end{aligned} \quad (1)$$

where A is a convex subset of some vector space V and $f : A \rightarrow \mathbb{R}$ is a convex function. Convex duality principles consist in pairing this problem with a *dual* maximization (concave) problem:

$$\text{(D)} : \quad \max g(w) \quad \text{sub } w \in B \quad (2)$$

where B is a convex subset of some other vector space W (possibly $W = V$) and $g : B \rightarrow \mathbb{R}$ is a concave function.

In general, by applying a duality principle, we usually try to

1. find a lower bound for the value of the primal problem, or, better
2. find the value of the primal problem, or, even better
3. find the solutions, if any, of the primal problem.

Different duality principles differ in the way the dual problem is built. Two main principles are Lagrange duality and Fenchel duality. Even though they are formally equivalent, at least in the finite-dimensional case, they provide different insights into the problem. We will see below how the Lagrange and Fenchel duality principles practically accomplish the tasks 1 to 3 above.

For the topics to be presented below, comprehensive references are [4] and [1] for the finite-dimensional case ([1] also provides an extensive account of numerical methods) and [2] for the infinite-dimensional case.

Lagrange Duality in Finite-dimensional Problems

We consider finite-dimensional problems, that is, $V = \mathbb{R}^N$ for some $N \geq 1$. We denote $v \cdot w$ the inner product between two vectors $v, w \in \mathbb{R}^N$ and use $v \geq 0$ as a shorthand for $v_n \geq 0 \quad \forall n$. Let $f, h_1, \dots, h_M : C \rightarrow \mathbb{R}$ be $M + 1$ convex functions, where $C \subseteq \mathbb{R}^N$ is a convex set. Setting $h = (h_1, \dots, h_M)$, so that h is a convex function from C to $\mathbb{R}^M$, we consider, as the primal problem, the minimization of f under M inequality constraints:

$$\begin{aligned} \text{(P)} : \quad & \min f(v) \quad \text{sub } v \in A \\ & = \{v \in C : h(v) \leq 0\} \subset \mathbb{R}^N \end{aligned} \quad (3)$$

To build a dual problem, we define the so-called Lagrangian function

$$\begin{aligned} \mathcal{L}(v, w) &:= f(v) + w \cdot h(v) \\ v &\in C, \quad w \in \mathbb{R}^M \end{aligned} \quad (4)$$

and note that $f(v) = \sup_{w \geq 0} \mathcal{L}(v, w)$ for any $v \in A$. As a consequence, we can write the primal problem in terms of $\mathcal{L}$:

$$\text{(P)} : \quad \inf_{v \in C} \sup_{w \geq 0} \mathcal{L}(v, w) \quad (5)$$

The dual problem is then defined by switching the supremum with the infimum

$$\text{(D)} : \quad \sup_{w \geq 0} \inf_{v \in C} \mathcal{L}(v, w) \quad (6)$$

In the terminology of the introductory section, the dual problem is then

$$(D) : \quad \max g(w) \quad \text{sub } w \in B \\ = \{w \in D : w \geq 0\} \subset \mathbb{R}^M \quad (7)$$

where

$$g(w) = \inf_{v \in C} \mathcal{L}(v, w) \quad (8)$$

and $D = \{w \in \mathbb{R}^M : g(w) > -\infty\}$ is the domain of g . It can be proved that D is a convex set and g is a concave function on D even if f is not convex: therefore *the dual problem is always concave*, even when the primal problem is not convex.

We assume throughout *primal and dual feasibility*, that is, A and B are assumed to be nonempty. Dual feasibility would however be ensured under Slater conditions for A (see below). Let $p = \inf_A f$ and $d = \sup_B g$ be the (possibly infinite) values of the primal and the dual. A *primal (dual) solution* is $\hat{v} \in A$ ($\hat{w} \in B$), if any, such that $f(\hat{v}) = p$ ($g(\hat{w}) = d$); a *solution pair* is a feasible pair $(\hat{v}, \hat{w}) \in A \times B$ made by a primal and a dual solution.

Lagrange Duality Theorem

1. Weak duality

Primal boundedness ($p > -\infty$) implies dual boundedness ($d < +\infty$) and

$$p \geq d \quad (p - d \geq 0 \text{ is called the duality gap}) \quad (9)$$

Moreover, if there is no duality gap ($p = d$), then $(\hat{v}, \hat{w}) \in A \times B$ is a solution pair if and only if

$$\hat{w} \cdot h(\hat{v}) = 0 \quad \text{and} \quad \mathcal{L}(\hat{v}, \hat{w}) = g(\hat{w}) \quad (10)$$

In this case, $\hat{w}$ is usually called a Lagrange multipliers vector.

2. Strong duality

If, in addition, there exists $v \in C$ such that $h_m(v) < 0$ for all m (Slater condition), then there is no duality gap and there exists a dual solution.

See [4] or [1] for a proof.

Weak duality, whose proof is trivial, holds under very general conditions: in particular, the primal problem need not be convex. It gives a lower bound for the value of the primal problem, which is useful in many

practical situations, “branch and bound” algorithms in integer programming being a prominent example. It also provides a workable condition that characterizes a solution pair, at least when there is no duality gap.

Strong duality, on the contrary, requires a precise topological assumption: the interior of the constraint set has to be nonempty (Slater condition). We note, however, that this condition is satisfied in most cases, at least in the present finite-dimensional setting. The proof is then based on a separating hyperplane theorem, that in turn requires convexity assumptions about f and h . When strong duality holds, and provided we are able to actually solve the dual problem, we obtain the exact value of the primal (no duality gap).

We can add a finite number (say L) of linear equality constraints to (P), obtaining

$$(P) : \quad \min f(v) \quad \text{sub } v \in A \\ = \{v \in C : h(v) \leq 0, Qv = r\} \subset \mathbb{R}^N \quad (11)$$

where Q is an $L \times N$ matrix and $r \in \mathbb{R}^L$. The Lagrangian is defined as

$$\mathcal{L}(v, w) = f(v) + w^{\text{in}} \cdot h(v) + w^{\text{eq}} \cdot (Qv - r) \\ v \in C, w = (w^{\text{in}}, w^{\text{eq}}) \in \mathbb{R}^{M \times L} \quad (12)$$

in such a way that

$$\inf_{v \in A} f(v) = \inf_{v \in C} \sup_{w^{\text{in}} \geq 0, w^{\text{eq}} \in \mathbb{R}^L} \mathcal{L}(v, w) \quad (13)$$

The dual problem is then

$$(D) : \quad \max g(w) \quad \text{sub } w \in B \\ = \{w \in D : w^{\text{in}} \geq 0\} \subset \mathbb{R}^{M \times L} \quad (14)$$

where, as before, $g(w) = \inf_{v \in C} \mathcal{L}(v, w)$, and D is the domain of g . It is worth noting that if the primal problem has equality constraints only, then the only constraint of the dual problem is $w \in D$.

A Lagrange duality theorem can then be stated and also proved in this case, reaching similar conclusions. We have just to replace $\hat{w}$ with $\hat{w}^{\text{in}}$ in the first condition in (10), and modify the Slater condition as follows:

- *There exists $v \in \text{ri}(C)$ such that $h_m(v) < 0$ for all m and $Qv = r$* (15)

The relative interior $\text{ri}(C)$ is the interior of the convex set C relative to the affine hull of C . For instance, if $C = [0, 1] \times \{0\} \subset \mathbb{R}^2$, then $\text{ri}(C) = (0, 1) \times \{0\}$ (because the affine hull of C is $\mathbb{R} \times \{0\}$), while the interior of C is clearly empty (see [4] for more on relative interiors and related topics about convex sets).

In many concrete problems, C is a polyhedron, that is, it is the (convex and closed) set defined by a certain finite set of linear inequalities, and all the functions h_m are affine. If we assume, in addition, that f may be extended to a finite convex function over *all* $\mathbb{R}^N$, Farkas Lemma allows us to prove strong duality without requiring any Slater condition. Remarkably, if f is linear too, then the existence of a primal solution is ensured.

The Lagrange duality theorem provides us a simple criterion for the existence of a dual solution and a set of conditions characterizing a possible primal solution. It is, however, not directly concerned with the *existence* of a primal solution. To ensure this, one has to assume stronger conditions such as compactness of C or coercivity of f . A third condition (f linear) has been described above.

We have seen that the dual problem usually looks much better than the primal: it is always concave and its solvability is guaranteed under mild assumptions about the primal. This fact is particularly useful in designing numerical procedures. Moreover, even when the primal is solvable, the dual often proves easier to handle. We provide a simple example that should clarify the point.

A standard linear programming (LP) problem comes, by definition, in the form

$$(P) : \quad \min c \cdot v \quad \text{sub } Qv = r, \quad v \geq 0, \quad v \in \mathbb{R}^N \quad (16)$$

where $c \in \mathbb{R}^N$, Q is a $L \times N$ matrix and $r \in \mathbb{R}^L$. An easy computation shows that the dual problem is (T denotes transposition)

$$(D) : \quad \max r \cdot w \quad \text{sub } Q^T w \leq c, \quad w \in \mathbb{R}^L \quad (17)$$

We know that strong duality holds in this case, and that the existence of a solution pair is guaranteed. In particular, $(Q^T \hat{w} - c) \cdot \hat{v} = 0$ is a necessary condition for a pair $(\hat{v}, \hat{w})$ to be a solution. The dual problem, however, has L variables and N constraints and thus can often be more tractable than the primal

if N is much larger than L . This is the basis for great enhancements in existing numerical methods.

A last remark concerns the word “duality”: any dual problem can be turned into an equivalent minimization primal problem. It turns out that the bidual, that is, the dual of this new primal problem, seldom coincides with the original primal problem. LP problems are an important exception: the bidual of an LP problem is the problem itself.

Fenchel Duality in Finite-dimensional Problems

Fenchel duality, that we will derive from Lagrange duality, may be applied to primal problems in the form

$$(P) : \quad \min \{f_1(v) - f_2(v)\} \\ \text{sub } v \in A = C_1 \cap C_2 \subset \mathbb{R}^N \quad (18)$$

where $C_1, C_2 \subseteq \mathbb{R}^N$ are convex, $f_1 : C_1 \rightarrow \mathbb{R}$ is convex, and $f_2 : C_2 \rightarrow \mathbb{R}$ is concave.

Consider the function $f(x, y) = f_1(x) - f_2(y)$ defined on $\mathbb{R}^{2N}$ and clearly convex. We can restate the primal as

$$(P) : \quad \min f(x, y) \quad \text{sub } (x, y) \in A' \\ = \{(x, y) \in C_1 \times C_2 : x = y\} \subset \mathbb{R}^{2N} \quad (19)$$

where the N fictitious linear constraints $(x_n = y_n \forall n)$ allow us to apply the Lagrange duality machinery. The Lagrangian function is $\mathcal{L}(x, y, w) = f_1(x) - f_2(y) + w \cdot (x - y)$ and, using some simple algebra, we compute

$$g(w) = \inf_{x \in C_1, y \in C_2} \mathcal{L}(x, y, w) = f_2^*(w) - f_1^*(w) \quad (20)$$

where

$$f_1^*(w) = \sup_{x \in C_1} \{w \cdot x - f_1(x)\} \quad (21)$$

is, by definition, the convex conjugate (indeed, f_1^* is convex) of the convex function f_1 , and

$$f_2^*(w) = \inf_{y \in C_2} \{w \cdot y - f_2(y)\} \quad (22)$$

is the concave conjugate (indeed, f_2^* is concave) of the concave function f_2 . As a consequence, the dual problem is

$$(D) : \max \{f_2^*(w) - f_1^*(w)\} \\ \text{sub } w \in B = C_1^* \cap C_2^* \subset \mathbb{R}^N \quad (23)$$

where C_1^* and C_2^* are the domains of f_1^* and f_2^* , respectively. Assuming primal feasibility and boundedness, the Lagrange duality theorem yields the Fenchel duality theorem.

Fenchel Duality Theorem

1. Weak duality

If there is no duality gap, $(\hat{v}, \hat{w})$ is a solution pair if and only if

$$\hat{v} \cdot \hat{w} = f_1(\hat{v}) + f_1^*(\hat{w}) = f_2(\hat{v}) + f_2^*(\hat{w}) \quad (24)$$

2. Strong duality

There is no duality gap between the primal and the dual, and there is a dual solution, provided one of the following conditions is satisfied:

- (a) $\text{ri}(C_1) \cap \text{ri}(C_2)$ is nonempty
- (b) C_1 and C_2 are polyhedra and f_1 (resp. f_2) may be extended to a finite convex (concave) function over all $\mathbb{R}^N$

See [4] or [1] for a proof.

We say that a convex function f is closed if, for any $a \in \mathbb{R}$, the set $\Gamma_a = \{v : f(v) \leq a\}$ is closed; a similar definition applies to concave functions, where the inequality inside Γ_a is reversed. A sufficient, though not necessary condition for f to be closed is continuity on all C . A celebrated result (the Fenchel–Moreau theorem) states that $(f^*)^* \equiv f$, provided f is a closed (convex or concave) function. Therefore, if in the primal problem f_1 and f_2 are closed, then the dual problem of the dual coincides with the primal, and the duality is therefore *complete*. Thanks to this fact, an application of the Fenchel duality theorem to the dual problem allows us to state that the primal has a solution provided one of the following conditions is satisfied:

1. $\text{ri}(C_1^*) \cap \text{ri}(C_2^*)$ is nonempty.
2. C_1^* and C_2^* are polyhedra, and f_1^* (resp. f_2^*) may be extended to a finite convex (concave) function over all $\mathbb{R}^N$.

Fenchel duality can sometimes be effectively used for general problems in the form

$$(P) : \min f(v) \quad \text{sub } v \in C \subset \mathbb{R}^N \quad (25)$$

where f and C are convex. Indeed, such a problem can be cast in the form (18) provided we set $f_1 = f$, $f_2 = 0$ (concave), $C_1 = \mathbb{R}^N$, and $C_2 = C$. The dual problem is given by equation (23), where

$$f_1^*(w) = \sup_{v \in \mathbb{R}^N} \{w \cdot v - f(v)\} \quad (26)$$

is an unconstrained problem and

$$f_2^*(w) = \inf_{v \in C} w \cdot v \quad (27)$$

has a simple goal function.

We have derived Fenchel duality as a by product of Lagrange duality. However, it is possible to go in the opposite direction, by first proving Fenchel duality (unsurprisingly, using hyperplane separation arguments, see [2]) and then writing a Lagrange problem in the Fenchel form, so that Lagrange duality can be derived (see [3]). Therefore, at least in the finite-dimensional setting, Lagrange and Fenchel duality are formally equivalent.

Duality in Infinite-dimensional Problems

For infinite-dimensional problems, Lagrange or Fenchel duality exhibit a large formal similarity with the finite-dimensional counterparts we have described so far. Nevertheless, the technical topological assumptions, which are needed to ensure duality, become much less trivial when the space $V = \mathbb{R}^N$ is replaced by an infinite-dimensional Banach space. We give a brief account of these differences.

Let V be a Banach space and consider the primal problem

$$(P) : \min f(v) \quad \text{sub } v \in A \\ = \{v \in C : h(v) \leq 0\} \subset V \quad (28)$$

where $C \subseteq V$ is a convex set, and $f : C \rightarrow \mathbb{R}$ and $h : C \rightarrow \mathbb{R}^M$ are convex functions. Then, by mimicking the finite-dimensional case, the dual problem is

$$(D) : \max g(w) \quad \text{sub } w \in B \\ = \{w \in D : w \geq 0\} \subset \mathbb{R}^M \quad (29)$$

where $g(w) = \inf_{v \in C} \{f(v) + w \cdot h(v)\}$, and D is the domain of g . We can note that the dual is finite-dimensional, but the definition of g involves an infinite-dimensional problem. A perfect analog of the finite-dimensional Lagrange duality theorem may be derived in this more general case too (see [2]) with essentially the same Slater condition (existence of some $v \in C$ such that $h_m(v) < 0$ for any m). We can also introduce a finite set of linear inequalities: this case can be handled in exactly the same way as in the finite-dimensional case. However, the hypothesis $\text{ri}(C) \neq \emptyset$ is not completely trivial here.

Fenchel duality too can be much generalized. Indeed, let V be a Banach space, $W = V^*$ its dual space (the Banach space of continuous linear forms on V), and denote by $\langle v, v^* \rangle$ the action of $v^* \in V^*$ on $v \in V$. Consider the primal problem

$$\begin{aligned} (\text{P}) : \quad & \min \{f_1(v) - f_2(v)\} \quad \text{sub } v \in A \\ & = C_1 \cap C_2 \subset V \end{aligned} \quad (30)$$

where $C_1, C_2 \subseteq V$ are convex sets, f_1 is convex on C_1 , and f_2 is concave on C_2 . Then, again by mimicking the finite-dimensional case, we associate the primal with the dual

$$\begin{aligned} (\text{D}) : \quad & \max \{f_2^*(v^*) - f_1^*(v^*)\} \quad \text{sub } v^* \in B \\ & = C_1^* \cap C_2^* \subset V^* \end{aligned} \quad (31)$$

where

$$\begin{aligned} f_1^*(v^*) &= \sup_{v \in C_1} \{\langle v, v^* \rangle - f_1(v)\} \quad \text{and} \quad f_2^*(v^*) \\ &= \inf_{v \in C_2} \{\langle v, v^* \rangle - f_2(v)\} \end{aligned} \quad (32)$$

are the convex and concave conjugates of f_1 and f_2 , respectively, and C_1^* and C_2^* are their domains. Then, with obvious formal modifications, Fenchel duality theorem holds in this case, too (see again [2]). However, to obtain strong duality, we must supplement conditions (a) or (b) with the following

- Either $\{(v, a) \in V \times \mathbb{R} : f_1(v) \leq a\}$
or $\{(v, a) \in V \times \mathbb{R} : f_2(v) \geq a\}$
has a nonempty interior.

This latter condition, which, in the finite-dimensional setting, follows from (a) or (b), must be checked separately in the present case.

References

- [1] Bertsekas, D.P. (1995). *Nonlinear Programming*, Athena Scientific, Belmont.
- [2] Luenberger, D.G. (1969). *Optimization by Vector Space Methods*, Wiley, New York.
- [3] Magnanti, T.L. (1974). Fenchel and Lagrange duality are equivalent, *Mathematical Programming* **7**, 253–258.
- [4] Rockafellar, R.T. (1970). *Convex Analysis*, Princeton University Press, Princeton.

Related Articles

Capital Asset Pricing Model; Expected Utility Maximization; Expected Utility Maximization: Duality Methods; Minimal Entropy Martingale Measure; Model Calibration; Optimization Methods; Risk–Return Analysis; Robust Portfolio Optimization; Stochastic Control; Utility Function; Utility Indifference Valuation.

GIACOMO SCANDOLO

Squared Bessel Processes

Squares of Bessel processes enjoy both an additivity property and a scaling property, which are, arguably, the main reasons why these processes occur naturally in a number of Brownian, or linear diffusion, studies. This survey is written in a minimalist manner; the aim is to refer the reader to a few references where many facts and formulae are discussed in detail.

Squared Bessel (BESQ) Processes

A squared Bessel (BESQ) process $(X_t^{(x,\delta)}, t \geq 0)$ may be defined (in law) as the solution of the stochastic differential equation:

$$X_t = x + 2 \int_0^t \sqrt{X_s} d\beta_s + \delta t, \quad X_t \geq 0 \quad (1)$$

where x is the starting value: $X_0 = x$, δ is the so-called dimension of X , and $(\beta_s)_{s \geq 0}$ is standard Brownian motion. For any integer dimension δ , $(X_t, t \geq 0)$ may be obtained as the square of the Euclidean norm of δ -dimensional Brownian motion.

The general theory of stochastic differential equations (SDEs) ensures that equation (1) enjoys pathwise uniqueness, hence uniqueness in law, and consequently the strong Markov property. Denoting by Q_x^δ the law of $(X_t)_{t \geq 0}$, solution of equation (1), on the canonical space $C_+ \equiv C(\mathbb{R}_+, \mathbb{R}_+)$, where $(Z_u, u \geq 0)$ is taken as the coordinate process, there is the convolution property:

$$Q_x^\delta * Q_{x'}^{\delta'} = Q_{x+x'}^{\delta+\delta'} \quad (2)$$

which holds for all $x, \delta \geq 0$ ([7]); in other terms, adding two independent BESQ processes yields another BESQ process, whose starting point, respectively dimension, is the sum of the starting points, respectively dimensions.

It follows from equation (2) that for any positive measure $\mu(du)$ on $\mathbb{R}_+$ such that $\int \mu(du)(1+u) < \infty$, then, if $I_\mu = \int \mu(du)Z_u$,

$$Q_x^\delta \left(\exp - \frac{1}{2} I_\mu \right) = (A_\mu)^\delta (B_\mu)^x \quad (3)$$

with $A_\mu = (\phi_\mu(\infty))^{1/2}$; $B_\mu = \exp(\phi'_\mu(0+))$ for ϕ_μ , the unique decreasing solution of the Sturm–Liouville equation: $\phi'' = \phi\mu$; $\phi(0) = 1$.

Equation (3) may be considered as the (generalized) Laplace transform (with argument μ) of the probability Q_x^δ , while as Q_x^δ , for any fixed δ and x , is infinitely divisible, the next formula is the Lévy Khintchine representation of Q_x^δ :

$$Q_x^\delta \left(\exp - \frac{1}{2} I_\mu \right) = \exp \left(- \int_{C_+} M_{x,\delta}(dz) \left(1 - e^{-\frac{1}{2} I_\mu(z)} \right) \right) \quad (4)$$

where $M_{x,\delta} = xM + \delta N$, for M and N two σ -finite measures on C_+ , which are described in detail in, for example, [5] and [6].

Brownian Local Times and BESQ Processes

The Ray–Knight theorems for Brownian local times $(L_t^y; y \in \mathbb{R}, t \geq 0)$ express the laws of $(L_T^y; y \in \mathbb{R})$ for some very particular stopping times in terms of certain Q_x^δ 's, namely,

1. if $T = T_a$ is the first hitting time of a by Brownian motion then $\tilde{Z}_y^{(a)} \equiv L_{T_a}^{a-y}$, $y \geq 0$, satisfies the following:

$$\tilde{Z}_y = 2 \int_0^y \sqrt{\tilde{Z}_z} d\beta_z + 2(y \wedge a) \quad (5)$$

2. if $T = \tau_\ell$ is the first time $(L_t^0, t \geq 0)$ the Brownian local time at level 0 reaches ℓ , then $(L_{\tau_\ell}^y, y \geq 0)$ and $(L_{\tau_\ell}^{-y}, y \geq 0)$ are two independent BESQ processes, distributed as Q_ℓ^0 .

An Implicit Representation in Terms of Geometric Brownian Motions

Lamperti [3] showed a one-to-one correspondence between Lévy processes $(\xi_t, t \geq 0)$ and semistable Markov processes $(\Sigma_u, u \geq 0)$ via the (implicit) formula:

$$\exp(\xi_t) = \Sigma \int_0^t ds \exp(\xi_s), \quad t \geq 0 \quad (6)$$

In the particular case where $\xi_t = 2(B_t + \nu t)$, $t \geq 0$, formula (6) becomes

$$\exp(2(B_t + \nu t)) = X^{(1,\delta)} \int_0^t ds \exp(2(B_s + \nu s)) \quad (7)$$

where, in agreement with our notation, $(X_u^{(1,\delta)}, u \geq 0)$ denotes a BESQ process starting from 1 with dimension $\delta = 2(1 + \nu)$. We note that in equation (7), δ may be negative, that is, $\nu < -1$; however, formula (7) reveals $(X_u^{(1,\delta)})$ for $u \leq T_0(X^{(1,\delta)})$ the first hitting time of 0 by $(X^{(1,\delta)})$. Nonetheless, the study of BESQ^δ , for any $\delta \in \mathbb{R}$, has been developed in [1].

Absolute continuity relationships between the laws of different BESQ processes may be derived from equation (7), combined with the Cameron–Martin relationship between the laws of $(B_t + \nu t, t \geq 0)$ and $(B_t, t \geq 0)$.

Precisely, one obtains thus, for $\delta \geq 2$:

$$\mathcal{Q}_{x|Z_u}^\delta = \left(\frac{Z_u}{x}\right)^{\frac{\nu}{2}} \exp\left(-\frac{\nu^2}{2} \int_0^u \frac{ds}{Z_s}\right) \cdot \mathcal{Q}_{x|Z_u}^2 \quad (8)$$

where $Z_u \equiv \sigma\{Z_s, s \leq u\}$, and $\nu = \frac{\delta}{2} - 1$. The combination of equations (7) and (8) may be used to derive results about $(B_t + \nu t, t \geq 0)$ from results about $X^{x,\delta}$ (and *vice versa*). In particular, the law of

$$A_{T_\lambda}^{(\nu)} := \int_0^{T_\lambda} ds \exp(2(B_s + \nu s)) \quad (9)$$

where T_λ denotes an independent exponential time, was derived in ([8], Paper #2) from this combination.

Some Explicit Formulae for BESQ Functionals

Formula (3), when μ is replaced by $\lambda\mu$, for any scalar $\lambda \geq 0$, yields the explicit Laplace transform

of I_μ , provided the function $\phi_{\lambda,\mu}$ is known explicitly, which is the case for $\mu(dt) = at^\alpha 1_{(t \leq A)} dt + b\varepsilon_A(dt)$ and many other examples.

Consequently, the semigroup of BESQ may be expressed explicitly in terms of Bessel functions, as well as the Laplace transforms of first hitting times (see, for example, [2]) and distributions of last passage times (see, for example, [4]). Chapter XI of [6] is entirely devoted to Bessel processes.

References

- [1] Goïng-Jaeschke, A. & Yor, M. (2003). A survey and some generalizations of Bessel processes, *Bernoulli* **9**(2), 313–350.
- [2] Kent, J. (1978). Some probabilistic properties of Bessel functions, *The Annals of Probability* **6**, 760–770.
- [3] Lamperti, J. (1972). Semi-stable Markov processes, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **22**, 205–225.
- [4] Pitman, J. & Yor, M. (1981). Bessel processes and infinitely divisible laws, in *Stochastic Integrals*, D. Williams, ed., LNM 851, Springer, pp. 285–370.
- [5] Pitman, J. & Yor, M. (1982). A decomposition of Bessel Bridges, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **59**, 425–457.
- [6] Revuz, D. & Yor, M. (1999). *Continuous Martingales and Brownian Motion*, 3rd Edition, Springer.
- [7] Shiga, T. & Watanabe, S. (1973). Bessel diffusions as a one-parameter family of diffusion processes, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **27**, 37–46.
- [8] Yor, M. (2001). *Exponential Functionals of Brownian Motion and Related Processes*, Springer-Finance.

Related Articles

Affine Models; Cox–Ingersoll–Ross (CIR) Model; Heston Model; Simulation of Square-root Processes.

MARC J. YOR

Semimartingale

Semimartingales form an important class of processes in probability theory, especially in the theory of **stochastic integration** and its applications. They serve as natural models for asset pricing, since under no-arbitrage assumptions a price process must be a semimartingale [1, 3].

Let $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t \geq 0}, P)$ be a complete probability space that satisfies the *usual assumptions* (i.e., $\mathcal{F}_0$ contains all P -null sets of $\mathcal{F}$ and the filtration $\mathbb{F}$ is right continuous). A càglàd, adapted process X is called a *semimartingale* if it admits a decomposition

$$X_t = X_0 + A_t + M_t \quad (1)$$

where X_0 is $\mathcal{F}_0$ -measurable, A is a process with finite variation, M is a local martingale, and $A_0 = M_0 = 0$. If, moreover, A is predictable (i.e., measurable with respect to the σ -algebra generated by all left-continuous processes), X is called a *special semimartingale*. In this case, the decomposition (1) is unique and we call it the *canonical decomposition*. Clearly, the set of all semimartingales is a vector space.

For any $a > 0$, a semimartingale X can be further decomposed as

$$X_t = X_0 + A_t + D_t + N_t \quad (2)$$

where D and N are local martingales such that D is a process with finite variation and the jumps of N are bounded by $2a$ (see [6] p. 126).

Alternatively, semimartingales can be defined as a class of “good integrators”. Let S be a collection of all simple predictable processes equipped with the uniform convergence in (t, ω) . A process H is called *simple predictable* if it has the representation

$$H_t = H_0 1_{\{0\}}(t) + \sum_{i=1}^n H_i 1_{(T_i, T_{i+1}]}(t) \quad (3)$$

where $0 = T_1 \leq \dots \leq T_{n+1} < \infty$ are stopping times, H_i are $\mathcal{F}_{T_i}$ -measurable and $|H_i| < \infty$ almost surely. Let L^0 be the space of (finite-valued) random variables topologized by convergence in probability. For a given process X , we define a linear mapping (*stochastic integral*) $I_X : S \rightarrow L^0$ by

$$I_X(H) = H_0 X_0 + \sum_{i=1}^n H_i (X_{T_{i+1}} - X_{T_i}) \quad (4)$$

A process X is defined to be a semimartingale if it is càglàd, adapted, and the mapping $I_X : S \rightarrow L^0$ is continuous. Such processes are “good integrators”, because they satisfy the following bounded convergence theorem: the uniform convergence of H^n to H (in S) implies the convergence in probability of $I_X(H^n)$ to $I_X(H)$. As a consequence, when X is a semimartingale, the domain of the stochastic integral I_X can be extended to the space of all predictable processes H (*see Stochastic Integrals*).

Indeed, these two definitions are equivalent. This result is known as the *Bichteler–Dellacherie theorem* [2, 4].

Examples

- Càglàd adapted processes with finite variation are semimartingales.
- All càglàd, adapted martingales, submartingales, and supermartingales are semimartingales.
- Brownian motion is a continuous martingale. Hence, it is a semimartingale.
- **Lévy processes** are semimartingales.
- Itô diffusions of the form

$$X_t = X_0 + \int_0^t a_s ds + \int_0^t \sigma_s dW_s \quad (5)$$

where W is a Brownian motion, are (continuous) semimartingales. In particular, solutions of stochastic differential equations of the type $dX_t = a(t, X_t)dt + \sigma(t, X_t)dW_t$ are semimartingales.

Quadratic Variation of Semimartingales

Quadratic variation is an important characteristic of a semimartingale. It is also one of the crucial objects in financial econometrics as it serves as a measure of the variability of a price process.

Let X, Y be semimartingales. The quadratic variation process $[X, X] = ([X, X]_t)_{t \geq 0}$ is given as

$$[X, X]_t = X_t^2 - X_0^2 - 2 \int_0^t X_{s-} dX_s \quad (6)$$

where $X_{s-} = \lim_{u < s, u \rightarrow s} X_u$ ($X_{0-} = X_0$). The quadratic covariation of X and Y is defined by

$$[X, Y]_t = X_t Y_t - X_0 Y_0 - \int_0^t X_{s-} dY_s - \int_0^t Y_{s-} dX_s \quad (7)$$

which is also known as the *integration by parts* formula (see [5] p. 51). Obviously, the operator $(X, Y) \rightarrow [X, Y]$ is symmetric and bilinear. We therefore have the polarization identity

$$[X, Y] = \frac{1}{2}([X + Y, X + Y] - [X, X] - [Y, Y]) \quad (8)$$

The quadratic (co-)variation process has the following properties:

1. $\Delta[X, Y] = \Delta X \Delta Y$ with $\Delta Z_s = Z_s - Z_{s-}$ ($\Delta Z_0 = 0$) for any càglàd process Z .
2. $[X, Y]$ has finite variation and $[X, X]$ is an increasing process.
3. Let A, B be càglàd, adapted processes. Then it holds that

$$\left[\int_0^t A_s dX_s, \int_0^t B_s dY_s \right]_t = \int_0^t A_s B_s d[X, Y]_s \quad (9)$$

Furthermore, the quadratic variation process can be written as a sum of its continuous and discontinuous parts:

$$[X, X]_t = [X, X]_t^c + \sum_{0 \leq s \leq t} |\Delta X_s|^2 \quad (10)$$

where $[X, X]^c$ denotes the continuous part of $[X, X]$. A semimartingale X is called *quadratic pure jump* if $[X, X]^c = 0$.

For any subdivision $0 = t_0^n < \dots < t_{k_n}^n = t$ with $\max_i |t_i^n - t_{i-1}^n| \rightarrow 0$, it holds that

$$\sum_{i=1}^{k_n} (X_{t_i^n} - X_{t_{i-1}^n})(Y_{t_i^n} - Y_{t_{i-1}^n}) \xrightarrow{P} [X, Y]_t \quad (11)$$

The latter suggests the *realized variance* as a natural consistent estimator of the quadratic variation (see **Realized Volatility and Multipower Variation**).

Stability Properties of Semimartingales

Semimartingales turn out to be invariant under change of measure. Indeed, if Q is a probability measure that is absolutely continuous with respect to P , then every P -semimartingale is a Q -semimartingale. When X is a P -semimartingale with decomposition (1) and P, Q are equivalent probability measures, then X is a Q -semimartingale with the decomposition $X_t = X_0 + \tilde{A}_t + \tilde{M}_t$, where

$$\tilde{M}_t = M_t - \int_0^t \frac{1}{Z_s} d[Z, M]_s \quad (12)$$

$Z_t = E_P\left(\frac{dQ}{dP} | \mathcal{F}_t\right)$ and $\tilde{A}_t = X_t - X_0 - \tilde{M}_t$. The latter result is known as *Girsanov's Theorem* (see [6] p. 133).

Furthermore, semimartingales are stable under certain changes of filtration. Let X be a semimartingale for the filtration $\mathbb{F}$. If $\mathbb{G} \subset \mathbb{F}$ is a subfiltration and X is adapted to $\mathbb{G}$, then X is a semimartingale for $\mathbb{G}$ (*Stricker's Theorem*). Semimartingales are also invariant to certain enlargement of filtration. Let $\mathcal{A} \subset \mathcal{F}$ be a collection of events such that $A, B \in \mathcal{A}$, $A \neq B$, implies $A \cap B = \emptyset$. Let $\mathcal{H}_t$ be generated by $\mathcal{F}_t$ and $\mathcal{A}$. Then every $(\mathbb{F}, P)$ -semimartingale is a $(\mathbb{H}, P)$ -semimartingale (*Jacod's Countable Expansion*).

Itô's Formula

Semimartingales are stable under C^2 -transformation. Let $X = (X^1, \dots, X^d)$ be a d -dimensional semimartingale and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a function with continuous second-order partial derivatives. Then $f(X)$ is again a semimartingale and the **Itô's Formula** holds:

$$\begin{aligned} f(X_t) - f(X_0) &= \sum_{i=1}^d \int_0^t \frac{\partial f}{\partial x_i}(X_{s-}) dX_s^i \\ &+ \frac{1}{2} \sum_{i,j=1}^d \int_0^t \frac{\partial^2 f}{\partial x_i \partial x_j}(X_{s-}) d[X^i, X^j]_s^c \\ &+ \sum_{0 \leq s \leq t} \left(f(X_s) - f(X_{s-}) \right. \\ &\quad \left. - \sum_{i=1}^d \frac{\partial f}{\partial x_i}(X_{s-}) \Delta X_s^i \right) \end{aligned} \quad (13)$$

One of the most interesting applications of **Itô's formula** is the so-called *Doléans–Dade exponential* (see **Stochastic Exponential**). Let X be a (one-dimensional) semimartingale with $X_0 = 0$. Then there exists a unique semimartingale Z that satisfies the equation $Z_t = 1 + \int_0^t Z_{s-} dX_s$. This solution is denoted by $\mathcal{E}(X)$ (the Doléans–Dade exponential) and is given by

$$\begin{aligned} \mathcal{E}(X)_t = & \exp\left(X_t - \frac{1}{2}[X, X]_t\right) \prod_{0 \leq s \leq t} (1 + \Delta X_s) \\ & \times \exp\left(-\Delta X_s + \frac{1}{2}|\Delta X_s|^2\right) \end{aligned} \quad (14)$$

Moreover, we obtain the identity $\mathcal{E}(X)\mathcal{E}(Y) = \mathcal{E}(X + Y + [X, Y])$.

An important example is $X_t = at + \sigma W_t$, where W denotes the Brownian motion and a, σ are constant. In this case, the continuous solution $\mathcal{E}(X)_t = \exp\left(\left(a - \frac{\sigma^2}{2}\right)t + \sigma W_t\right)$ is known as the *Black–Scholes model*.

References

- [1] Back, K. (1991). Asset prices for general processes, *Journal of Mathematical Economics* **20**(4), 371–395.

- [2] Bichteler, K. (1981). Stochastic integration and L_p -theory of semimartingales, *Annals of Probability* **9**, 49–89.
 [3] Delbaen, F. & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing, *Mathematische Annalen* **300**, 463–520.
 [4] Dellacherie, C. (1980). Un survol de la théorie de l'intégrale stochastique, *Stochastic Processes and their Applications* **10**, 115–144.
 [5] Jacod, J. & Shiryaev, A.N. (2003). *Limit Theorems for Stochastic Processes*, 2nd Edition, Springer-Verlag.
 [6] Protter, P.E. (2005). *Stochastic Integration and Differential Equations*, 2nd Edition, Springer-Verlag.

Further Reading

- Revuz, D. & Yor, M. (2005). *Continuous Martingales and Brownian Motion*, 3rd Edition, Springer-Verlag.

Related Articles

Doob–Meyer Decomposition; Equivalence of Probability Measures; Filtrations; Itô's Formula; Martingales; Poisson Process; Stochastic Exponential; Stochastic Integrals.

MARK PODOLSKIJ

Capital Asset Pricing Model

The 1990 Nobel Prize winner William Sharpe [49, 50] introduced one cornerstone of the modern finance theory with his seminal capital asset pricing model (CAPM) for which Black [9], Lintner [35, 36], Mossin [43], and Treynor [54] proposed analogous and extended versions. He then proposed an answer to the financial theory's question about the uncertainty surrounding any investment and any financial asset. Indeed, financial theory raised the question of how risk impacts the fixing of asset prices in the financial market (*see* **Modern Portfolio Theory**), and William Sharpe proposed an explanation of the link prevailing between risky asset prices and market equilibrium. The CAPM therefore proposes a characterization of the link between the risk and return of financial assets, on one side, and market equilibrium, on the other side. This fundamental relationship establishes that the expected excess return of a given risky asset (*see* **Expectations Hypothesis; Risk Premia**) corresponds to the expected market risk premium (i.e., market price of risk) times a constant parameter called *beta* (i.e., a proportionality constant). The beta is a measure of the asset's relative risk and represents the asset price's propensity to move with the market. Indeed, the beta assesses the extent to which the asset's price follows the market trend simultaneously. Namely, the CAPM explains that, on an average basis, the unique source of risk impacting the returns of risky assets comes from the broad financial market to which all the risky assets belong and on which they are all traded. The main result is that the global risk of a given financial asset can be split into two distinct components, namely, a market-based component and a specific component. This specific component vanishes within well-diversified portfolios so that their global risk summarizes to the broad market influence.

Framework and Risk Typology

The CAPM provides a foundation for the theory of market equilibrium, which relies on both the utility theory (*see* **Utility Theory: Historical Perspectives**) and the portfolio selection theory (*see* **Markowitz, Harry**). The main focus consists of analyzing and

understanding the behaviors and transactions of market participants on the financial market. Under this setting, market participants are assumed to act simultaneously so that they can invest their money in only two asset classes, namely, risky assets, which are contingent claims, and nonrisky assets such as the risk-free asset. The confrontation between the supply and demand of financial assets in the market allows, therefore, for establishing an equilibrium price (for each traded asset) once the supply of financial assets satisfies the demand of financial assets. The uncertainty surrounding contingent claims is so that the general equilibrium theory explains risky asset prices by the equality between the supply and demand of financial assets. Under this setting, Sharpe [49, 50] assumes that the returns of contingent claims depend on each other only due to a unique exogenous market factor called the *market portfolio*. The other potential impacting factors are assumed to be random.

Hence, the CAPM results immediately from Markowitz [37, 38] setting since it represents an equilibrium model of financial asset prices (*see* **Markowitz, Harry**). Basically, market participants hold portfolios, which are composed of the risk-free asset and the market portfolio (representing the set of all traded risky assets). The market portfolio is moreover a mean–variance efficient portfolio, which is optimally diversified and satisfies equilibrium conditions (*see* **Efficient Markets Theory: Historical Perspectives; Efficient Market Hypothesis; Risk–Return Analysis**). Consequently, holding a risky asset such as a stock is equivalent to holding a combination of the risk-free asset and the market portfolio, the market portfolio being the unique market factor.

The Capital Asset Pricing Model

Specifically, Sharpe [49, 50] describes the uncertainty underlying contingent claims with a one-factor model—the CAPM. The CAPM illustrates the establishment of financial asset prices under uncertainty and under market equilibrium. Such equilibrium is partial and takes place under a set of restrictive assumptions.

Assumptions

1. Markets are *perfect* and without frictions: no tax, no transaction costs (*see* **Transaction Costs**),

- and no possibility of manipulating asset prices in the market (i.e., perfect market competition).
2. Information is instantaneously and perfectly available in the market so that investors simultaneously access the same information set without any cost.
 3. Market participants invest over *one time period* so that we consider a one-period model setting.
 4. Financial assets are infinitely divisible and liquid.
 5. Lending and borrowing processes apply the risk-free rate (same rate of interest), and there is no short sale constraint.
 6. Asset returns are *normally distributed* so that expected returns and corresponding standard deviations are sufficient to describe the assets' behaviors (i.e., their probability distributions). The Gaussian distribution assumption is equivalent to a quadratic utility setting.
 7. Investors are *risk averse and rational*. Moreover, they seek to maximize the expected utility of their future wealth/of the future value of their investment/portfolio (*see Expected Utility Maximization: Duality Methods; Expected Utility Maximization*; and the two-fund separation theorem of Tobin [52]).
 8. Investors build *homogeneous expectations* about the future variation of interest rates. All the investors build the same forecasts about the expected returns and the variance-covariance matrix of stock returns. Therefore, there is a unique set of optimal portfolios. Basically, investors share the same opportunity sets, which means they consider the same sets of accessible and "interesting" portfolios.
 9. The combination of two distinct and independent risk factors drives the evolution of any risky return over time, namely, the broad financial market and the fundamental/specific features of the asset under consideration. Basically, the risk level embedded in asset returns results from the trade-off between a market risk factor and an idiosyncratic risk factor.

The market risk factor is also called *systematic risk factor* and *nondiversifiable risk factor*. It represents a risk factor, which is common to any traded financial asset. Specifically, the market risk factor represents the global evolution of the financial market and the economy (i.e., trend of the broad market, business cycle), and impacts any risky asset. Indeed,

it characterizes the systematic fluctuations in asset prices, which result from the broad market. In a complementary way, the specific risk factor is also called *idiosyncratic risk factor*, *unsystematic risk factor*, or *diversifiable risk factor*. It represents a component, which is peculiar to each financial asset or to each financial asset class (e.g., small or large caps). This specific component in asset prices has no link with the broad market. Moreover, the systematic risk factor is priced by the market, whereas the idiosyncratic risk factor is not priced by the market. Specifically, market participants ascribe a nonzero expected return to the market risk factor, whereas they ascribe a zero expected return to the specific risk factor. This feature results from the fact that the idiosyncratic risk can easily be mitigated within a well-diversified portfolio, namely, a portfolio with a sufficient number of heterogeneous risky assets so that their respective idiosyncratic risks cancel each other. Thus, a diversified portfolio's global risk (i.e., total variance) results only from the market risk (i.e., systematic risk).

CAPM equation

Under the previous assumptions, the CAPM establishes a linear relationship between a portfolio's expected risk premium and the expected market risk premium as follows:

$$E[R_P] = r_f + \beta_P \times (E[R_M] - r_f) \quad (1)$$

where R_M is the return of the market portfolio; R_P is the return of portfolio P (which may also correspond to a given stock i); r_f is the risk-free interest rate; β_P is the beta of portfolio P ; and $E[R_M] - r_f$ is the market price of risk. The market portfolio M is composed of all the available and traded assets in the market. The weights of market portfolio's components are proportional to their corresponding market capitalization relative to the global broad market capitalization. Therefore, the market portfolio is representative of the broad market evolution and its related systematic risk. Finally, β_P is a systematic risk measure also called *Sharpe coefficient* since it quantifies the sensitivity of portfolio P or stock i to the broad market. Basically, the portfolio's beta is written as

$$\beta_P = \frac{\text{Cov}(R_P, R_M)}{\text{Var}(R_M)} = \frac{\sigma_{PM}}{\sigma_M^2} \quad (2)$$

where $\text{Cov}(R_P, R_M) = \sigma_{PM}$ is the covariance between the portfolio's return and the market return,

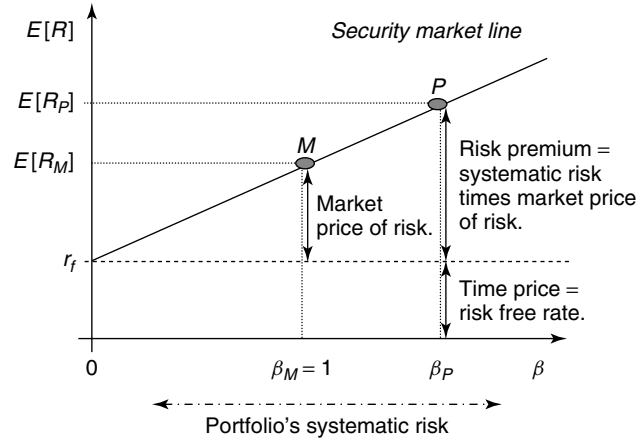


Figure 1 Security market line

and $\text{Var}(R_M) = \sigma^2_M$ is the market return's variance over the investment period. In other words, beta is the risk of covariation between the portfolio's and the market's returns normalized by the market return's variance. Therefore, beta is a relative risk measure. Under the Gaussian return assumption, the standard deviation, or equivalently the variance, is an appropriate risk metric for measuring the dispersion risk of asset returns.

Therefore, under equilibrium, the portfolio's expected return R_P equals the risk-free rate increased by a risk premium. The risk premium is a linear function of the systematic risk measure as represented by the beta and the market price of risk as represented by the expected market risk premium. Such a relationship is qualified as the security market line (SML; see Figure 1). Since idiosyncratic risk can be diversified away, only the systematic risk component in asset returns matters.^a Intuitively, diversified portfolios cannot get rid of their respective dependency to the broad market. From a portfolio management prospect, the CAPM relationship then focuses mainly on diversified portfolios, namely, portfolios or stocks with no idiosyncratic risk.

It then becomes useless to keep any idiosyncratic risk in a given portfolio since such a risk is not priced by the market. The beta parameter becomes subsequently the only means to control the portfolio's risk since the CAPM relationship (1) establishes the premium investors require to bear the portfolio's systematic risk. Indeed, the higher the dependency on the broad financial market is, the greater the risk premium

required by investors becomes. Consequently, the beta parameter allows investors to classify assets as a function of their respective systematic risk level (see Table 1).

Assets with negative beta values are usually specific commodity securities such as gold-linked assets. Moreover, risk-free securities such as cash or Treasury bills, Treasury bonds, or Treasury notes belong to the zero-beta asset class. Risk-free securities are independent from the broad market and exhibit a zero variance, or equivalently a zero standard deviation. However, the class of zero-beta securities includes also risky assets, namely, assets with a nonzero variance, which are not correlated with the market.

Table 1 Systematic risk classification

Beta level	Classification
$\beta > 1$	Offensive, cyclical asset amplifying market variations
$0 < \beta < 1$	Defensive asset absorbing market variations
$\beta = 1$	Market portfolio or asset mimicking market variations
$\beta = 0$	Asset with no market dependency
β lies between -1 and 1	Asset with low systematic risk level
$ \beta $ lies above 1	Asset with a higher risk level than the broad market's risk

Estimation and Usefulness

The CAPM theory gives a partial equilibrium relationship, which is assumed to be stable over time. However, how can we estimate such a linear relationship in practice and how do we estimate a portfolio's beta? How useful is this theory to market participants and investors?

Empirical Estimation

As a first point, under the Gaussian return assumption, beta coefficients can be computed while considering the covariance and variance of asset returns over the one-period investment horizon (see equation (2)). However, this way of computing beta coefficients does not work in a non-Gaussian world. Moreover, beta estimates depend on the selected market index, the studied time window, and the frequency of historical data [8].

As a second point, empirical estimations of the CAPM consider historical data and select a stock market index as a proxy for the CAPM market portfolio. Basically, the CAPM is tested while running two possible types of regressions based on observed asset returns (i.e., past historical data). Therefore, stocks' and portfolios' betas are estimated by regressing past asset returns on past market portfolio returns. We therefore focus on the potential existence of a linear relationship between stock/asset returns and market returns. The first possible estimation method corresponds to the market model regression as follows:

$$R_{it} - r_f = \alpha_i + \beta_i \times (R_{Mt} - r_f) + \varepsilon_{it} \quad (3)$$

where R_{it} is the return of asset i at time t ; R_{Mt} is the market portfolio's return at time t , namely, the systematic risk factor as represented by the chosen market benchmark, which is the unique explanatory factor; r_f is the short-term risk-free rate; ε_{it} is a Gaussian white noise with a zero expectation and a constant variance $\sigma_{\varepsilon_i}^2$; α_i is a constant trend coefficient; and the slope coefficient β_i is simply the beta of asset i . The trend coefficient α_i measures the distance of the asset's average return to the security market line, namely, the propensity of asset i to overperform (i.e., $\alpha_i > 0$) or to underperform (i.e., $\alpha_i < 0$) the broad market. In other words, α_i is the difference between the expected return forecast provided by the security market line and the average return observed on past history. The error term ε_{it} represents the diversifiable/idiosyncratic risk factor

describing the return of asset i . Therefore, R_{Mt} and ε_{it} are assumed to be independent, whereas (ε_{it}) are supposed to be mutually independent. Regression equation (3) is simply the ex-post form of the CAPM relationship, namely, the application of CAPM to past observed data [27].

The second method for estimating CAPM betas is the characteristic line so that we consider the following regression:

$$R_{it} = a_i + b_i \times R_{Mt} + \varepsilon_{it} \quad (4)$$

where a_i and b_i are constant trend and slope regression coefficients, respectively [51]. Moreover, such coefficients have to satisfy the following constraints:

$$\alpha_i = a_i - (1 - b_i) \times r_f \quad (5)$$

$$\beta_i = b_i \quad (6)$$

Regression equations (3) and (4) are only valid under the strong assumptions that α_i and β_i coefficients are stationary over time (e.g., time stability), and that each regression equation is a valid model over each one-period investment horizon.

In practice, the market model (3) is estimated over a two-year window of weekly data, whereas the characteristic line (4) is estimated over a five-year window of monthly data. Basically, the market model and the characteristic line use, as a market proxy, well-chosen stock market indexes such as NYSE index and S&P500 index, respectively, which are adapted to the frequency of the historical data under consideration.

Practical Use

A sound estimation process is very important insofar as the CAPM relationship intends to satisfy investors' needs. From this viewpoint, the main goal of CAPM estimation is first to use past-history beta estimates to forecast future betas. Specifically, the main objective consists of extracting information from past history to predict future betas. However, extrapolating past beta estimates to build future beta values may generate estimation errors resulting from outliers due to firm-specific events or structural changes either in the broad market or in the firm [10].

Second, the CAPM is a benchmark tool helping investors' decision. Specifically, the SML is used to identify overvalued (i.e., above SML) and undervalued (i.e., below SML) stocks under a fundamental

analysis setting. Indeed, investors compare observed stock returns with CAPM required returns and then assess the performance of the securities under consideration. Therefore, the CAPM relationship provides investors with a tool for investment decisions and trading strategies since it provides buy and sell signals, and drives asset allocation across different asset classes.

Third, the CAPM allows for building classical performance measures such as Sharpe ratio (*see Sharpe Ratio*), Treynor index, or Jensen's alpha (*see Style Analysis; Performance Measures*). Finally, the CAPM theory can be transposed to firm valuation insofar as the equilibrium value of the firm is the discounted value of its future expected cash flows. The discounting factor is just mitigated by one identified risk factor affecting equity [20, 29, 30, 47]. According to the theorem proposed by Modigliani and Miller [40–42] (*see Modigliani–Miller Theorem*), the cost of equity capital for an indebted firm corresponds to the risk-free rate increased by an operating risk premium (independent from the firm's debt) times a leverage-specific factor. The firm's risk is therefore measured by the beta of its equity (i.e., equity's systematic risk), which also depends on the beta of the firm's assets and on the firm's leverage. Indeed, the leverage increases the beta of equity in a perfect market and therefore increases the firm's risk, which represents the probability of facing a default situation. However, an optimal capital structure may result from market imperfections such as taxes, agency costs, bankruptcy costs, and information asymmetry among others. For example, there exists a trade-off between the costs incurred by a financial distress (i.e., default) and the potential tax benefits inferred from leverage (i.e., debt). Consequently, applying the CAPM to establish the cost of capital allows for budget planning and capital budgeting insofar as choosing an intelligent debt level allows for maximizing the firm value. Namely, there exists an optimal capital structure.

Limitations and Model Extensions

However, CAPM is only valid under its strong semi-anal assumptions and exhibits a range of shortcomings as reported by Banz [6], for example. However, in practice and in the real financial world, many of these assumptions are violated. As a result, the CAPM suffers from various estimation problems that impact its

efficiency. Indeed, Campbell *et al.* [14] show the poor performance of CAPM over the 1990s investment period in the United States. Such a result does have several possible explanations among which missing explanatory factors, heteroscedasticity in returns or autocorrelation patterns, time-varying or nonstationary CAPM regression estimates. For example, heteroscedastic return features imply that the static estimation of the CAPM is flawed under the classic setting (e.g., ordinary least squares linear regression). One has, therefore, to use appropriate techniques while running the CAPM regression under heteroscedasticity or non-Gaussian stock returns (*see* [7], for example, and *see also Generalized Method of Moments (GMM); GARCH Models*).

General Violations

Basic CAPM assumptions are not satisfied in the market and engender a set of general violations. First, lending and borrowing rates of interest are different in practice. Generally speaking, it is more expensive to borrow money than to lend money in terms of interest rate level. Second, the risk-free rate is not constant over time but one can focus on its arithmetic mean over the one-period investment horizon. Moreover, the choice of the risk-free rate employed in the CAPM has to be balanced with the unit-holding period under consideration. Third, transactions costs are often observed on financial markets and constitute part of the brokers' and dealers' commissions. Fourth, the market benchmark as well as stock returns are often nonnormally distributed and skewed [44]. Indeed, asset returns are skewed, leptokurtic [55], and they exhibit volatility clusters (i.e., time-varying volatility) and long memory patterns [2, 45]. Moreover, the market portfolio is assumed to be composed of all the risky assets available on the financial market so as to represent the portfolio of all the traded securities. Therefore, the broad market proxy or market benchmark should encompass stocks, bonds, human capital, real estate assets, and foreign assets (*see the critique of Roll* [46]). Fifth, financial assets are not infinitely divisible so that only fixed amounts or proportions of shares, stocks, and other traded financial instruments can be bought or sold.

Finally, the static representation of CAPM is at odds with the dynamic investment decision process. This limitation gives birth to multiperiodic extensions of CAPM. Extensions are usually called *intertemporal capital asset pricing models (ICAPMs)*,

and extend the CAPM framework to several unit-holding periods (see [11, 39]).

Trading, Information, and Preferences

Insider trading theory assumes that some market participants hold some private information. Specifically, information asymmetry prevails so that part of existing information is not available to all investors. Under such setting, Easley and O'Hara [22] and Wang [56] show that the trade-off between public and private information affects any firm's cost of capital as well as the related return required by investors. Namely, the existence of private information increases the return required by uninformed investors. Under information asymmetry, market participants exchange indeed information through observed trading prices [18]. Moreover, heterogeneity prevails across investors' preferences. Namely, they exhibit different levels of risk tolerance, which drives their respective investments and behaviors in the financial market. Finally, homogeneous expectations are inconsistent with the symmetry in the motives of transaction underlying any given trade. For a transaction to take place, the buy side has to meet the sell side. Indeed, Anderson *et al.* [4] show that heterogeneous beliefs play a nonnegligible role in asset pricing.

Nonsynchronous Trading

Often, the market factor of risk and stocks are not traded at the same time on the financial market, specifically at the daily frequency level. This stylized fact engenders the so-called nonsynchronous trading problem. When the market portfolio is composed of highly liquid stocks, the nonsynchronism problem is reduced within the portfolio as compared to an individual stock. However, for less liquid stocks or less liquid financial markets, the previous stylized fact becomes an issue under the CAPM estimation setting. To bypass this problem, the asset pricing theory introduces one-lag systematic risk factor(s) as additional explanatory factor(s) to describe asset returns [13, 21, 48].

Missing Factors

The poor explanatory power of the CAPM setting [14] comes from the lack of information describing stock returns in the market among others. The broad market's uncertainty is described by a unique risk

factor: the market portfolio. Indeed, considering the market portfolio as the unique source of systematic risk, or equivalently as the unique systematic risk information source is insufficient. To bypass this shortcoming, a wide academic literature proposes to add complementary factors to the CAPM in order to better forecast stock returns (*see Arbitrage Pricing Theory; Predictability of Asset Prices; Factor Models*). Those missing factors are often qualified as asset pricing anomalies [5, 24, 26, 31]. Namely, the absence of key explanatory factors generates misestimations in computed beta values.

For example, Fama and French [25] propose to consider two additional factors such as the issuing firm's size and book-to-market characteristics. Further, Carhart [16] proposes to add a fourth complementary factor called *momentum*. The stock momentum represents the significance of recent past stock returns on the current observed stock returns. Indeed, investors' sentiment and preferences may explain expected returns to some extent. In this prospect, momentum is important since investors make the difference between poor and high performing stocks over a recent past history. More recently, Li [34] proposed two additional factors to the four previous ones, namely, the earnings-to-price ratio and the share turnover as a liquidity indicator. Indeed, Acharya and Pedersen [1], Brennan and Subrahmanyam [12], Chordia *et al.* [19], and Keene and Peterson [32] underlined the importance of liquidity as an explanatory factor in asset pricing. Basically, the trading activity impacts asset prices since the degree of transactions' fluidity drives the continuity of observed asset prices. In other words, traded volumes impact market prices, and the impact's magnitude depends on the nature of market participants [17].

Time-varying Betas

Some authors like Tofallis [53] questioned the soundness of CAPM while assessing and forecasting stock returns' performance. Indeed, the CAPM relationship is assumed to remain stable over time insofar as it relies on constant beta estimates over each unit-holding period (i.e., reference time window). Such a process assumes implicitly that beta estimates remain stable in the near future so that ex-post beta estimates are good future risk indicators. However, time instability is a key feature of beta estimates. For example,

Gençay *et al.* [28] and Koutmos and Knif [33] support time-varying betas in CAPM estimation.

Moreover, CAPM-type asset pricing models often suffer from error-in-variables problems coupled with time-varying parameters features [15]. To solve such problems, authors like Amman and Verhoeven [3], Ellis [23], and Wang [57] among others advocate using conditional versions of the CAPM. Moreover, Amman and Verhofen [3] and Wang [57] show the efficiency of conditional asset pricing models and exhibit the superior performance of the conditional CAPM setting as compared to other asset pricing models.

End Notes

^aSpecifically, the systematic risk represents that part of returns' global risk/variance, which is common to all traded assets, or equivalently, which results from the broad market's influence.

References

- [1] Acharya, V.V. & Pedersen, L.H. (2005). Asset pricing with liquidity risk, *Journal of Financial Economics* **77**(2), 375–410.
- [2] Adrian, T. & Rosenberg, J. (2008). *Stock Returns and Volatility: Pricing the Short-run and Long-run Components of Market Risk*, Staff Report No 254, Federal Reserve Bank of New York.
- [3] Amman, M. & Verhofen, M. (2008). Testing conditional asset pricing models using a Markov chain Monte Carlo approach, *European Financial Management* **14**(3), 391–418.
- [4] Anderson, E.W., Ghysels, E. & Juergens, J.L. (2005). Do heterogeneous beliefs matter for asset pricing? *Review of Financial Studies* **18**(3), 875–924.
- [5] Avramov, D. & Chordia, T. (2006). Asset pricing models and financial market anomalies, *Review of Financial Studies* **19**(3), 1001–1040.
- [6] Banz, R. (1981). The relationship between return and market value of common stocks, *Journal of Financial Economics* **9**(1), 3–18.
- [7] Barone Adesi, G., Gagliardini, P. & Urga, G. (2004). Testing asset pricing models with coskewness, *Journal of Business and Economic Statistics* **22**(4), 474–495.
- [8] Berk, J. & DeMarzo, P. (2007). *Corporate Finance*, Pearson International Education, USA.
- [9] Black, F. (1972). Capital market equilibrium with restricted borrowing, *Journal of Business* **45**(3), 444–455.
- [10] Bossaerts, P. & Hillion, P. (1999). Implementing statistical criterion to select return forecasting models: what do we learn? *Review of Financial Studies* **12**(2), 405–428.
- [11] Breeden, D. (1979). An intertemporal capital asset pricing model with stochastic consumption and investment opportunities, *Journal of Financial Economics* **7**(3), 265–296.
- [12] Brennan, M.J. & Subrahmanyam, A. (1996). Market microstructure and asset pricing: on the compensation for illiquidity in stock returns, *Journal of Financial Economics* **41**(3), 441–464.
- [13] Busse, J.A. (1999). Volatility timing in mutual funds: evidence from daily returns, *Review of Financial Studies* **12**(5), 1009–1041.
- [14] Campbell, J.Y., Lettau, M., Malkiel, B.G. & Xu, Y. (2001). Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk, *Journal of Finance* **56**(1), 1–43.
- [15] Capiello, L. & Fearnley, T.A. (2000). *International CAPM with Regime Switching GARCH Parameters*. Graduate Institute of International Studies, University of Geneva. Research Paper No 17.
- [16] Carhart, M.M. (1997). On persistence in mutual fund performance, *Journal of Finance* **52**(1), 57–82.
- [17] Carpenter, A. & Wang, J. (2007). Herding and the information content of trades in the Australian dollar market, *Pacific-Basin Finance Journal* **15**(2), 173–194.
- [18] Chan, H., Faff, R., Ho, Y.K. & Ramsay, A. (2006). Asymmetric market reactions of growth and value firms with management earnings forecasts, *International Review of Finance* **6**(1–2), 79–97.
- [19] Chordia, T., Roll, R. & Subrahmanyam, A. (2001). Trading activity and expected stock returns, *Journal of Financial Economics* **59**(1), 3–32.
- [20] Cohen, R.D. (2008). *Incorporating default risk into Hamada's equation for application to capital structure*, *Wilmott Magazine* March, 62–68.
- [21] Dimson, E. (1979). Risk measurement when shares are subject to infrequent trading, *Journal of Financial Economics* **7**(2), 197–226.
- [22] Easley, D. & O'Hara, M. (2004). Information and the cost of capital, *Journal of Finance* **59**(4), 1553–1583.
- [23] Ellis, D. (1996). A test of the conditional CAPM with simultaneous estimation of the first and second conditional moments, *Financial Review* **31**(3), 475–499.
- [24] Faff, R. (2001). An Examination of the Fama and French three-factor model using commercially available factors, *Australian Journal of Management* **26**(1), 1–17.
- [25] Fama, E.F. & French, K.R. (1993). Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* **33**(1), 3–56.
- [26] Fama, E.F. & French, K.R. (1996). Multi-factor explanations of asset pricing anomalies, *Journal of Finance* **51**(1), 55–84.
- [27] Friend, I. & Westerfield, R. (1980). Co-skewness and capital asset pricing, *Journal of Finance* **35**(4), 897–913.
- [28] Gençay, R., Selçuk, F. & Whitcher, B. (2003). Systematic risk and timescales, *Quantitative Finance* **3**(1), 108–116.

- [29] Hamada, R. (1969). Portfolio analysis market equilibrium and corporation finance, *Journal of Finance* **24**(1), 13–31.
- [30] Hamada, R. (1972). The effect of the firm's capital structure on the systematic risk of common stocks, *Journal of Finance* **27**(2), 435–451.
- [31] Hu, O. (2007). Applicability of the Fama-French three-factor model in forecasting portfolio returns, *Journal of Financial Research* **30**(1), 111–127.
- [32] Keene, M.A. & Peterson, D.R. (2007). The importance of liquidity as a factor in asset pricing, *Journal of Financial Research* **30**(1), 91–109.
- [33] Koutmos, G. & Knif, J. (2002). Estimating systematic risk using time-varying distributions, *European Financial Management* **8**(1), 59–73.
- [34] Li, X. (2001). *Performance Evaluation of Recommended Portfolios of Individual Financial Analysts*. Working Paper, Owen Graduate School of Management, Vanderbilt University.
- [35] Lintner, J. (1965). The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets, *Review of Economics and Statistics* **47**(1), 13–37.
- [36] Lintner, J. (1969). The aggregation of investor's diverse judgments and preferences in purely competitive security markets, *Journal of Financial and Quantitative Analysis* **4**(4), 347–400.
- [37] Markowitz, H.W. (1952). Portfolio selection, *Journal of Finance* **7**(1), 77–91.
- [38] Markowitz, H.W. (1959). *Portfolio Selection. Efficient Diversification of Investment*, John Wiley & Sons, New York.
- [39] Merton, R.C. (1973). An intertemporal capital asset pricing model, *Econometrica* **41**(5), 867–887.
- [40] Modigliani, F. & Miller, M.H. (1958). The cost of capital, corporation finance and the theory of investment, *American Economic Review* **48**(3), 261–297.
- [41] Modigliani, F. & Miller, M.H. (1963). Corporate income taxes and the cost of capital: a correction, *American Economic Review* **53**(3), 433–443.
- [42] Modigliani, F. & Miller, M.H. (1966). Some estimates of the cost of capital to the utility industry 1954–7, *American Economic Review* **56**(3), 333–391.
- [43] Mossin, J. (1966). Equilibrium in a capital asset market, *Econometrica* **34**(4), 768–783.
- [44] Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: a new approach, *Econometrica* **59**(2), 347–370.
- [45] Oh, G., Kim, S. & Eom, C. (2008). Long-term memory and volatility clustering in high-frequency price changes, *Physica A: Statistical Mechanics and Its Applications* **387**(5–6), 1247–1254.
- [46] Roll, R. (1977). A critique of the asset pricing theory's tests: Part one: on past and potential testability of the theory, *Journal of Financial Economics* **4**(1), 129–176.
- [47] Rubinstein, M. (1973). A mean variance synthesis of corporate financial theory, *Journal of Finance* **38**(1), 167–181.
- [48] Scholes, M. & Williams, J. (1977). Estimating betas from non synchronous data, *Journal of Financial Economics* **5**(3), 309–327.
- [49] Sharpe, W.F. (1963). A simplified model of portfolio analysis, *Management Science* **9**(2), 227–293.
- [50] Sharpe, W.F. (1964). Capital asset prices: a theory of market equilibrium under risk, *Journal of Finance* **19**(3), 425–442.
- [51] Smith, K.V. & Tito, D.A. (1969). Risk-return measures of ex post portfolio performance, *Journal of Financial and Quantitative Analysis* **4**(4), 449–471.
- [52] Tobin, J. (1958). Liquidity preferences as behavior towards risk, *Review of Economic Studies* **25**(1), 65–86.
- [53] Tofallis, C. (2008). Investment volatility: a critique of standard beta estimation and a simple way forward, *European Journal of Operational Research* **187**(3), 1358–1367.
- [54] Treynor, J. (1961). Toward a theory of the market value of risky assets, in *Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics*, Korajczyk, Robert A., ed., Risk Books, London, pp. 15–22. Unpublished Manuscript. Recently published in 1999 as the Chapter 2 of editor).
- [55] Verhoeven, P. & McAleer, M. (2004). Fat tails and asymmetry in financial volatility models, *Mathematics and Computers in Simulation* **64**(3–4), 351–361.
- [56] Wang, J. (1993). A model of intertemporal asset prices under asymmetric information, *Review of Economic Studies* **60**(2), 249–282.
- [57] Wang, K.Q. (2003). Asset pricing with conditioning information: a new test, *Journal of Finance* **58**(1), 161–196.

Related Articles

Arbitrage Pricing Theory; Efficient Markets Theory; Historical Perspectives; Markowitz, Harry; Modigliani, Franco; Sharpe, William F.

HAYETTE GATFAOUI

Arbitrage Pricing Theory

The arbitrage pricing theory (APT) was introduced by Ross [10] as an alternative to the capital asset pricing model (CAPM). The model derives a multibeta representation of expected returns relative to a set of K reference variables under assumptions that may be described roughly as follows:

1. There exists no mean–variance arbitrage.
2. The asset returns follow a K -factor model.
3. The reference variables and the factors are non-trivially correlated.^a

The first assumption implies that there are no portfolios with arbitrarily large expected returns and unit variance. The second one assumes that the returns are a function of K factors common to all assets, and noise term specific to each asset. The third one identifies the sets of reference variables for which the model works.

The model predictions may have approximation errors. However, these errors are small for each portfolio that its weight on each asset is small (a well-diversified portfolio).

Early versions of the model unnecessarily assumed that the factors are equal to the reference variables. The extension of the model to arbitrary sets of reference variables comes at the cost of increasing the bound on the approximation errors by a multiplicative factor. However, when focusing on pricing of only well-diversified portfolios, this seems to be unimportant because each of the approximation error is small and a multiplicative factor does not change much the size of the error.

Factor Representation

Consider a finite sequence of random variables $\{Z_i; i = 1, \dots, N\}$ with finite variances that will be held fixed throughout the article. It is regarded as representing the excess^b returns of a given set of assets (henceforth “assets $i = 1, \dots, N$ ”). Without any further assumptions

$$Z_i = b_{i,0} + \sum_{k=1}^K b_{i,k} f_k + e_i; \quad i = 1, \dots, N$$

where $f_1, \dots, f_K$ are the first K factors in the principal component analysis (PCA) of the sequence

$\{Z_i; i = 1, \dots, N\}$. The $b_{i,k}$ are the factor loadings and the e_i are the residuals from projecting the Z_i on the factors.

The $K + 1$ largest eigenvalue of the covariance matrix of the Z_i , denoted by $\Lambda^2(K)$, is interpreted as a measure of the extent to which our sequence of assets has a K -factor representation. The PCA selects the f_k so that $\Lambda^2(K)$ is minimized. In addition, $\Lambda^2(K)$ is also the largest eigenvalue of the covariance matrix of the e_i .

Diversified Portfolios

Let $w \in R^N$ be a portfolio in assets $i = 1, \dots, N$. Its excess return is

$$Z_w = \sum_{i=1}^N w_i Z_i.$$

Its representation as a linear function of the factors is

$$Z_w = b_{w,0} + \sum_{k=1}^K b_{w,k} f_k + e_w$$

where $b_{w,k} = \sum_{i=1}^N w_i b_{i,k}$ are the factor loadings and $e_w = \sum_{i=1}^N w_i e_i$ is the residual which satisfies

$$\text{Var}[e_w] < \Lambda^2(K) \sum_{i=1}^N w_i^2$$

A portfolio $w = (w_1, \dots)$ is called an (approximate) well-diversified portfolio if

$$\sum_{i=1}^N w_i^2 \approx 0 \quad (1)$$

Intuitively, a well-diversified portfolio is one with a large number of assets that has a small weight in many of them, and, in addition, there is no single asset for which the weight is not small.

The variance of the residual of a well-diversified portfolio is small and thus its excess return is approximately a linear function of the factors; that is,

$$Z_w \approx b_{w,0} + \sum_{k=1}^K b_{w,k} f_k \quad (2)$$

Although $\sum_{i=1}^N w_i^2 \approx 0$, Z_w may not be small. For example, let $w_i = 1/N$, then we have $\sum_{i=1}^N w_i^2 = 1/N$, and $b_{w,k} = (1/N) \sum_{i=1}^N b_{i,k}$

A further discussion on well-diversified portfolios can be found in [4].

Multibeta Representation

Throughout the article we consider a fixed set of K reference variables $\{g_1, \dots, g_K\}$ with respect to which we derive an approximate *multibeta representation* defined as

$$E[Z_i] = \sum_{k=1}^K B_{i,k} \lambda_k + \rho_i \quad (3)$$

where

$$B_{i,k} = \text{Cov}(Z_i, g_k) \quad (4)$$

This means that

$$E[Z_i] \approx \sum_{k=1}^K B_{i,k} \lambda_k \quad (5)$$

where ρ_i is the approximation error in pricing asset i . The sum of the squares of these approximation errors, that is,

$$\sum_{i=1}^N \rho_i^2 = \Omega^2 \quad (6)$$

determines the quality of the approximation.

The APT Bound

Huberman [3] showed that Ω is finite for an infinite sequence of excess returns but did not derive a bound. Such bounds were derived by Chamberlain & Rothschild [1], in the case where the reference variables are the factors and by Reisman [7], in the general case. Reisman showed that

$$\Omega \leq \Lambda S \Phi \Psi V \quad (7)$$

where $\Lambda^2(K)$ is the $K + 1$ largest eigenvalue of the covariance matrix of the Z_i ; S is the lowest upper bound on expected excess return among portfolios with unit variance; $\Phi^2 = 1 - R^2$ of the regression of the tangency portfolio on the reference variables; Ψ is an increasing function of the largest eigenvalue of $(G'G)^{-1}$, where $G = \text{Corr}(f_n, g_m)_{n,m=1,\dots,K}$ is the cross-correlation matrix of the factors and the reference variables; and V^2 is a bound on the variances of the Z_i . See [5, 8] for further details.

What is important about the bound is that neither Φ nor Ψ depends on the number of assets, N . This means that the size of the bound depends on the number of assets N , only through $\Lambda(K)$, S , and V , which may be bounded as this number increases to infinity.

The Pricing Errors

The pricing error of any portfolio w ,

$$\sum_{k=1}^K w_k \rho_k = \rho_w \quad (8)$$

satisfies

$$|\rho_w|^2 \leq \Omega \sum_{i=1}^N w_i^2 \quad (9)$$

Provided Ω is not large and N is large, the pricing error on each well-diversified portfolio is small. For a single asset i , we only get that most of the ρ_i are small. However, for a few of the assets the ρ_i may not be small.

Example

Assume that each Z_i is given by

$$Z_i = a_i + b_i f + e_i$$

where the e_i are mutually uncorrelated and have zero mean, and f has a zero mean and unit variance and is uncorrelated with all the e_i .

The APT implies that every random variable g for which $\text{cov}(g, f)$ is not zero can serve as a reference variable. Thus there exists a constant λ so that

$$E[Z_i] = \text{cov}(Z_i, g) \lambda + \rho_i \text{ for each } i$$

In addition, for each well-diversified portfolio w , we have

$$E[Z_w] \approx \text{cov}(Z_w, g) \lambda$$

In this example, $\Psi = 1/\text{corr}(f, g)^2$; $\Lambda(1)$ and S and Φ may take arbitrary values.

Empirical Studies

Empirical studies attempted to find the sets of reference variables for which the hypothesis that

$$E[Z_i] = \sum_{k=1}^K B_{i,k} \lambda_k$$

cannot be rejected. Roll and Ross [9] identified sets of macroeconomic variables that are believed to be responsible for stock price movements and tested whether they explain expected returns in the major US markets. Trzcinka [13] applied PCA to identify the factors. He showed that a small number of factors

may explain most of the variation of the market. Then he tested the multibeta representation with these factors as reference variables.

Equilibrium APT

The CAPM implies that the market portfolio is mean–variance efficient. If the market portfolio is a well-diversified one, then it is spanned by the factors. In that case, we get that if the reference variables are the factors, then Φ is small, which implies that the approximation error for each asset in the sequence is small. Connor [2] and Wei [14] derived a related result which is called equilibrium APT.

Arbitrage and APT

S measures the extent to which arbitrage in the mean–variance sense exists. It is equal to the maximal expected excess return per unit variance of portfolios in the Z_i . A large S can be interpreted as some form of no arbitrage. However it is not an arbitrage in the standard sense as there are examples in which S is finite and arbitrage exists. See Reisman [6].

Testability

It was pointed out by Shanken [11, 12] that an inequality of the type given in equation (7) is a tautology. That is, it is a mathematical statement and thus cannot be rejected.

Assume that we performed statistical tests that imply that the probability that the bound in equation (7) holds, is small. Then the only explanation can be that it was a bad sample. Since equation (7) is a tautology, there is no other explanation.

Nevertheless, this does not imply that the bound is not useful. The bound translates prior beliefs on the sizes of Λ , S , and Ψ , into a prior belief on a bound on the size of the approximation error of each well-diversified portfolio.

The relationship between the sizes of Λ , S , and Ψ , and the model assumptions is illustrated in the next section.

The APT Assumptions

The model is derived under assumptions on the extent to which there exists

1. a factor structure with K factors;
2. no mean–variance arbitrage;
3. nontrivial correlation between our set of reference variables and the first K factors in the PCA.

The parameters Λ , S , and Ψ are measures of the extent to which each of the above assumptions holds. The larger it is, the larger is the extent to which the related assumption does not hold.

What this says is that the model translates our beliefs on the extent to which the model assumptions hold to a belief on a bound on the size of the approximation errors in pricing well-diversified portfolios.

Summary

The APT implies that each (approximate) well-diversified portfolio is (approximately) priced by a set of K reference variables.

What distinguishes this model from the K -factor CAPM is the set of reference variables that is implied by each of the models.

In the CAPM, the market portfolio is mean–variance efficient and its return must be equal to a linear function of the set of reference variables.

In contrast, in the APT, the reference variables are any set that is nontrivially correlated with the common factors of the returns and it may not span the mean–variance frontier.

End Notes

^aThe cross-correlation matrix is nonsingular.

^bThe excess return is the return minus the risk-free rate.

References

- [1] Chamberlain, G. & Rothschild, M. (1983). Arbitrage, factor structure, and mean variance analysis on large asset markets, *Econometrica* **51**, 1281–1304.
- [2] Connor, G. (1984). A unified beta pricing theory, *Journal of Economic Theory* **34**, 13–31.
- [3] Huberman, G. (1982). A simple approach to arbitrage pricing, *Journal of Economic Theory* **28**, 183–191.
- [4] Ingersoll Jr J.E. (1984). Some results in the theory of arbitrage pricing, *Journal of Finance* **39**, 1021–1039.
- [5] Nawalkha, S.K. (1997). A multibeta representation theorem for linear asset pricing theories, *Journal of Financial Economics* **46**, 357–381.
- [6] Reisman, H. (1988). A general approach to the Arbitrage Pricing Theory (APT), *Econometrica* **56**, 473–476.

- [7] Reisman, H. (1992). Reference variables, factor structure, and the approximate multibeta representation, *Journal of Finance* **47**, 1303–1314.
- [8] Reisman, H. (2002). Some comments on the APT, *Quantitative Finance* **2**, 378–386.
- [9] Roll, R. & Ross, S.A. (1980). An empirical investigation of the arbitrage pricing theory, *Journal of Finance* **35**, 1073–1103.
- [10] Ross, S.A. (1976). The arbitrage theory of capital asset pricing, *Journal of Economic Theory* **13**, 341–360.
- [11] Shanken, J. (1982). The arbitrage pricing theory: is it testable? *Journal of Finance* **37**, 1129–1140.
- [12] Shanken, J. (1992). The current state of the arbitrage pricing theory, *Journal of Finance* **47**, 1569–1574.
- [13] Trzcinka, C. (1986). On the number of factors in the arbitrage pricing model, *Journal of Finance* **41**, 347–368.
- [14] Wei, K. & John, C. (1988). An asset-pricing theory unifying CAPM and APT, *Journal of Finance*, **43**, 881–892.

Related Articles

Capital Asset Pricing Model; Correlation Risk; Factor Models; Risk–Return Analysis; Ross, Stephen; Sharpe, William F.

HAIM REISMAN

Efficient Market Hypothesis^a

The topic of capital market efficiency plays a central role in introductory instruction in finance. After investigating the risk–return trade-off and the selection of optimal portfolios, instructors find it natural to go on to raise the question of what information is incorporated in the estimates of risk and expected return that underlie portfolio choices. Information that is “fully reflected” in security prices (and therefore in investors’ estimates of expected return and risk) cannot be used to construct successful trading rules, which are defined as those with an abnormally high expected return for a given risk. In contrast, information that is not fully reflected in security prices can be so used. Students appear to find this material plausible and intuitive, and this is the basis of its appeal. Best of all, the idea of capital market efficiency appears not to depend on the validity of particular models, implying that students can grasp the major ideas without wading through the details of finance models.

However, those who are accustomed to relying on formal models to discipline their thinking find that capital market efficiency has the disadvantage of its advantage: the fact that market efficiency is not grounded in a particular model (unlike, e.g., portfolio theory) means that it is not so easy to determine what efficiency really means. To see this, consider the assertion of Fama [8] that capital market efficiency can only be tested in conjunction with a particular model of returns. This statement implies that there exist two independent sources of restrictions on the data that are being tested jointly: the assumed model and market efficiency. Analysts who are used to deriving all restrictions being tested from the assumed model find this puzzling: what is the additional source of information that is separate from the model?

This question was not addressed clearly in the major expositions of market efficiency offered by its proponents. One way to resolve this ambiguity is to look at the empirical tests that are interpreted as supporting or contradicting market efficiency. Most of the empirical evidence that Fama [7] interpreted as supporting market efficiency is based on a particular model: expected returns conditional on some pre-specified information set are constant. For example, return autocorrelatedness is evidence against market

efficiency only if market efficiency is identified with constancy of expected returns. On this reading, the additional restriction implied by market efficiency might consist of the assumption that investors have rational expectations. The market model explains asset prices based on investors’ subjective perceptions of their environment; the assumption of rational expectations is needed to connect these subjective perceptions with objective correlations. Admittedly, it is pure conjecture to assume that proponents intend this identification of market efficiency with rational expectations—as Berk [1] pointed out, there is no mention of rational expectations in [7, 8].

In many settings, conditional expected returns are constant over time when agents are risk neutral. If agents are risk averse, expected returns will generally differ across securities, as is clear from the capital asset pricing model (*see Capital Asset Pricing Model*), and will change over time according to the realizations of the conditioning variables even in stationary settings [14, 19]. Hence, if investors are risk averse, the assumption of rational expectations will not generally lead to returns that are fair games.

Analysts who understood that constancy of expected returns requires the assumption of risk neutrality (or some other even more extreme assumption, such as that growth rates of gross domestic product are independently and identically distributed over time) were skeptical about the empirical evidence offered in support of market efficiency. From the fact that high-risk assets generate higher average returns than low-risk assets—or from the fact that agents purchase insurance even at actuarially unfavorable prices, or from a variety of other considerations—we know that investors are risk averse. If so, there is no reason to expect that conditional expected returns will be constant.

One piece of evidence offered in the 1970s, which appeared to contradict the consensus in support of market efficiency, had to do with the volatility of security prices and returns. If conditional expected returns are constant, then the volatility of stock prices depends entirely on the volatility of dividends (under some auxiliary assumptions, such as exclusion of bubbles). This observation led LeRoy and Porter [16] and Shiller [23] to suggest that bounds on the volatility of stock prices and returns can be derived from the volatility of dividends. These authors concluded that stock prices appear to be more

volatile than can be justified by the volatility of dividends. This finding corroborated the informal opinion (that was subsequently confirmed by Cutler *et al.* [6]) that large moves in stock prices generally cannot be convincingly associated with contemporaneous news that would materially affect expected future dividends.

Connecting the volatility of stock prices with that of dividends required a number of auxiliary econometric specifications. These were supplied differently by LeRoy–Porter and Shiller. However, both sets of specifications turned out to be controversial (see [9] for a survey of the econometric side of the variance-bounds tests). Some analysts, such as Marsh and Merton [20], concluded that the appearance of excess volatility was exactly what should be expected in an efficient market, although the majority opinion was that resolving the econometric difficulties reduces but does not eliminate the excess volatility [25].

It was understood throughout that the variance bounds were implications of the assumption that expected returns are constant. As noted, this was the same model that was implicitly assumed in the market efficiency tests summarized by Fama. The interest in the variance-bounds tests derived from the fact that the results of the two sets of tests of the same model appeared to be so different. In the late 1980s, there was a growing realization that small but persistent autocorrelations in returns could explain the excess volatility of prices [24]. This connection is particularly easy to understand if we employ the Campbell–Shiller log-linearization. Defining r_{t+1} as the log stock return from t to $t + 1$, p_t as the log stock price at t , and d_t as the log dividend level, we have

$$p_t \cong k + p_{dt} + p_{rt} \quad (1)$$

where p_{dt} and p_{rt} are given by

$$p_{dt} = E_t \left(\sum_{j=1}^{\infty} \rho^j [(1 - \rho) d_{t+j}] \right) \quad (2)$$

and

$$p_{rt} = -E_t \left(\sum_{j=1}^{\infty} \rho^j r_{t+j} \right) \quad (3)$$

(see [2–4]). Here, k and ρ are parameters associated with the log-linearization. Thus p_{dt} and p_{rt} capture price variations induced by expected dividend variations and expected return variations, respectively.

The attractive feature of the log-linearization is that expectations of future dividends and expectations of future returns appear symmetrically and additively in relation (1). Without the log-linearization, dividends would appear in the numerator of the present-value relation and returns in the denominator, rendering the analysis less tractable.

As noted, the market-efficiency tests of Fama and the variance bounds are implications of the hypothesis that p_{rt} is a constant. If p_{rt} is, in fact, random and positively correlated with p_{dt} , then the assumption of constancy of expected returns will bias the implied volatility of p_t downward. Campbell and Shiller found that if averages of future returns are regressed on current stock prices, a significant proportion of the variation can be explained, contradicting the specification that expected returns are constant.

Campbell *et al.* noted that as economists came to understand the connection between return autocorrelatedness and price and return volatility, the variance-bounds results seemed less controversial:

LeRoy and Porter [16] and Shiller [23] started a heated debate in the early 1980s by arguing that stock prices are too volatile to be rational forecasts of future dividends discounted at a constant rate. This controversy has since died down, partly because it is now more clearly understood that a rejection of constant-discount-rate models is not the same as a rejection of Efficient Capital Markets, and partly because regression tests have convinced many financial economists that expected stock returns are time-varying rather than constant ([2] p. 275).

This passage, in implying that the return autocorrelation results provide an explanation for excess stock price volatility, is a bit misleading. The log-linearized present-value relation (1) is not a theoretical model with the potential to explain price volatility. Rather, it is very close to an identity (the only respect in which equation (1) imposes substantive restrictions lies in the assumption that the infinite sums converge; this rules out bubbles). The Campbell–Shiller exercise amounts to decomposing price variation into dividend variation, return variation, and a covariance term and observing that the latter two terms are not negligible quantitatively. This, although useful, is a restatement of the variance-bounds result, not an explanation of it. Explaining excess volatility would involve accounting in economic terms for the fact that expected returns have the time structure that they do. Campbell and Shiller have not done

this—nor has anyone else. LeRoy–Porter’s conclusion from the variance-bounds tests was that we do not understand why asset prices move as they do. That conclusion is no less true now than it was when the variance-bounds results were first reported.

Fama’s assertion that market efficiency is testable, but only in conjunction with a model of market returns, can be given another reading. Rather than identifying market efficiency with the proposition that investors have rational expectations—alternatively, with the decision to model investors as having rational expectations—one can associate market efficiency with the proposition that asset prices behave as one would expect if security markets were entirely frictionless. In such markets, prices respond quickly to information, implying that investors cannot use publicly available information to construct profitable trading rules because that information is reflected in security prices as soon as it becomes available. In contrast, the presence of major frictions in asset markets is held to imply that prices may respond slowly to information. In that case, the frictions prevent investors from exploiting the resulting trading opportunities.

In the foregoing argument, it is presumed that trading frictions and transactions costs are analogous to adjustment costs. In the theory of investment, it is sometimes assumed that investment in capital goods induces costs that motivate firms to change quantities—in this case, physical capital—more slowly than they would otherwise. It appears natural to assume that prices are similar. For example, real estate prices are held to respond slowly to relevant information because the costs implied by the illiquidity of real estate preclude the arbitrages that would otherwise bring about rapid price adjustment.

Recent work on the valuation of assets in the presence of market frictions raises questions as to the appropriateness of the analogy between quantity adjustment and price adjustment. It is correct that, if prices respond slowly to information, investors may be unable to construct the trades that exploit the mispricing because of frictions. This, however, does not establish that markets clear in settings where prices adjust slowly. Equilibrium models that characterize asset prices in the presence of frictions suggest that in equilibrium prices respond quickly to shocks, just as in the absence of frictions. For example, Krainer [11] and Krainer and LeRoy [13] analyzed equilibrium prices of illiquid assets such

as real estate in a model that accounts explicitly for illiquidity in terms of search and matching. In a similar setting, Krainer [12] introduced economy-wide shocks and found that, despite the illiquidity of real estate, prices adjust instantaneously to the shocks, just as in liquid markets.

A similar result was demonstrated by Lim [17]. He considered the determination of asset prices when short sales are restricted. Lintner [18] and Miller [21], among others, proposed that short sale restrictions cause securities to trade at higher prices than they would otherwise. This is held to occur because investors with negative information may be unable to trade based on their information, whereas those with positive information can buy without restriction. Empirical evidence is held to support this result [5, 10, 22]. Lim showed that this outcome will not occur if investors have rational expectations about the extent of short sales restrictions. Under rational expectations, prices in Lim’s model follow a martingale under the natural probabilities (reflecting assumed risk neutrality), just as they would in the absence of short sales restrictions.

These results were derived in settings that imposed strong restrictions, and it is not clear how general they are. However, the preliminary conclusion is that if market efficiency is defined as the absence of frictions, empirical evidence of quick adjustment of prices to information cannot necessarily be interpreted as supporting market efficiency, since that outcome would occur in the presence of frictions.

It could be objected that none of these considerations supports distinguishing between the implications of an asset pricing model and market efficiency, however defined. All testable restrictions are derived from an assumed model; so, the question is, what can be gained by identifying some of these restrictions with something called *market efficiency*? This is particularly debatable, given the ambiguity in the usage of this term now. Berk [1] suggested dropping the term “market efficiency” from financial economics, and this might be the best course.

End Notes

^a. An evaluation of the idea of capital market efficiency has been presented elsewhere [15]. In this essay, repetition of material found there has been avoided as much as possible.

References

- [1] Berk, J. (2007). *A Critique of the Efficient Capital Markets Hypothesis*. Reproduced, Haas School of Business, University of California, Berkeley.
- [2] Campbell, J.Y., Lo, A.W. & MacKinlay, A.C. (1996). *The Econometrics of Financial Markets*, Princeton University Press, Princeton, NJ, 275.
- [3] Campbell, J.Y. & Shiller, R.J. (1988). The dividend-price ratio and expectations of future dividends and discount factors, *Review of Financial Studies* **1**, 195–228.
- [4] Campbell, J.Y. & Shiller, R. (1988). Stock prices, earnings, and expected dividends, *Journal of Finance* **43**, 661–676.
- [5] Cheng, J.W., Chang, E.C. & Yu, Y. (2007). Short-sales constraints and price discovery: evidence from the Hong Kong market, *Journal of Finance* **62**(5), 2097–2121.
- [6] Cutler, D., Poterba, J. & Summers, L. (1989). What moves stock prices? *Journal of Portfolio Management* **15**, 4–12.
- [7] Fama, E.F. (1970). Efficient capital markets: a review of theory and empirical work, *Journal of Finance* **25**, 283–417.
- [8] Fama, E.F. (1991). Efficient capital markets: II, *Journal of Finance* **46**, 1575–1617.
- [9] Gilles, C. & LeRoy, S.F. (1991). Econometric aspects of the variance-bounds tests: a survey, *Review of Financial Studies* **4**, 753–791.
- [10] Jones, C. & Lamont, O. (2002). Short-sale constraints and stock returns, *Journal of Financial Economics* **66**, 207–239.
- [11] Krainer, J. (1997). *Pricing Illiquid Assets with a Matching Model*. Reproduced, University of Minnesota.
- [12] Krainer, J. (2001). A theory of liquidity in residential real estate markets, *Journal of Urban Economics* **13**, 32–53.
- [13] Krainer, J. & LeRoy, S.F. (2002). Equilibrium valuation of illiquid assets, *Economic Theory* **19**, 223–242.
- [14] LeRoy, S.F. (1973). Risk aversion and the martingale model of stock prices, *International Economic Review* **14**, 436–446.
- [15] LeRoy, S.F. (1989). Efficient capital markets and martingales, *Journal of Economic Literature* **27**, 1583–1621.
- [16] LeRoy, S.F. & Porter, R.D. (1981). The present value relation: tests based on implied variance bounds, *Econometrica* **49**, 555–574.
- [17] Lim, B. (2007). *Short-sales Constraints and Price Bubbles*. Reproduced, University of California, Santa Barbara.
- [18] Lintner, J. (1969). The aggregation of investors' diverse judgments and preferences in purely competitive security markets, *Journal of Financial and Quantitative Economics* **4**(4), 347–400.
- [19] Lucas, R.E. (1978). Asset prices in an exchange economy, *Econometrica* **46**, 1429–1445.
- [20] Marsh, T.A. & Merton, R.C. (1986). Dividend variability and variance bounds tests for the rationality of stock market prices, *American Economic Review* **76**, 483–498.
- [21] Miller, E.M. (1977). Risk, uncertainty, and divergence of opinion, *Journal of Finance* **32**(4), 1151–1168.
- [22] Ofek, E. & Richardson, M. (2003). Dotcommania: the rise and fall of internet stock prices, *Journal of Finance* **58**(3), 1113–1137.
- [23] Shiller, R.J. (1981). Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* **71**, 421–436.
- [24] Summers, L. (1986). Does the stock market rationally reflect fundamental values, *Journal of Finance* **41**, 591–600.
- [25] West, K.D. (1988). Bubbles, fads and stock price volatility: a partial evaluation, *Journal of Finance* **43**, 636–656.

Related Articles

Expectations Hypothesis; Predictability of Asset Prices; Risk Aversion; Risk Premia; Transaction Costs.

STEPHEN F. LEROY

Expectations Hypothesis

If the attractiveness of an economic hypothesis is measured by the number of papers which statistically reject it, the expectations theory of the term structure is a knockout [43].

The term *expectations hypothesis (EH)* stands for numerous statements that link yields, returns on bonds, and forward rates of different maturities and periods. The EH has been the basis of empirical and theoretical work in fixed income following the work of Macaulay [54]. These hypotheses were developed for understanding the returns and yields on long- *versus* short-term bonds and the time series movements of the term structure. The literature distinguishes between the *pure expectations hypothesis (PEH)*, which postulates that (i) expected excess returns on long-term over short-term bonds are zero, (ii) yield term premia are zero, or (iii) forward term premia are zero, from the EH, which postulates that (i) expected excess returns are constant over time, (ii) yield term premia are constant, or (iii) forward term premia are constant over time.

We review the literature related to the EH. We present the different forms of both the PEH and the less strong EH. We show that their mathematical expressions depend on the researchers' choice of model—continuous time *versus* discrete time—and their choice of frequency of compounding returns—continuous (log-return) *versus* discrete (simple return). Depending on these choices, we may or may not have equivalence among the several forms of the (pure) EH. In addition, we examine which of the statements can be derived from a no-arbitrage general equilibrium model. Lastly, we present empirical evidence against the EH mainly from the US data, and the less strong rejection of the hypotheses when using non-US data.

Notation

To formulate the different forms of the EH, we need to introduce the basic fixed income assets and concepts associated with them. Even though all the empirical research is done using discrete time models, the theoretical literature predominantly uses continuous time models mainly for tractability

and simplicity reasons. In comparing the expected returns on two bonds of different maturities, however, the returns may be compounded in any of four natural ways: continuously, to the shorter bond's maturity, to the longer bond's maturity, or to the nearest available future date. For these reasons, in the following, we introduce notation that is flexible enough to accommodate the description of discrete as well as continuous time models and all possible ways that compounding may take place.

A *zero-coupon bond* or *discount bond*—the simplest fixed income security—promises a single fixed payment at a specified date in the future known as *maturity date*. The size of this payment is called *face value* of the bond. Example of such securities is the Treasury bills, which are bonds issued by the US government with maturities up to a year.

We denote the price of a zero-coupon bond that matures $\tau \in \mathbb{R}^+$ periods from now and pays 1 unit at maturity as $P_t^{(\tau)}$. Call the yield to maturity—compounded once per period—for this zero-coupon bond as $Y_t^{(\tau)}$. Then prices and yields are connected through the following equation:

$$P_t^{(\tau)} = \frac{1}{(1 + Y_t^{(\tau)})^\tau} \quad (1)$$

It is common in the empirical finance literature to work with log or continuously compounded variables. This has the usual advantage of linearizing exponential affine equations that arise frequently in asset pricing and of defining comparable yield values independent of the remaining horizon value τ . Using lowercase letters for logs, the relationship between log yield and log price is

$$y_t^{(\tau)} = -\frac{P_t^{(\tau)}}{\tau} \quad (2)$$

The collection of all these yields for different maturities is called the *zero term structure of interest rates*. Buying this bond at time t and reselling it at time $t + s$ generate a holding period return of

$$R_{t \rightarrow t+s}^{(\tau)} = \frac{P_{t+s}^{(\tau-s)}}{P_t^{(\tau)}} = \frac{(1 + Y_t^{(\tau)})^\tau}{(1 + Y_{t+s}^{(\tau-s)})^{\tau-s}} \quad (3)$$

and a log holding period return of

$$\begin{aligned} r_{t \rightarrow t+s}^{(\tau)} &= p_{t+s}^{(\tau-s)} - p_t^{(\tau)} \\ &= s y_t^{(\tau)} - (\tau - s)(y_{t+s}^{(\tau-s)} - y_t^{(\tau)}) \end{aligned} \quad (4)$$

Clearly, the holding period s cannot exceed the time to maturity τ , $s \leq \tau$. The above equation shows that the holding period return on a zero-coupon bond is not known at time t unless the holding period coincides with the lifetime of the bond. In this case, the holding period return is the yield to maturity. Otherwise, the return is a random variable that depends on the future evolution of yields.

Even though returns are unknown, bonds can be combined to guarantee a fixed interest rate on an investment to be made in the future; the interest rate on this investment is called a *forward rate*. The forward and log forward rates guaranteed at time t for an investment made at time $t + s$ until time $t + \tau$ where $s \leq \tau$ are given as

$$1 + F_t^{(s,\tau)} = \frac{1}{\tau - s} \frac{1}{(P_t^{(\tau)}/P_t^{(s)})} = \frac{1}{\tau - s} \frac{(1 + Y_t^{(\tau)})^\tau}{(1 + Y_t^{(s)})^s} \quad (5)$$

$$f_t^{(s,\tau)} = \frac{P_t^{(s)} - P_t^{(\tau)}}{\tau - s} = y_t^{(s)} + \frac{\tau}{\tau - s} (y_t^{(\tau)} - y_t^{(s)}) \quad (6)$$

Finally, the short-term interest rate is the limit of yields to maturity as maturity approaches, $r_t = \lim_{\tau \downarrow 0} y_t^{(\tau)}$.

Bond Pricing

Bonds are usually priced with the use of the so-called *risk-neutral probability measure* $\mathbb{Q}$, which is equivalent to the *true* or *physical* or *data-generating* measure $\mathbb{P}$.^a The pricing under $\mathbb{P}$ is done with the use of a pricing kernel M , which is the result of the no-arbitrage assumption. For an asset that promises payoff $S(T)$ at time T , its price now at time t is given by the expected discounted cash flows equation:

$$S(t) = \mathbb{E}_t^{\mathbb{P}} \left[\frac{M(T)}{M(t)} S(T) \right] \quad (7)$$

It can be shown that given the shocks of the economy $dz(t)$, M takes the following form:

$$\frac{dM(t)}{M(t)} = -r(t) dt - \Lambda(t)^\top dz(t) \quad (8)$$

and the two measures $\mathbb{P}$ and $\mathbb{Q}(\Lambda)$ are connected through the Radon–Nikodym derivative

$$\frac{d\mathbb{Q}(\Lambda)}{d\mathbb{P}} = \xi_T^\Lambda = \exp \left(-\frac{1}{2} \int_0^T \Lambda(s)^\top \Lambda(s) ds - \int_0^T \Lambda(s) ds \right) \quad (9)$$

This gives rise to the following pricing equation under both measures:

$$S(t) = \mathbb{E}_t^{\mathbb{P}} \left[\frac{M(T)}{M(t)} S(T) \right] = \mathbb{E}_t^{\mathbb{Q}} \left[e^{-\int_t^T r(s) ds} S(T) \right] \quad (10)$$

Easy manipulations of the above equation can prove that under $\mathbb{Q}$, the instantaneous expected returns for all assets are equal to the risk-free rate. For this reason, the measure $\mathbb{Q}$ is also called the *risk-neutral measure*. Specializing the above equation for a zero-coupon bond that matures at time $t + \tau$ and promises a payoff of \$1 gives

$$P_t^{(\tau)} = \mathbb{E}_t^{\mathbb{P}} \left[\frac{M(T)}{M(t)} \right] = \mathbb{E}_t^{\mathbb{Q}} \left[e^{-\int_t^T r(s) ds} \right] \quad (11)$$

From the above pricing equations, we observe that the key variables that govern the bond prices dynamics are (i) the interest rate $r(t)$ and (ii) the prices of risk $\Lambda(t)$. Different assumptions about the functional form of these variables imply different data-generating processes for bond yields. A large part of the current fixed income literature is devoted to studying these different models and the goodness of fitting the raw bond yield data crosssectionally and in time series. We examine below the features of the most representative bond pricing models.

The Expectations Hypothesis

The term *expectations hypothesis* stands for numerous statements that link yields, returns on bonds, and forward rates of different maturities and periods. It is important to note that initially, starting with Hicks [49], Lutz [53], and Macaulay [54], these statements were not formally derived from any fully specified equilibrium model, but rather merely hypothesized. For this reason, the term *expectations hypothesis* is not associated with only one mathematical statement.

These hypotheses were developed for understanding the returns and yields on long- *versus* short-term bonds, and the time series movements of the term structure. Later, researchers developed theoretical models that give rise to some of the hypothesized equations associated with the EH [20, 21, 27, 57].

The literature distinguishes between the PEH, which postulates that (i) expected excess returns on long-term over short-term bonds are zero, (ii) yield term premia are zero, or that (iii) forward term premia are zero, from the EH, which postulates that (i) expected excess returns are constant over time, (ii) yield term premia are constant, or (iii) forward term premia are constant over time. In the following, all the forms of the PEH in discrete time with discrete and continuous compounding and in continuous time (continuous compounding) are presented. We will see that the PEH expressions derived in all these models are not equivalent across models as well as within each model.

Pure Expectations Hypothesis in Discrete Time

Discrete Compounding

The first form of the PEH equates the expected returns on one-period (short-term) and n -period (long-term) bonds, or equivalently, expected excess returns on long-term over short-term bonds are zero:

$$\begin{aligned} (1 + Y_t^{(1)}) &= \mathbb{E}_t^{\mathbb{P}} [1 + R_{t \rightarrow t+1}^{(n)}] \\ &= (1 + Y_t^{(n)})^n \cdot \mathbb{E}_t^{\mathbb{P}} \left[(1 + Y_{t+1}^{(n-1)})^{-n+1} \right] \end{aligned} \quad (12)$$

The second form of the PEH equates the n -period expected returns on the one-period and n -period bonds, or equivalently, yield term premia are zero^b:

$$\begin{aligned} (1 + Y_t^{(n)})^n &= \mathbb{E}_t^{\mathbb{P}} \left[(1 + Y_t^{(1)}) (1 + Y_{t+1}^{(1)}) \right. \\ &\quad \left. \cdots (1 + Y_{t+n-1}^{(1)}) \right] \end{aligned} \quad (13)$$

The third form of the PEH equates the expected future one-period spot rate with the current forward rate

of that future period, or equivalently, forward term premia are zero^c:

$$1 + F_t^{(n-1,n)} = \frac{(1 + Y_t^{(n)})^n}{(1 + Y_t^{(n-1)})^{n-1}} = \mathbb{E}_t^{\mathbb{P}} [1 + Y_{t+n-1}^{(1)}] \quad (14)$$

The last form of the PEH equates the n -period bond return with the one-period bond and $n - 1$ period bond:

$$(1 + Y_t^{(n)})^n = (1 + Y_t^{(1)}) \mathbb{E}_t^{\mathbb{P}} \left[(1 + Y_{t+1}^{(n-1)})^{n-1} \right] \quad (15)$$

Even though the above expressions describe different forms of the PEH, they are not mutually equivalent. Assuming that the above expressions are true for all t and n , it can be shown that (i) equation (13) is equivalent to equation (15), (ii) equation (14) implies equation (13) (therefore equation (15)), but the opposite is not true unless we make the additional assumption that $(1 + Y_{t+j}^{(1)})_{j=1}^{\infty}$ are uncorrelated with each other, and (iii) equations (12) and (15) are inconsistent, because the expected value of the inverse of a random variable is not in general equal to the inverse of its expected value.

To summarize, the PEH cannot hold in both its one-period form and its n -period form, and, essentially there are three different (competing) forms of the PEH in discrete time, the excess return expression (12), the yield premia expression (13), and the forward premia expression (14).

Imposing more structure in the term structure model by assuming that the interest rate is lognormal and homoscedastic, we can quantify the effect of Jensen's inequality. Under this additional assumption, the excess one-period bond returns under the different hypotheses can be shown to be of $\frac{1}{2} \text{Var}[r_{t \rightarrow t+1}^{(n)} - y_t^{(1)}]$ order. Therefore, the difference between the one-period excess bond returns of different PEH forms is $\text{Var}[r_{t \rightarrow t+1}^{(n)} - y_t^{(1)}]$. Using sample means and standard deviations, we can get an estimate and a standard error of the above quantity. This magnitude is very small for short-term bonds and becomes significant only for long-term bonds; hence, the differences between different forms of the PEH are small except for very long term zero-coupon bonds. Thus, the data reject all forms of the PEH at the short end, but reject no forms of the PEH at the long end of the

term structure. In this sense, the distinction between the different forms of the PEH is not critical for evaluating this hypothesis.

Continuous Compounding

Most empirical research, though, uses neither of the above PEH forms, but a log form of them. Once the PEH is formulated in logs, all the forms of the PEH become equivalent. Using log returns, the counterparts of equations (12), (13), and (14) are

$$y_t^{(1)} = \mathbb{E}_t^{\mathbb{P}}[r_{t \rightarrow t+1}^{(n)}] \quad (16)$$

$$y_t^{(n)} = (1/n) \sum_{i=0}^{n-1} \mathbb{E}_t^{\mathbb{P}}[y_{t+i}^{(1)}] \quad (17)$$

$$f_t^{(\tau-1, \tau)} = \mathbb{E}_t^{\mathbb{P}}[y_{t+\tau-1}^{(1)}] \quad (18)$$

The empirical literature uses equations (17) and (18) in order to construct two related notions of term premia that have played a prominent role in the literature of expected bond returns: the yield premium,

$$c_t^{(n)} \equiv y_t^{(n)} - \frac{1}{n} \sum_{i=0}^{n-1} \mathbb{E}_t^{\mathbb{P}}[y_{t+i}^{(1)}] \quad (19)$$

and the forward term premium,

$$p_t^{(n)} \equiv f_t^{(n, n+1)} - \mathbb{E}_t^{\mathbb{P}}[y_{t+n}^{(1)}]. \quad (20)$$

Derivations of PEH- and EH-tested formulas follow below.

Pure Expectations Hypothesis in Continuous Time

Cox *et al.* [27] restate the PEH forms in continuous time and prove that the different forms are incompatible. The equivalent of expression (12) in continuous time is created by assuming that the holding period is the shortest possible, that is, infinitesimal. In this case, the PEH takes the following form:

$$\frac{\mathbb{E}_t^{\mathbb{P}}[dP_t^{(\tau)}]}{P_t^{(\tau)}} = r(t) dt \quad (21)$$

This expression states that all bonds have the same expected infinitesimal return, equal to the short-term interest rate. However, the above expression is

also known to hold under $\mathbb{Q}$; under the risk-neutral measure, all assets have the same expected return, equal to the risk-free rate. This implies that this form of the PEH postulates that $\mathbb{P} = \mathbb{Q}$. The expression's (13) continuous time equivalent is

$$\frac{1}{P_t^{(\tau)}} = \mathbb{E}_t^{\mathbb{P}} \left[e^{\int_t^{t+\tau} r(s) ds} \right] \quad (22)$$

This statement equates the guaranteed return from holding any zero-coupon bond to maturity with the total return expected from rolling over a series of short-term period bonds. The continuous time equivalent of equation (14) is

$$\frac{-\partial P_t^{(\tau)} / \partial \tau}{P_t^{(\tau)}} = \mathbb{E}_t^{\mathbb{P}}[r(t + \tau)] \quad (23)$$

The left-hand side of the equation is the current infinitesimal forward rate at time $t + \tau$, and the right-hand side is the expected future spot rate at $t + \tau$. Integrating the last equation and applying the boundary condition $P_t^{(0)} = 1$ gives

$$-\ln[P_t^{(\tau)}] = \int_t^{t+\tau} \mathbb{E}_t^{\mathbb{P}}[r(s)] ds \quad (24)$$

Formulating the PEH in continuous time makes the pairwise incompatibility of equations (21), (22), and (24) transparent. If we define the random variable $\tilde{X} \equiv \exp\left(-\int_t^{t+\tau} r(s) ds\right)$, then these equations can be rewritten as

$$P = \mathbb{E}_t^{\mathbb{P}}[\tilde{X}] \quad (25)$$

$$P^{-1} = \mathbb{E}_t^{\mathbb{P}}[\tilde{X}^{-1}] \quad (26)$$

$$\ln(P) = \mathbb{E}_t^{\mathbb{P}}[\ln \tilde{X}] \quad (27)$$

By invoking Jensen's inequality, one can show that the yields to maturity implied from equations (21), (22), and (24) satisfy the relationship (with some abuse of notation):

$$y_t^{(\tau)(21)} \leq y_t^{(\tau)(22)} \leq y_t^{(\tau)(24)} \quad (28)$$

In this model, it is also easy to see that the expected excess returns are positive in all hypotheses except in equation (21).

Perhaps the most impacting result of Cox *et al.* [27] is the characterization of the PEH forms that can be the result of a (no-arbitrage) equilibrium model.

They examine whether there exist pricing kernels (i.e., prices of risk) that can satisfy the resulting pricing PDE in the economy and at the same time satisfy the form of the PEH under examination. They conclude that only equation (21) can be sustained by an equilibrium model. By definition, equation (21) implies that $\mathbb{Q} = \mathbb{P}$; therefore, selecting $\Lambda(t) = \mathbf{0}$ gives rise to a valid pricing kernel. Cox *et al.* [27] prove that the other forms do not give rise to a valid pricing kernel. However, McCulloch [57] later showed that their claim is incorrect. Working in generalizing a preexisting discrete time model to continuous time, he shows that there exists an equilibrium economy that also gives rise to equation (23).

Expectation Hypothesis

As described above, the difference between the EH and the PEH is that the term premia under the PEH are assumed to be zero, whereas under the EH they are assumed to be constant. Therefore, to formulate the different forms of the EH we need to add in each form of the PEH a constant term that depends only upon the remaining time to maturity of the corresponding bond considered in each form of the EH.

Even though the different forms of the PEH are generally incompatible, Campbell [21] showed that the different forms of the EH *are not* incompatible and he derived a general equilibrium model that sustained several forms of the EH at the same time. His model is set up in continuous time. In addition, special cases of the models examined in [16] and [47] provide equilibrium models that give rise to constant term premia [36].

Next, we show the most commonly tested equations of the EH in the literature.

Tests of the Expectations Hypothesis

The EH has been under scrutiny at least since the work of Macaulay [54]. In this study, Macaulay emphasizes the low (given the EH is true) correlation between forward rates and subsequent spot rates. Since then, the EH has been tested in hundreds of studies, and in all of them—with only few exceptions—has been rejected. Some of the early papers that test the EH are those of Sutch [74], Shiller [69, 70, 71], Modigliani and Shiller [59], Sargent [67, 68].

Fama [39, 40] and Fama and Bliss [41] also present challenges of the EH where they find evidence of rich patterns of variation in expected returns across time and maturities. Keim and Stambaugh [50], Fama and French [42], and Campbell and Ammer [23] show that yield spreads help to forecast excess return on bonds as well as on other long-term assets.

Perhaps the most widely cited tests of the EH are the Campbell and Shiller [24] regressions based on the equations:

$$\mathbb{E}_t^{\mathbb{P}} \left[y_{t+m}^{(n-m)} - y_t^{(n)} \right] = \alpha_{nm} + \frac{m}{n-m} (y_t^{(n)} - y_t^{(m)}) \quad (29)$$

which are a more general form of the regressions in [30] based on the following equations:

$$\mathbb{E}_t^{\mathbb{P}} \left[y_{t+1}^{(n-1)} - y_t^{(n)} \right] = \alpha_{n1} + \frac{1}{n-1} (y_t^{(n)} - r_t) \quad (30)$$

that are used to assess the goodness of fit of the different term structure models. The derivations of the above equations are shown in detail in the above-mentioned papers and in [73]. In short, from equation (4) we have that the one-period excess bond continuously compounded return is equal to

$$\begin{aligned} \mathbb{E}_t^{\mathbb{P}} \left[r_{t \rightarrow t+1}^{(n)} - r_t \right] &= -(n-1) \mathbb{E}_t^{\mathbb{P}} \left[y_{t+1}^{(n-1)} - y_t^{(n)} \right] \\ &\quad + (y_t^{(n)} - r_t) \end{aligned} \quad (31)$$

and, it can also be shown that it is equal to

$$\begin{aligned} \mathbb{E}_t^{\mathbb{P}} \left[r_{t \rightarrow t+1}^{(n)} - r_t \right] &= -(n-1) \mathbb{E}_t^{\mathbb{P}} \left[c_{t+1}^{(n-1)} - c_t^{(n)} \right] \\ &\quad + p_t^{(n-1)} \end{aligned} \quad (32)$$

where $c_t^{(n)}$ and $p_t^{(n)}$ are the yield and forward premia defined in equations (19) and (20), respectively. The last expression implies that if the PEH holds (i.e., $c_t^{(n)} = 0$, $p_t^{(n)} = 0$) then the expected excess returns are zero, whereas if the EH holds (i.e., $c_t^{(n)} = c(n)$, $p_t^{(n)} = p(n)$), then the expected excess returns are constants that depend on the time to maturity n . Combining equations (31) and (32) gives rise to equation (30), the well-known LPY regressions of Dai and Singleton [30].

Campbell and Shiller [24] and Dai and Singleton [30], among others, document the failure of both the

regressions (31) and (32), which are true under the EH. According to these equations, the coefficients of the nonconstant terms when regressing $y_{t+m}^{(n-m)} - y_t^{(n)}$ onto $\frac{m}{n-m}(y_t^{(n)} - y_t^{(m)})$, or $\frac{1}{n-1}(y_t^{(n)} - r_t)$ if $m = 1$, should be equal to unity. Not only are the estimated coefficients not unity but also they are often statistically significantly negative, particularly for large n . This means that the EH fails more significantly for long-term bonds. The intuition of the EH is that increases in the slope of term structure ($y_t^{(n)} - r_t$) reflect expectations of rising future short spot rates. For the “buy an n -period bond and hold it to maturity” investment strategy to match, on average, the returns from rolling over short rates in a rising short rate environment, the price of the long bond should decrease such that the yield increases ($y_{t+1}^{(n-1)} > y_t^{(n)}$). The regression results suggest that the slope of the yield curve does not even forecast correctly the direction of the changes in the long-bond yields. The underreaction of long rates to spread term changes has also been the study in [56]. Elaborating and further documenting this underreaction, Campbell [22] finds that yield spreads do not forecast short-run changes in long yields (against EH) but do forecast long-run changes in short yields (consistent with EH).

Backus *et al.* [7] tested the EH by running regressions based on analogous equations for the forward rates:

$$\mathbb{E}_t^{\mathbb{P}} \left[f_{t+1}^{(n-1,n)} - r_t \right] = \alpha + (f_t^{(n-1,n)} - r_t) \quad (33)$$

They also find that the regression coefficients of $(f_t^{(n-1,n)} - r_t)$ are not unity as the null hypothesis sets, but slightly less than one and significantly different than one. They also show that the small differences of the estimated coefficients with unity in the above regressions do not constitute separate or weaker findings against the EH than the deviations of the Campbell–Shiller coefficients from unity, but are actually the same. They constructed a one-factor model and showed that the small differences in the coefficients of the Backus *et al.* [7] regressions from their null value translate into large negative values for the Campbell–Shiller coefficients.

Similar to the above forward regressions are the forward term regressions tested in [43, 72] Shiller *et al.* [72] use a log-linearized model [70, 71] that allows them to test several models of the EH without having to use discount rates (which are

easily available for maturities up to a year, but for longer maturities the rates have to be constructed using spline methods) but with the use of the easily observable coupon-bond yields. One of their regressions is based on the EH equation:

$$\mathbb{E}_t^{\mathbb{P}} \left[y_{t+m}^{(n-m)} - y_t^{(m)} \right] = \alpha_{n,m} + (f_t^{(m,n)} - y_t^{(m)}) \quad (34)$$

Using coupon-bond yields does not change the results that future bond yield changes cannot be predicted by the current term spread of forward spread. Still, even the direction cannot be predicted correctly. They suggest time-varying risk premia as a plausible solution to the failure of the EH. Froot [43] also tests the same equation trying to understand whether its failure is the result of time varying term premia or that expected future spot rates under- or overreact to changes in short rates. He finds that for short maturities its failure is due to variation in term premia, but this is not true for long maturities.

A recent paper that has received a lot of attention is Cochrane and Piazzesi [26]. Cochrane and Piazzesi [26] have revisited the forecasting regressions of Fama and Bliss using the term structure of forward rates instead of a single forward rate. Their most notable finding is that the coefficients from regressing excess bond returns over one year onto the one year forward rates for the next five years exhibit a tentlike shape for all maturity bonds. The tentlike shape similarities for all bond maturities suggest that a single common factor may be underlying the predictability of excess bond returns for all maturities. Another very interesting fact in [26] is the high R^2 s generated in the above regressions. The R^2 s range between 36% and 39%. This is substantially more predictability than in [41] using a single forward factor.

A series of other papers have also examined alternative reasons for the failure of the EH: Mankiew and Miron [55] find that interest rate movements were more predictable before the founding of the Federal Reserve in 1913, and the downward bias appears to be smaller in that period. Campbell and Ammer [23] emphasize that long-term bond yields vary primarily in response to changing expected inflation. Rudebusch [65] argues that contemporary Federal Reserve operating procedures lead to predictable interest rate movements in the very short run and the very long run, but tend to smooth away predictable movements in the medium run. Balduzzi *et al.* [8]

argue that spreads between short-term rates and the overnight federal reserve funds rate are mainly driven by expectations of changes in the target, and not by the transitory dynamics of the overnight rate around the target. Hence, the bias in tests of the EH that they document can be mainly attributed to the erroneous anticipation of future changes in monetary policy.

Several studies have examined the small-sample bias of the regression coefficients. Bekaert and Hodrick [10] argue that the past use of large sample critical regions, instead of their small-sample counterparts, may have overstated the evidence against the expectations theory. They find that the evidence against the EH for these interest rates and exchange rates is much less strong than under asymptotic inference. Other studies, though, such as Backus *et al.* [7] and Bekaert, Hodrick, and Marshall [11], find that the small-sample properties of the regressions like the ones shown in this article are actually biased upward; this means the *true* Campbell–Shiller coefficients are more negative than the ones estimated in the regressions, heightening the puzzle related with the failure of the EH.

Researchers have also looked at the validity of the EH outside the United States. The tendency in these studies is to find Campbell–Shiller coefficients that are less than zero but less negative than the US data results. Some of those studies that are done primarily for European countries and show mixed results are Bekaert and Hodrick [10], Boero and Torricelli [14], Evans [37], Gerlach and Smets [44], Hardouvelis [48], Kugler [51].

Conclusion

The EH constitutes several hypotheses that were generated to understand bond returns and their yields through the help of other maturity bond returns or investment strategies and forward rates. We showed that these hypotheses can be formulated in many different ways and using different models (continuous time *vs* discrete and continuous compounding *vs* discrete). The different hypotheses are not equivalent. Therefore to test the validity of the EH numerous different expressions have to be tested.

The consensus is that the EH fails in the US data. Its failure, though, is less strong or mixed for the non-US data. Researchers challenging and

trying to understand the failure of the EH have hinted on different reasons that may give rise to time-varying expected excess returns on bonds. Even though the understanding of the failure of the EH is not complete, part of the literature is devoted to creating models that better capture this failure and that better replicate the data.

In this strand we can put the papers of reduced form (affine or nonaffine, with macro or without macrovariables) term structure models, such as Ahn *et al.* [1], Ang and Bekaert [2], Bansal and Zhou [9], Bikbov and Chernov [12, 13], Buraschi *et al.* [17], Dai and Singleton [29, 30, 32], Diebold *et al.* [33], Duarte [34], Evans [38], Leippold and Wu [52], and Naik and Lee [59], and those of the structural form models with or without macrovariables, such as, Ang *et al.* [3, 6], Ang and Piazzesi [4, 5], Brandt and Wang [15], Buraschi and Jiltsov [18, 19], Dai [28], Greenwood and Vayanos [45], Guibaud *et al.* [46], Piazzesi [62], Piazzesi and Schneider [63], Rudebusch and Wu [65, 66], Vayanos and Vila [75], and Wachter [76].

Acknowledgments

The author thanks Aggie Moon for providing research assistantship. The author takes the responsibility for errors if any.

End Notes

^aCochrane [25], Dai and Singleton [31], Duffie [35], Nielsen [60], Piazzesi [61], Singleton [73].

^bIn the EH literature the term *yield premium* is used to denote the difference of the *n*th root of the terms in the left- and right-hand side of equation (13).

^cIn the EH literature the term *forward premium* is used to denote the difference of the terms in the left- and right-hand side of equation (14).

References

- [1] Ahn, D.-H., Dittmar, R.F. & Gallant, A.R. (2002). Quadratic term structure models: theory and evidence, *Review of Financial Studies* **15**, 243–288.
- [2] Ang, A. & Bekaert, G. (2002). Regime switches in interest rates, *Journal of Business and Economic Statistics* **20**, 163–182.
- [3] Ang, A., Dong, S. & Piazzesi, M. (2007). *No-Arbitrage Taylor Rules*, National Bureau of Economic Research.

-
- [4] Ang, A. & Piazzesi, M. (2003a). A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables, *Journal of Monetary Economics* **50**, 745–787.
- [5] Ang, A. & Piazzesi, M. (2003b). A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables, *Journal of Monetary Economics* **50**, 745–787.
- [6] Ang, A., Piazzesi, M. & Wei, M. (2006). What does the yield curve tell us about the GDP growth? *Journal of Econometrics* **131**, 359–403.
- [7] Backus, D., Foresi, S., Mozumbar, A. & Wu, L. (2001). Predictable changes in yields and forward rates, *Journal of Financial Economics* **59**, 281–311.
- [8] Balduzzi, P., Bertola, G. & Foresi, S. (1997). A model of target changes and the term structure of interest rates, *Journal of Monetary Economics* **39**, 223–249.
- [9] Bansal, R. & Zhou, H. (2002). Term structure of interest rates with regime shifts, *Journal of Finance* **57**, 1997–2044.
- [10] Bekaert, G. & Hodrick, R.J. (2001). Expectations hypothesis tests, *Journal of Finance* **56**, 1357–1394.
- [11] Bekaert, G., Hodrick, R.J. & Marshall, D.A. (1997). On biases in tests of the expectations hypothesis of the term structure of interest rates, *Journal of Financial Economics* **44**, 309–348.
- [12] Bikbov, R. & Chernov, M. (2005). *Term Structure and Volatility: Lessons from the Eurodollar Futures and Options*. Working Paper, London Business School, London.
- [13] Bikbov, R. & Chernov, M. (2006). *No-Arbitrage Macroeconomic Determinants*. Working Paper, London Business School.
- [14] Boero, G. & Torricelli, C. (1997). *The Expectations Hypothesis of the Term Structure: Evidence for Germany*. Working Paper CRENoS 1997/4. Centre for North South Economic Research, University of Cagliari and Sassari, Sardinia, revised.
- [15] Brandt, M.W. & Wang, K.Q. (2003). Time-varying risk aversion and unexpected inflation, *Journal of Monetary Economics* **50**, 1457–1498.
- [16] Breeden, D. (1986). Consumption, production and interest rates: a synthesis, *Journal of Financial Economics* **7**, 265–296.
- [17] Buraschi, A., Cieslak, A. & Trojani, F. (2007). *Correlation Risk and the Term Structure of Interest Rates*. Working Paper, Imperial College, U.K.
- [18] Buraschi, A. & Jiltsov, A. (2005). Inflation risk premia and the expectations hypothesis, *Journal of Financial Economics* **75**, 429–490.
- [19] Buraschi, A. & Jiltsov, A. (2007). Term structure of interest rates implications of habit persistence, *Journal of Finance* **62**, 3009–3063.
- [20] Campbell, J.Y. (1986a). Bond and stock returns in a simple exchange model, *Quarterly Journal of Economics* **101**, 785–803.
- [21] Campbell, J.Y. (1986b). A defense of traditional hypotheses about the term structure of interest rates, *Journal of Finance* **41**, 183–193.
- [22] Campbell, J.Y. (1995). Some lessons from the yield curve, *Journal of Economic Perspectives* **9**, 129–152.
- [23] Campbell, J.Y. & Ammer, J. (1993). What moves the stock and bond markets? A variance decomposition for long-term asset returns, *Journal of Finance* **48**, 3–37.
- [24] Campbell, J.Y. & Shiller, R.J. (1991). Yield spreads and interest rate movements: A Bird’s eye view, *Review of Economic Studies* **58**, 495–514.
- [25] Cochrane, J. (2000). *Asset Pricing*, Princeton University Press, Princeton.
- [26] Cochrane, J. & Piazzesi, M. (2005). Bond risk premia, *American Economic Review* **95**, 138–160.
- [27] Cox, J.C., Ingersoll, J.C. & Ross, S.A. (1981). A re-examination of traditional hypotheses about the term structure of interest rates, *Journal of Finance* **36**, 769–799.
- [28] Dai, Q. (2003). *Term Structure Dynamics in a Model with Stochastic Internal Habit*. Working Paper, New York University.
- [29] Dai, Q. & Singleton, K. (2000). Specification analysis of affine term structure models, *Journal of Finance* **55**, 1943–1978.
- [30] Dai, Q. & Singleton, K. (2002). Expectations puzzles, time-varying risk premia, and affine models of the term structure, *Journal of Financial Economics* **63**, 415–442.
- [31] Dai, Q. & Singleton, K. (2003a). Fixed-income pricing, in *Handbook of Economics and Finance*, C. Constantinides, M. Harris & R. Stulz, eds, North Holland, Amsterdam.
- [32] Dai, Q. & Singleton, K. (2003b). Term structure dynamics in theory and reality, *Review of Financial Studies* **16**, 631–678.
- [33] Diebold, F., Rudebusch, G. & Aruoba, B. (2006). The macroeconomy and the yield curve: a dynamic latent factor approach, *Journal of Econometrics* **131**, 309–338.
- [34] Duarte, J. (2004). Evaluating an alternative risk preference in affine term structure models, *Review of Financial Studies* **17**, 370–404.
- [35] Duffie, D. (1996). *Dynamic Asset Pricing Theory*, Princeton University Press, Princeton.
- [36] Dunn, K.B. & Singleton, K.J. (1986). Modeling the term structure of interest rates under non-separable utility and durability of goods, *Journal of Financial Economics* **17**, 27–55.
- [37] Evans, M.D. (2000). *Regime Shifts, Risk, and the Term Structure*. Working Paper, Georgetown University.
- [38] Evans, M.D. (2003). Real risk, inflation risk, and the term structure, *The Economic Journal* **113**, 345–389.
- [39] Fama, E.F. (1984a). The information in the term structure, *Journal of Financial Economics* **13**, 509–528.
- [40] Fama, E.F. (1984b). Term premiums in bond returns, *Journal of Financial Economics* **13**, 529–546.
- [41] Fama, E.F. & Bliss, R.R. (1987). The information in long-maturity forward-rates, *American Economic Review* **77**, 680–692.

- [42] Fama, E.F. & French, K.R. (1989). Business conditions and expected returns on stocks and bonds, *Journal of Financial Economics* **29**, 23–49.
- [43] Froot, K.A. (1989). New hope for the expectations hypothesis of the term structure of interest rates, *Journal of Finance* **44**, 283–305.
- [44] Gerlach, S. & Smets, F. (1997). The term structure of Euro-rates: some evidence in support of the expectations hypothesis, *Journal of International Money and Finance* **16**, 305–321.
- [45] Greenwood, R. & Vayanos, D. (2008). *Bond Supply and Excess Bond Returns*. Working Paper, London School of Economics.
- [46] Guibaud, S., Nosbusch, Y. & Vayanos, D. (2007). *Preferred Habitat and the Optimal Maturity Structure of Government Debt*. Working Paper, London School of Economics.
- [47] Hansen, L.P. & Singleton, K. (1983). Stochastic consumption, risk aversion, and the temporal behavior of asset returns, *Journal of Political Economy* **91**, 249–268.
- [48] Hardouvelis, G. (1994). The term structure spread and future changes in long and short rates in G7 countries, *Journal of Monetary Economics* **33**, 255–283.
- [49] Hicks, J.R. (1939) *Value and Capital*, Oxford University Press, Oxford.
- [50] Keim, D.B. & Stambaugh, R.F. (1986). Predicting returns in the stock and bond markets, *Journal of Financial Economics* **17**, 357–390.
- [51] Kugler, P. (1997). Central bank policy reaction and the expectations hypothesis of the term structure, *International Journal of Financial Economics* **2**, 164–181.
- [52] Leippold, M. & Wu, L. (2003). Design and estimation of quadratic term structure models, *European Finance Review* **7**, 47–73.
- [53] Lutz, F.A. (1940). The structure of interest rates, *The Quarterly Journal of Economics* **55**, 36–63.
- [54] Macaulay, F.R. (1938). *Some Theoretical Problems Suggested by the Movements of Interest Rates, Bond Yields, and Stock Prices in the United States Since 1856*. NBER Working Paper Series, New York.
- [55] Mankiew, G.N. & Miron, J.A. (1986). The changing behavior of the term structure of interest rates, *Quarterly Journal of Economics* **101**, 211–228.
- [56] Mankiew, G.N. & Summers, L.H. (1984). Do long-term interest rates overreact to short-term rates? *Brookings Papers on Economic Activity* **1**, 223–242.
- [57] McCulloch, H.J. (1993). A reexamination of traditional hypotheses about the term structure: a comment, *Journal of Finance* **48**, 779–789.
- [58] Modigliani, F. & Shiller, R.J. (1973). Inflation, rational expectations, and the term structure of interest rates, *Economica* **40**, 12–43.
- [59] Naik, V. & Lee, M.H. (1997). *Yield Curve Dynamics with Discrete Shifts in Economic Regimes: Theory and Estimation*. Unpublished Working Paper, Faculty of Commerce, University of British Columbia.
- [60] Nielsen, L.T. (1999). *Pricing and Hedging of Derivative Securities*, Oxford University Press, Oxford.
- [61] Piazzesi, M. (2003). Affine term structure models, *Handbook of Financial Econometrics*, Elsevier, p. 828.
- [62] Piazzesi, M. (2005). Bond yields and the federal reserve, *Journal of Political Economy* **113**, 311–344.
- [63] Piazzesi, M. & Schneider, M. (2007). Equilibrium yield curves, *NBER/Macroeconomics Annual* **21**, 389–442.
- [64] Rudebusch, G.D. (1995). Federal reserve interest rate targeting, rational expectations, and the term structure, *Journal of Monetary Economics* **35**, 245–274.
- [65] Rudebusch, G.D. & Wu, T. (2004a). *A Macro-Finance Model of the Term Structure, Monetary Policy, and the Economy*. Working Paper, Federal Reserve Bank of San Francisco.
- [66] Rudebusch, G.D. & Wu, T. (2004b). *The Recent Shift in Term Structure Behavior from a No-Arbitrage Macro-Finance Perspective*. Working Paper, Federal Reserve Bank of San Francisco.
- [67] Sargent, T.J. (1972). Rational expectations and the term structure of interest rates, *Journal of Money, Credit and Banking* **4**, 74–97.
- [68] Sargent, T.J. (1979). A note on maximum likelihood estimation of the rational expectations model of the term structure, *Journal of Monetary Economics* **5**, 133–143.
- [69] Shiller, R.J. (1972). *Rational Expectations and the Term Structure of Interest Rates*. Ph.D. Dissertation, MIT.
- [70] Shiller, R.J. (1979). The volatility of long-term interest rates and expectations models of the term structure, *Journal of Political Economy* **87**, 1190–1219.
- [71] Shiller, R.J. (1981). Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review* **71**, 421–436.
- [72] Shiller, R.J., Campbell, J.Y. & Schoenholtz, K.L. (1983). Forward rates and future policy: interpreting the term structure of interest rates, *Brookings Papers on Economic Activity* **14**(1), 173–224.
- [73] Singleton, K.J. (2006). *Empirical Dynamic Asset Pricing*, Princeton University Press, Princeton.
- [74] Sutch, R.C. (1970). Expectations, risk, and the term structure of interest rates, *Journal of Finance* **25**, 703.
- [75] Vayanos, D. & Vila, J.-L. (2007). *A Preferred-Habitat Model of the Term Structure of Interest Rates*. Working Paper, London School of Economics.
- [76] Wachter, J.A. (2006). A consumption-based model of the term structure of interest rates, *Journal of Financial Economics* **79**, 365–399.

Further Reading

- Longstaff, F.A. (2000). The term structure of very short-term rates: new evidence for the expectations hypothesis, *Journal of Financial Economics* **58**, 397–415.
- Sutch, R.C. (1968). *Expectations, Risk, and the Term Structure of Interest Rates*. Dissertation, MIT.

ANTONIOS SANGVINATSOS

Stochastic Discount Factors

Economic agents make investment decisions within active and liquid financial markets. Capital is allocated today in exchange for some future income stream. If there is no uncertainty regarding the future payoff of an investment opportunity, the yield that will be asked on the investment will equal the risk-free interest rate prevailing for the time period covering the time of investment until the time of the payoff. However, in the presence of any payoff uncertainty at the time of undertaking an investment venture, economic agents will typically ask for *risk compensation*, and thus for some *investment-specific* yield, which will discount the expected future payoff stream. The yields that particular agents ask for depend both on their statistical views on possible future outcomes, as well as their attitudes toward risk.

Yields vary across different investment opportunities and their interrelations are difficult to explain. For the *same* agent, a different discounting factor has to be used for every separate valuation occasion. If, however, one is ready to accept discounting that varies *randomly* with the possible outcomes, and therefore accepts the concept of a *stochastic* discount factor, then a very economically consistent theory can be developed. Asset valuation becomes a matter of randomly discounting payoffs under different states of nature and weighing them according to the agent's probability structure. The advantages of this approach are obvious, since a *single* discounting mechanism suffices to describe how *any* asset is priced by the agent.

We discuss the theory of stochastic discount factors first in a discrete-time, finite state space and then in the more practical case of Itô-process models.

Stochastic Discount Factors in Discrete Probability Spaces

We start by introducing all relevant ideas in a very simple one-time-period framework and finite states of the world. There are plenty of textbooks with vast exposition on these and other related themes (e.g., [1] or the first chapters of [9]—see also [5] for the general state-space case).

The Setup

Consider a very simplistic example of an economy, where there are only two dates of interest, represented by times $t = 0$ (today) and $t = T$ (the financial-planning horizon). There are several states of nature possible at time T and, for the time being, these are represented as a *finite* set Ω . Only one $\omega \in \Omega$ will be revealed at time T , but this is not known in advance today.

In the market, there is a *baseline* asset with a price process $S^0 = (S^0_t)_{t=0,T}$. Here, S^0_0 is a strictly positive constant and $S^0_T(\omega) > 0$ for all $\omega \in \Omega$. The process $\beta := S^0_0/S^0_T$ is called the *deflator*. It is customary to regard this baseline asset as *riskless*, providing a simple annualized interest rate $r \in \mathbb{R}_+$ for investment from today to time T ; in this case, $S^0_0 = 1$ and $S^0_T = 1 + rT$. This viewpoint is *not* adapted here, since it is unnecessary.

Together with the baseline asset, there exist d other liquid *traded* assets whose prices S^i_t , $i = 1, \dots, d$ today are known constants, but the prices S^i_T , $i = 1, \dots, d$, at day T depend on the outcome $\omega \in \Omega$, that is, they are random variables.

Agent Portfolio Selection via Expected Utility Maximization

Consider an economic agent in the market as described above. Faced with inherent uncertainty, the agent postulates some likelihood on the possible outcomes, modeled via a *probability measure* $\mathbb{P} : \Omega \mapsto [0, 1]$ with $\sum_{\omega \in \Omega} \mathbb{P}[\omega] = 1$. This gives rise to a probability $\mathbb{P}$ on the subsets of Ω defined via $\mathbb{P}[A] = \sum_{\omega \in A} \mathbb{P}(\omega)$ for all $A \subseteq \Omega$. This probability can either be *subjective*, that is, coming from views that are agent specific, or *historical*, that is, arising from statistical considerations via some estimation procedure.

Economic agents act in the market and optimally invest to maximize their satisfaction. Each agent has some *preference structure* on the possible future random payoffs that is represented here via the *expected utility* paradigm.^a There exists a *continuously differentiable*, *increasing*, and *strictly concave* function $U : \mathbb{R} \mapsto \mathbb{R}$, such that the agent will prefer a random payoff $\xi : \Omega \mapsto \mathbb{R}$ from another random payoff $\zeta : \Omega \mapsto \mathbb{R}$ at time T if and only if $\mathbb{E}^{\mathbb{P}}[U(\xi)] \leq \mathbb{E}^{\mathbb{P}}[U(\zeta)]$, where $\mathbb{E}^{\mathbb{P}}$ denotes expectation with respect to the probability $\mathbb{P}$.

Starting with capital $x \in \mathbb{R}$, an economic agent chooses at day zero a strategy $\theta \equiv (\theta^1, \dots, \theta^d) \in \mathbb{R}^d$, where θ^j denotes the units from the j th asset held in the portfolio. What remains, $x - \sum_{i=1}^d \theta^i S_0^i$, is invested in the baseline asset. If $X^{(x, \theta)}$ is the wealth generated starting from capital x and investing according to θ , then $X_0^{(x, \theta)} = x$ and

$$\begin{aligned} X_T^{(x, \theta)} &= \left(x - \sum_{i=1}^d \theta^i S_0^i \right) \frac{S_T^0}{S_0^0} + \sum_{i=1}^d \theta^i S_T^i \\ &= x \frac{S_T^0}{S_0^0} + \sum_{i=1}^d \theta^i \left(S_T^i - \frac{S_T^0}{S_0^0} S_0^i \right) \end{aligned} \quad (1)$$

or, in deflated terms, $\beta_T X_T^{(x, \theta)} = x + \sum_{i=1}^d \theta^i (\beta_T S_T^i - S_0^i)$. The agent's objective is to choose a strategy in such a way as to *maximize expected utility*, that is, find θ_* such that

$$\mathbb{E}^\mathbb{P} \left[U \left(X_T^{(x, \theta_*)} \right) \right] = \sup_{\theta \in \mathbb{R}^d} \mathbb{E}^\mathbb{P} \left[U \left(X_T^{(x, \theta)} \right) \right] \quad (2)$$

The above problem will indeed have a solution if and only if no arbitrage exists in the market. By definition, an *arbitrage* is a wealth generated by some $\theta \in \mathbb{R}^d$ such that $\mathbb{P}[X_T^{(x, \theta)} \geq 0] = 1$ and $\mathbb{P}[X_T^{(x, \theta)} > 0] > 0$. It is easy to see that arbitrage exists in the market if and only if $\sup_{\theta \in \mathbb{R}^d} \mathbb{E}^\mathbb{P}[U(X_T^{(x, \theta)})]$ is *not* attained by some $\theta_* \in \mathbb{R}^d$. Assuming, then, the *no-arbitrage* (NA) condition, concavity of the function $\mathbb{R}^d \ni \theta \mapsto \mathbb{E}^\mathbb{P}[U(X_T^{(x, \theta)})]$ will imply that the first-order conditions

$$\frac{\partial}{\partial \theta^i} \Big|_{\theta=\theta_*} \mathbb{E}^\mathbb{P} \left[U \left(X_T^{(x, \theta)} \right) \right] = 0, \quad \text{for all } i = 1, \dots, d \quad (3)$$

will provide the solution θ_* to the problem. Since the expectation is just a finite sum, the differential operator can pass inside, and then the first-order conditions for optimality are

$$\begin{aligned} 0 &= \mathbb{E}^\mathbb{P} \left[\frac{\partial}{\partial \theta^i} \Big|_{\theta=\theta_*} U \left(X_T^{(x, \theta)} \right) \right] \\ &= \mathbb{E}^\mathbb{P} \left[U' \left(X_T^{(x, \theta_*)} \right) \left(S_T^i - \frac{S_T^0}{S_0^0} S_0^i \right) \right], \\ i &= 1, \dots, d \end{aligned} \quad (4)$$

The above is a nonlinear system of d equations to be solved for d unknowns $(\theta_*^1, \dots, \theta_*^d)$. Under NA, the system 4 has a solution θ_* . Actually, under a trivial nondegeneracy condition in the market, the solution is unique; even if the optimal strategy θ_* is not unique, strict concavity of U implies that the optimal wealth $X_T^{(x, \theta_*)}$ generated is unique.

A little bit of algebra on equation (4) gives, for all $i = 1, \dots, d$,

$$\begin{aligned} S_0^i &= \mathbb{E}^\mathbb{P} [Y_T S_T^i], \quad \text{where} \\ Y_T &:= \frac{U' \left(X_T^{(x, \theta_*)} \right)}{\mathbb{E}^\mathbb{P} \left[(S_T^0/S_0^0) U' \left(X_T^{(x, \theta_*)} \right) \right]} \end{aligned} \quad (5)$$

Observe that since U is continuously differentiable and strictly increasing, U' is a strictly positive function, and therefore $\mathbb{P}[Y_T > 0] = 1$. Also, equation (5) also holds trivially for $i = 0$. Note that the random variable Y_T that was obtained above depends on the utility function U , the probability $\mathbb{P}$, as well as on the initial capital $x \in \mathbb{R}$.

Definition 1 *In the model described above, a process $Y = (Y_t)_{t=0, T}$ will be called a stochastic discount factor if $\mathbb{P}[Y_0 = 1, Y_T > 0] = 1$ and $S_0^i = \mathbb{E}^\mathbb{P} [Y_T S_T^i]$ for all $i = 0, \dots, d$.*

If Y is a stochastic discount factor, using equation (1), one can actually show that

$$\mathbb{E}^\mathbb{P} [Y_T X_T^{(x, \theta)}] = x, \quad \text{for all } x \in \mathbb{R} \text{ and } \theta \in \mathbb{R}^d \quad (6)$$

In other words, the process $Y X^{(x, \theta)}$ is a $\mathbb{P}$ -martingale for all $x \in \mathbb{R}$ and $\theta \in \mathbb{R}^d$.

Connection with Risk-neutral Valuation

Since $\mathbb{E}^\mathbb{P} [S_T^0 Y_T] = S_0^0 > 0$, we can define a probability mass $\mathbb{Q}$ by requiring that $\mathbb{Q}(\omega) = (S_T^0(\omega)/S_0^0) Y_T(\omega) \mathbb{P}(\omega)$, which defines a probability $\mathbb{Q}$ on subsets of Ω in the obvious way. Observe that, for any $A \subseteq \Omega$, $\mathbb{Q}[A] > 0$ if and only if $\mathbb{P}[A] > 0$; we say that the probabilities $\mathbb{P}$ and $\mathbb{Q}$ are *equivalent* and we denote this by $\mathbb{Q} \sim \mathbb{P}$. Now, rewrite equation (5) as

$$S_0^i = \mathbb{E}^\mathbb{Q} [\beta_T S_T^i], \quad \text{for all } i = 0, \dots, d \quad (7)$$

A probability $\mathbb{Q}$, equivalent to $\mathbb{P}$, with the property prescribed in equation (7) is called *risk-neutral* or an *equivalent martingale measure*. In this simple framework, stochastic discount factors and risk-neutral probabilities are in one-to-one correspondence. In fact, more can be said.

Theorem 1 [Fundamental Theorem of Asset Pricing] *In the discrete model as described previously, the following three conditions are equivalent:*

1. *There are no arbitrage opportunities.*
2. *A stochastic discount factor exists.*
3. *A risk-neutral probability measure exists.*

The fundamental theorem of asset pricing was first formulated by Ross [11] and it took 20 years to reach a very general version of it in general semimartingale models that are beyond the scope of our treatment here. The interested reader can check the monograph [3], where the history of the theorem and all its proofs are presented.

The Important Case of the Logarithm

The most well-studied case of utility on the real line is $U(x) = \log(x)$, both because of its computational simplicity and for the theoretical value that it has. Since the logarithmic function is only defined on the strictly positive real line, it does not completely fall in the aforementioned framework, but it is easy to see that the described theory is still valid.

Consider an economic agent with logarithmic utility that starts with initial capital $x = 1$. Call $X^* = X^{(1; \theta_*)}$ the optimal wealth corresponding to log-utility maximization. The fact that $U'(x) = 1/x$ allows to define a stochastic discount factor Y^* via $Y_0^* = 1$ and

$$Y_T^* = \frac{1}{X_T^* \mathbb{E}^{\mathbb{P}}[1/(\beta_T X_T^*)]} \quad (8)$$

From $\mathbb{E}^{\mathbb{P}}[Y_T^* X_T^*] = 1$, it follows that $\mathbb{E}^{\mathbb{P}}[1/(\beta_T X_T^*)] = 1$ and therefore $Y^* = 1/X^*$. This simple relationship between the log-optimal wealth and the stochastic discount factor that is induced by it is one of the keys to characterize the existence of stochastic discount factors in more complicated models and their relationship with absence of free lunches. It finds good use in the section Stochastic Discount Factors for Itô Processes for the case of models using Itô processes.

Arbitrage-free Prices

For a claim with random payoff H_T at time T , an *arbitrage-free* (AF) price H_0 is a price at time zero such that the extended market that consists of the original traded assets with asset prices S^i , $i = 0, \dots, d$, augmented by the new claim, remains AF. If the claim is *perfectly replicable*, that is, if there exists $x \in \mathbb{R}$ and $\theta \in \mathbb{R}^d$ such that $X_T^{(x; \theta)} = H_T$, it is easily seen that the *unique* AF price for the claim is x . However, it is frequently the case that a newly introduced claim is not perfectly replicable using the existing liquid assets. In that case, there exists more than one AF price for the claim; actually, the set of all the possible AF prices is $\{\mathbb{E}^{\mathbb{P}}[Y_T H_T] \mid Y \text{ is a stochastic discount factor}\}$. To see this, first pick a stochastic discount factor Y_T and set $H_0 = \mathbb{E}[Y_T H_T]$; then, Y remains a stochastic discount factor for the extended market, which therefore does not allow for any arbitrage opportunities. Conversely, if H_0 is an AF price for the new claim, we know from Theorem 1 that there exists a stochastic discount factor Y for the extended market, which satisfies $H_0 = \mathbb{E}[Y_T H_T]$ and is trivially a stochastic discount factor for the original market. The result we just mentioned justifies the appellation “Fundamental theorem of asset pricing” for Theorem 1.

Utility Indifference Pricing

Suppose that a new claim promising some random payoff at time T is issued. Depending on the claim’s present traded price, an economic agent might be inclined to take a long or short position—this will depend on whether the agent considers the market price low or high, respectively. There does exist a market price level of the claim that will make the agent indifferent to going long or short on an *infinitesimal*^b amount of asset. This price level is called *indifference price*. In the context of claim valuation, utility indifference prices have been introduced in [2];^c however, they had been widely used previously in the science of economics. Indifference prices depend on the particular agent’s views, preferences, as well as portfolio structure, and should not be confused with market prices, which are established using the forces of supply and demand.

Since the discussed preference structures are based on *expected utility*, it makes sense to try and understand quantitatively how *utility indifference prices* are

formed. Under the present setup, consider a claim with random payoff H_T at time T . The question we wish to answer is this: what is the indifference price H_0 of this claim today for an economic agent?

For the time being, let H_0 be any price set by the market for the claim. The agent will invest in the risky assets and will hold θ units of them, as well as the new claim, taking a position of ϵ units. Then, the agent's terminal payoff is

$$X_T^{(x;\theta,\epsilon)} := X_T^{(x;\theta)} + \epsilon \left(H_T - \frac{S_T^0}{S_0^0} H_0 \right) \quad (9)$$

The agent will again maximize expected utility, that is, will invest $(\theta_*, \epsilon_*) \in \mathbb{R}^d \times \mathbb{R}$ such that

$$\mathbb{E}^\mathbb{P} \left[U \left(X_T^{(x;\theta_*,\epsilon_*)} \right) \right] = \sup_{(\theta,\epsilon) \in \mathbb{R}^d \times \mathbb{R}} \mathbb{E}^\mathbb{P} \left[U \left(X_T^{(x;\theta,\epsilon)} \right) \right] \quad (10)$$

If H_0 is the agent's indifference price, it must follow that $\epsilon_* = 0$ in the above maximization problem; then, the agent's optimal decision regarding the claim would be not to buy or sell any units of the asset. In particular, the concave function $\mathbb{R} \ni \epsilon \mapsto \mathbb{E}^\mathbb{P} \left[U \left(X_T^{(x;\theta_*,\epsilon)} \right) \right]$ should achieve its maximum at $\epsilon = 0$. First-order conditions give that H_0 is the agent's indifference price if

$$\begin{aligned} 0 &= \frac{\partial}{\partial \epsilon} \Big|_{\epsilon=0} \mathbb{E}^\mathbb{P} \left[U \left(X_T^{(x;\theta_*,\epsilon)} \right) \right] \\ &= \mathbb{E}^\mathbb{P} \left[U' \left(X_T^{(x;\theta_*,0)} \right) \left(H_T - \frac{S_T^0}{S_0^0} H_0 \right) \right] \end{aligned} \quad (11)$$

A remark is in order before writing down the indifference-pricing formula. The strategy θ_* that has been appearing above represents the optimal holding in the liquid traded assets when *all* assets *and* the claim are available—it is *not*, in general, the agent's optimal asset holdings if the claim were *not* around. Nevertheless, *if* the solution of problem (10) is such that the optimal holdings in the claim are $\epsilon_* = 0$, *then* θ_* are also the agent's optimal asset holdings if there had been no claim to begin with. In other words, if $\epsilon_* = 0$, $X_T^{(x;\theta_*,0)}$ is exactly the same quantity $X_T^{(x;\theta_*)}$ that appears in equation (4). Remembering the definition of the stochastic discount factor Y_T of

equation (5), we can write

$$H_0 = \mathbb{E}^\mathbb{P} [Y_T H_T] \quad (12)$$

It is important to observe that Y_T depends on a number of factors, namely, the probability $\mathbb{P}$, the utility U , and the initial capital x , but *not* on the particular claim to be valued. Thus, we need only one evaluation of the stochastic discount factor and we can use it to find indifference prices with respect to all kinds of different claims.

State Price Densities

For a fixed $\omega \in \Omega$, consider an *Arrow–Debreu* security that pays off a unit of account at time T if the state of nature is ω , and pays off nothing, otherwise. The indifference price of this security for the economic agent is $p(\omega) := Y(\omega)P(\omega)$. Since Y appears as the density of the “state price” p with respect to the probability $\mathbb{P}$, stochastic discount factors are also termed *state price densities* in the literature. For two states of nature ω and ω' of Ω such that $Y(\omega) < Y(\omega')$, an agent who uses the stochastic discount factor Y would consider ω' a more unfavorable state than ω and would be inclined to pay more for insurance against adverse market movements.

Comparison with Real-world Valuation

Only for the purpose of what is presented here, assume that $S_0^0 = 1$ and $S_T^0 = 1 + rT$ for some $r \in \mathbb{R}_+$. Let Y be a stochastic discount factor; then, we have $1 = S_0^0 = \mathbb{E}^\mathbb{P} [Y_T S_T^0] = (1 + rT) \mathbb{E}^\mathbb{P} [Y_T]$. Pick any claim with random payoff H_T at time T and use $H_0 = \mathbb{E}^\mathbb{P} [Y_T H_T]$ to write

$$H_0 = \frac{1}{1 + rT} \mathbb{E}^\mathbb{P} [H_T] + \text{cov}^\mathbb{P} (Y_T, H_T) \quad (13)$$

where $\text{cov}^\mathbb{P}(\cdot, \cdot)$ is used to denote *covariance* of two random variables with respect to $\mathbb{P}$. The first term $(1 + rT)^{-1} \mathbb{E}^\mathbb{P} [H_T]$ of the above formula describes “real-world” valuation for an agent who would be neutral under his views $\mathbb{P}$ in facing the risk coming from the random payoff H_T . This risk-neutral attitude is usually absent: agents require compensation for the risk they undertake, or might even feel inclined to pay more for a security that will insure them in cases of unfavorable outcomes. This is exactly mirrored by the

correction factor $\text{cov}^{\mathbb{P}}(Y_T, H_T)$ appearing in equation (13). If the covariance of Y_T and H_T is negative, the claim tends to pay more when Y_T is low. By the discussion in the section State Price Densities, this means that the payoff will be high in states that are not greatly feared by the agent, who will therefore be inclined to pay *less* than what the real-world valuation gives. On the contrary, if the covariance of Y_T and H_T is positive, H_T will pay off higher in dangerous states of nature for the agent (where Y_T is also high), and the agent's indifference price will be *higher* than the real-world valuation.

Stochastic Discount Factors for Itô Processes

The Model

Uncertainty is modeled *via* a probability space $(\Omega, \mathcal{F}, \mathbf{F}, \mathbb{P})$, where $\mathbf{F} = (\mathcal{F}_t)_{t \in [0, T]}$ is a filtration representing the flow of information. The market consists of a *locally* riskless savings account whose price process S^0 satisfies $S^0_0 > 0$ and

$$\frac{dS^0_t}{S^0_t} = r_t dt, \quad t \in [0, T] \quad (14)$$

for some $\mathbf{F}$ -adapted, positive *short-rate* process $r = (r_t)_{t \in \mathbb{R}}$. It is obvious that $S^0_t = S^0_0 \exp(\int_0^t r_u du)$ for $t \in [0, T]$. We define the *deflator* β *via*

$$\beta_t = \frac{S^0_t}{S^0_0} = \exp\left(-\int_0^t r_u du\right), \quad t \in [0, T] \quad (15)$$

The movement of d risky assets will be modeled *via* Itô processes:

$$\frac{dS^i_t}{S^i_t} = b^i_t dt + \langle \sigma^i_t, dW_t \rangle, \quad t \in \mathbb{R}_+, \quad i = 1, \dots, d \quad (16)$$

Here, $b = (b^1, \dots, b^d)$ is the $\mathbf{F}$ -adapted d -dimensional process of *appreciation rates*, $W = (W^1, \dots, W^m)$ is an m -dimensional $\mathbb{P}$ -Brownian motion representing the *sources of uncertainty* in the market, and $\langle \cdot, \cdot \rangle$ denotes the usual inner product notation: $\langle \sigma^i_t, dW_t \rangle = \sum_{j=1}^m \sigma^{ji}_t dW^j_t$ where $(\sigma^{ji})_{1 \leq j \leq m, 1 \leq i \leq d}$ is the $\mathbf{F}$ -adapted $(m \times d)$ -matrix-valued process whose entry σ^{ji}_t represents the impact

of the j th source of uncertainty on the i th asset at time $t \in [0, T]$. With “ $\top$ ” denoting transposition, $c := \sigma^\top \sigma$ is the $d \times d$ local *covariation matrix*. To avoid degeneracies in the market, it is required that c_t has full rank for all $t \in [0, T]$, $\mathbb{P}$ almost surely (a.s.). This implies, in particular, that $d \leq m$ —there are more sources of uncertainty in the market than are liquid assets to hedge away the uncertainty risk. Models of this sort are classical in the quantitative finance literature—see, for example, [8].

Definition 2 *A risk premium is any m -dimensional, $\mathbf{F}$ -adapted process λ satisfying $\sigma^\top \lambda = b - r\mathbf{1}$, where $\mathbf{1}$ is the d -dimensional vector with all unit entries.*

The terminology “risk premium” is better explained for the case $d = m = 1$; then $\lambda = (b - r)/\sigma$ is the premium over the risk-free rate that investors require per unit of risk associated with the (only) source of uncertainty. In the general case, λ^j can be interpreted as the premium required for the risk associated with the j th source of uncertainty, represented by the Brownian motion W^j . In *incomplete* markets, when $d < m$, Proposition 1 shows all the different choices for λ . Each choice will parameterize the different risk attitudes of different investors. In other words, risk premia characterize the possible stochastic discount factors, as is revealed in Theorem 3.

If $m = d$, the equation $\sigma^\top \lambda = b - r\mathbf{1}$ has only one solution: $\lambda^* = \sigma c^{-1}(b - r\mathbf{1})$. If $d < m$ there are many solutions, but they can be characterized using easy linear algebra.

Proposition 1 *The risk premia are exactly all processes of the form $\lambda = \lambda^* + \kappa$, where $\lambda^* := \sigma c^{-1}(b - r\mathbf{1})$ and κ is any adapted process with $\sigma^\top \kappa = 0$.*

If $\lambda = \lambda^* + \kappa$ in the notation of Proposition 1, then $\langle \lambda^*, \kappa \rangle = (b - r\mathbf{1})^\top c^{-1} \sigma^\top \kappa = 0$. Then, $|\lambda|^2 = |\lambda^*|^2 + |\kappa|^2$, where $|\lambda^*|^2 = \langle b - r\mathbf{1}, c^{-1}(b - r\mathbf{1}) \rangle$.

Stochastic Discount Factors

The usual method of obtaining stochastic discount factors in continuous time is through risk-neutral measures. The fundamental theorem of asset pricing in the present Itô-process setting states that absence of free lunches with vanishing risk^d is equivalent to the existence of a probability $\mathbb{Q} \sim \mathbb{P}$ such that βS^i is (only) a *local* $\mathbb{Q}$ -martingale for all $i = 0, \dots, d$. (For

the definition of local martingales, check, e.g., [7].) In that case, by defining Y via $Y_t = \beta_t(d\mathbb{Q}/d\mathbb{P})|_{\mathcal{F}_t}$, YS^i is a local $\mathbb{P}$ -martingale for all $i = 0, \dots, d$. The last property is taken here as the *definition* of a stochastic discount factor.

Definition 3 Consider the above Itô-process setup, a stochastic process Y is called a stochastic discount factor if

- $Y_0 = 1$ and $Y_T > 0$, $\mathbb{P}$ a.s.
- YS^i is a local $\mathbb{P}$ -martingale for all $i = 0, 1, \dots, d$.

In the case where YS^0 is an actual martingale, that is, $\mathbb{E}^{\mathbb{P}}[Y_T S_T^0] = S_0^0$, a risk-neutral measure $\mathbb{Q}$ is readily defined via the recipe $d\mathbb{Q} = (Y_T S_T^0 / S_0^0) d\mathbb{P}$. However, this is not always the case, as Example 1 below will show. Therefore, existence of a stochastic discount factor is a weaker notion than existence of a risk-neutral measure. For some practical applications though, these differences are unimportant. There is further discussion of this point later in the section Stochastic Discount Factors and Equivalent Martingale Measures.

Example 1 Let $S^0 \equiv 1$ and S^1 be a three-dimensional Bessel process with $S_0^1 = 1$. If $\mathbf{F}$ is the natural filtration of S^1 , it can be shown that the *only* stochastic discount factor is $Y = 1/S^1$, which is a *strict local martingale* in the terminology of [4].

Credit Constraints on Investment

In view of the theoretical possibility of continuous trading, to avoid so-called *doubling strategies* (and for the fundamental theorem of asset pricing to hold), *credit constraints* have to be introduced. The wealth of agents has to be bounded from below by some constant, representing the credit limit. Shifting the wealth appropriately, one can assume that the credit limit is set to zero; therefore, only positive wealth processes are allowed in the market.

Since only strictly positive processes are considered, it is more convenient to work with *proportions* of investment, rather than absolute quantities as was the case in the section Stochastic Discount Factors in Discrete Probability Spaces. Pick some $\mathbf{F}$ -adapted process $\pi = (\pi^1, \dots, \pi^d)$. For $i = 1, \dots, d$ and $t \in [0, T]$, the number π_t^i represents the *percentage* of

capital in hand invested in asset i at time t . In that case, $\pi^0 = 1 - \sum_{i=1}^d \pi^i$ will be invested in the savings account. Denote by X^π the wealth generated by starting from unit initial capital ($X_0^\pi = 1$) and invest according to π . Then,

$$\frac{dX_t^\pi}{X_t^\pi} = \sum_{i=0}^d \pi_t^i \frac{dS_t^i}{S_t^i} = (r_t + \langle \pi_t, b_t - r_t \mathbf{1} \rangle) dt + \langle \sigma_t \pi_t, dW_t \rangle \quad (17)$$

To ensure that the above wealth process is well defined, we must assume that

$$\begin{aligned} \int_0^T |\langle \pi_t, b_t - r_t \mathbf{1} \rangle| dt < +\infty \text{ and} \\ \int_0^T \langle \pi_t, c_t \pi_t \rangle dt < +\infty, \quad \mathbb{P} \text{ a.s.} \end{aligned} \quad (18)$$

The set of all d -dimensional, $\mathbf{F}$ -adapted processes π that satisfy equation (18) is denoted by Π . A simple use of the integration-by-parts formula gives the following result:

Proposition 2 If Y is a stochastic discount factor, then YX^π is a local martingale for all $\pi \in \Pi$.

Connection with “No Free Lunch” Notions

The next line of business is to obtain an *existential* result about stochastic discount factors in the present setting, also connecting their existence to an NA-type notion. Remember, from the section The Important Case of the Logarithm, the special stochastic discount factor that is the reciprocal of the log-optimal wealth process. We proceed somewhat heuristically to compute the analogous processes for the Itô-process model. The linear stochastic differential equation (17) has the following solution, expressed in logarithmic terms:

$$\begin{aligned} \log X^\pi = \int_0^\cdot \left(r_t + \langle \pi_t, b_t - r_t \mathbf{1} \rangle - \frac{1}{2} \langle \pi_t, c_t \pi_t \rangle \right) dt \\ + \int_0^\cdot \langle \sigma_t \pi_t, dW_t \rangle \end{aligned} \quad (19)$$

Assuming that the local martingale term $\int_0^\cdot \langle \sigma_t \pi_t, dW_t \rangle$ in equation (19) is an actual martingale, the aim is to maximize the expectation of the drift

term. Notice that we can actually maximize the drift pathwise if we choose the portfolio $\pi_* = c^{-1}(b - r\mathbf{1})$. We need to ensure that π_* is in Π . It is easy to see that the equations in (18) are both satisfied if and only if $\int_0^T |\lambda_t^*|^2 dt < \infty$ $\mathbb{P}$ a.s., where $\lambda^* := \sigma c^{-1}(b - r\mathbf{1})$ is the special risk premium of Proposition 1. Under this assumption, $\pi_* \in \Pi$. Call $X^* = X^{\pi_*}$ and define

$$\begin{aligned} Y^* &:= \frac{1}{X^*} \\ &= \beta \exp \left(- \int_0^\cdot \langle \lambda_t^*, dW_t \rangle - \frac{1}{2} \int_0^\cdot |\lambda_t^*|^2 dt \right) \end{aligned} \quad (20)$$

Using the integration-by-parts formula, it is rather straightforward to check that Y^* is a stochastic discount factor. In fact, the ability to define Y^* is the way to establish that a stochastic discount factor exists, as the next result shows.

Theorem 2 *For the Itô process-model considered above, the following are equivalent.*

1. *The set of stochastic discount factors is nonempty.*
2. *$\int_0^T |\lambda_t^*|^2 dt < \infty$ $\mathbb{P}$ -a.s.; in that case, Y^* defined in equation (20) is a stochastic discount factor.*
3. *For any $\epsilon > 0$, there exists $\ell = \ell(\epsilon) \in \mathbb{R}_+$ such that $\mathbb{P}[X_T^\pi > \ell] < \epsilon$ uniformly over all portfolios $\pi \in \Pi$.*

The interest reader is referred to [6], where the property of the market described in statement 3 of the above theorem is termed *No Unbounded Profit with Bounded Risk*.

The next *structural* result about the stochastic discount factors in the Itô-process setting reveals the importance of Y^* as a building block.

Theorem 3 *Assume that $\mathbf{F}$ is the filtration generated by the Brownian motion W . Then, any stochastic discount factor Y in the previous Itô-process model can be decomposed as $Y = Y^* N^k$, where Y^* was defined in equation (20) and*

$$\begin{aligned} N_t^k &= \exp \left(- \int_0^t \langle \kappa_u, dW_u \rangle - \int_0^t |\kappa_u|^2 du \right), \\ &\forall t \in [0, T] \end{aligned} \quad (21)$$

where κ is an m -dimensional $\mathbf{F}$ -adapted process with $\sigma^\top \kappa = 0$.

If the assumption that $\mathbf{F}$ is generated by W is removed, one still obtains a similar result with N^k being replaced by *any* positive $\mathbf{F}$ -martingale N with $N_0 = 1$ that is *strongly* orthogonal to W . The specific representation obtained in Theorem 3 comes from the *martingale representation theorem* of Brownian filtrations; see, for example, [7].

Stochastic Discount Factors and Equivalent Martingale Measures

Consider an agent who uses a stochastic discount factor Y for valuation purposes. There is a possibility that YS^i could be a *strict local* $\mathbb{P}$ -martingale for some $i = 0, \dots, d$, which would mean that $S_0^i > \mathbb{E}^\mathbb{P}[Y_T S_T^i]$. The last inequality is puzzling in the sense that the agent's indifference price for the i th asset, which is $\mathbb{E}^\mathbb{P}[Y_T S_T^i]$, is *strictly* lower than the market price S_0^i . In such a case, the agent would be expected to wish to short some units of the i th asset. This is indeed what is happening; however, because of credit constraints, this strategy is infeasible. The following is a convincing example that establishes this fact. Before presenting the example, an important issue should be clarified. One would rush to state that such “inconsistencies” are tied to the notion of a stochastic discount factor as it appears in Definition 3, and that is *strictly* weaker than existence of a probability $\mathbb{Q} \sim \mathbb{P}$ that makes all discounted processes βS^i *local* $\mathbb{Q}$ -martingales for $i = 0, \dots, d$. Even if such a probability *did* exist, βS^i could be a *strict local* $\mathbb{Q}$ -martingale for some $i = 1, \dots, d$; in that case, $S_0^i > \mathbb{E}^\mathbb{Q}[\beta_T S_T^i]$ and the same mispricing problem pertains.

Example 2 Let $S^0 \equiv 1$, S^1 be the *reciprocal* of a three-dimensional Bessel process starting at $S_0^1 = 1$ under $\mathbb{P}$ and $\mathbf{F}$ be the filtration generated by S^1 . Here, $\mathbb{P}$ is the *only* equivalent local martingale measure and $1 = S_0^1 > \mathbb{E}^\mathbb{P}[S_T^1]$ for all $T > 0$. This is a *complete* market—an agent can start with capital $\mathbb{E}^\mathbb{P}[S_T^1]$ and invest in a way so that at time T the wealth generated is exactly S_T . Naturally, the agent would like to long as much as possible from this replicating portfolio and go as short as possible from the actual asset. However, in doing so, the possible downside risk is infinite throughout the life of the investment and the enforced credit constraints will disallow for such strategies.

In the context of Example 2, the law of one price fails, since the asset that provides payoff S_T^1 at time T has a market price S_0^1 and a replication price $\mathbb{E}^\mathbb{P}[S_T^1] < S_0^1$. Therefore, if the law of one price is to be valid in the market, one has to insist on existence of an equivalent (*true*) martingale measure $\mathbb{Q}$, where each discounted process βS^i is a *true* (and not only local) $\mathbb{Q}$ -martingale for all $i = 0, \dots, d$. For pricing purposes then, it makes sense to ask that the stochastic discount factor Y^κ that is chosen according to Theorem 3 is such that $Y^\kappa S^i$ is a true $\mathbb{P}$ -martingale for all $i = 0, \dots, d$. Such stochastic discount factors give rise to probabilities $\mathbb{Q}^\kappa$ that make all deflated asset-price-process $\mathbb{Q}^\kappa$ -martingales and can be used as pricing measures.

Let us now specialize to the important “diffusion” case where $r_t = r \in \mathbb{R}$ for all $t \in [0, T]$ and $\sigma_t = \eta(t, S_t)$ for all $t \in [0, T]$, where η is a nice function with values in the space of $(m \times d)$ -matrices. As long as a claim written only on the traded assets is concerned, the choice of $\mathbb{Q}^\kappa$ for pricing is irrelevant, since the asset prices under $\mathbb{Q}^\kappa$ have dynamics

$$\frac{dS_t^i}{S_t^i} = r_t dt + \langle \sigma_t^i, dW_t^\kappa \rangle, \quad \forall t \in [0, T], \quad i = 1, \dots, d \quad (22)$$

where W^κ is a $\mathbb{Q}^\kappa$ -Brownian motion. However, if one is interested in pricing a claim written on a nontraded asset whose price process Z has $\mathbb{P}$ -dynamics

$$dZ_t = a_t dt + \langle f_t, dW_t \rangle, \quad t \in [0, T] \quad (23)$$

for $\mathbf{F}$ -adapted a and $f = (f^1, \dots, f^m)$, then the $\mathbb{Q}^\kappa$ -dynamics of Z are

$$dZ_t = (a_t - \langle f_t, \lambda_t^* \rangle - \langle f_t, \kappa_t \rangle) dt + \langle f_t, dW_t^\kappa \rangle, \quad \forall t \in [0, T] \quad (24)$$

The dynamics of Z will be independent of the choice of κ only if the volatility structure of the process Z , given by f , is in the range of $\sigma^\top$. This will mean that $\langle f, \kappa \rangle = 0$ for all κ such that $\sigma^\top \kappa = 0$ and that Z is perfectly replicable using the traded assets. As long as there is any randomness in the movement in Z that cannot be captured by investing in the traded assets, that is, if there exists some κ with $\sigma^\top \kappa = 0$ and $\langle f, \kappa \rangle$ not being identically zero, perfect replicability fails and pricing becomes a more complicated

issue, depending on the preferences of the particular agent as given by the choice of κ to form the stochastic discount factor.

End Notes

^aOne can impose natural conditions on preference relations defined on the set of all possible outcomes that will lead to numerical representation of the preference relationship *via* expected utility maximization. This was axiomatized in [10]—see also Chapter 2 of [5] for a nice exposition.

^bWe stress “infinitesimal” because when the portfolio holdings of the agent change, the indifference prices also change; thus, for large sales or buys that will considerably change the portfolio structure, there might appear an incentive, that was not there before, to sell or buy the asset.

^cFor this reason, utility indifference prices are sometimes referred to as *Davis prices*.

^dFree lunches with vanishing risk is the suitable generalization of the notion of arbitrages to get a version of the fundamental theorem of asset pricing in continuous time. The reader is referred to [3].

^eThe inequality follows because positive local martingales are supermartingales—see, for example, [7].

References

- [1] Cochrane, J.H. (2001). *Asset Pricing*, Princeton University Press.
- [2] Davis, M.H.A. (1997). Option pricing in incomplete markets, in *Mathematics of Derivative Securities (Cambridge, 1995)*, Publications of the Newton Institute, Cambridge University Press, Cambridge, Vol. 15, pp. 216–226.
- [3] Delbaen, F. & Schachermayer, W. (2006). *The Mathematics of Arbitrage*, Springer Finance, Springer-Verlag, Berlin.
- [4] Elworthy, K.D. & Li, X.-M. & Yor, M. (1999). The importance of strictly local martingales; applications to radial Ornstein-Uhlenbeck processes, *Probability Theory and Related Fields* **115**, 325–355.
- [5] Föllmer, H. & Schied, A. (2004). *Stochastic Finance*, extended Edition, de Gruyter Studies in Mathematics, Walter de Gruyter & Co., Berlin, Vol. 27.
- [6] Karatzas, I. & Kardaras, C. (2007). The numéraire portfolio in semimartingale financial models, *Finance and Stochastics* **11**, 447–493.
- [7] Karatzas, I. & Shreve, S.E. (1991). *Brownian Motion and Stochastic Calculus*, 2nd Edition, Graduate Texts in Mathematics, Springer-Verlag, New York, Vol. 113.
- [8] Karatzas, I. & Shreve, S.E. (1998). *Methods of Mathematical Finance*, Applications of Mathematics (New York), Springer-Verlag, New York, Vol. 39.

-
- [9] Lamberton, D. & Lapeyre, B. (1996). *Introduction to Stochastic Calculus Applied to Finance*, Chapman & Hall, London. Translated from the 1991 French original by Nicolas Rabeau and François Mantion.
 - [10] von Neumann, J. & Morgenstern, O. (2007). *Theory of Games and Economic Behavior*, anniversary edition, Princeton University Press, Princeton, NJ. With an introduction by Harold W. Kuhn and an afterword by Ariel Rubinstein.
 - [11] Ross, S.A. (1976). The arbitrage theory of capital asset pricing, *Journal of Economic Theory* **13**, 341–360.

Related Articles

Arrow–Debreu Prices; Change of Numeraire; Complete Markets; Equivalent Martingale Measures; Fundamental Theorem of Asset Pricing; Pricing Kernels.

CONSTANTINOS KARDARAS

Utility Function

Behavior and Preferences

Modern utility theory studies preference orderings over choice sets and their numerical representations. Consider a decision maker (DM) who has to choose among a set X of alternatives. The set X is called the *DM's choice set*. In the deterministic case, which is our focus here, alternatives are certain, without any uncertainty. For example, in consumer theory, the DM is a consumer and X is the consumption set that he/she faces, that is, a subset of $\mathbb{R}^n$ whose elements $x = (x_1, \dots, x_n)$ represent the consumption bundles available to consumers. In intertemporal choice problems, X is a subset of $\mathbb{R}^\infty$, the space of sequences $\{x_t\}_{t=1}^\infty$, where x_t is the DM's outcome at time t . Alternatives become more complicated objects under uncertainty, such as random variables in one-period problems and stochastic processes in intertemporal problems. This more general case is not considered here.

DMs have some preferences over the elements of X ; they may like some alternatives more than others or may be indifferent among some of them. For example, in consumer theory consumers will rank consumption bundles in their consumption sets according to their tastes.

This motivates the introduction of preference orderings $\succsim$ defined on the choice set X . The ordering $\succsim$ has the following interpretation: for any two vectors x and y in X , we write $x \succsim y$ if the DM either strictly prefers x to y or is indifferent between the two.

The ordering $\succsim$ is the basic primitive of the theory. The following two relations are derived from $\succsim$:

1. for any two vectors x and y in X , we write $x \succ y$ if the DM strictly prefers x to y . Formally, $x \succ y$ if $x \succsim y$, but not $y \succsim x$;
2. for any two vectors x and y in X , we write $x \sim y$ if the DM is indifferent between x and y . Formally, $x \sim y$ if both $x \succsim y$ and $y \succsim x$.

On the preference ordering $\succsim$, which is the theory's "raw material," some properties are considered.

Axiom 1 (Transitivity). *For any three elements x, y , and z in X , if $x \succsim y$ and $y \succsim z$, then $x \succsim z$.*

Transitivity is a rationality assumption. Its violation generates cycles, for example, $x \succsim y \succsim z \succ x$. The most troublesome consequence of such cycles is that there might not exist a best element in the choice set X . For example, suppose that $X = \{x, y, z\}$ and that $x \succ y$ and $y \succ z$. If transitivity is violated, we get the cycle $x \succ y \succ z \succ x$ and there is no best element in X .

Axiom 2 (Completeness). *For any two elements x and y in X , $x \succsim y$, $y \succsim x$, or both.*

This is a simple, but not innocuous, property. A DM's preference $\succsim$ satisfies this property if, when faced with any two alternatives in X , he/she can always say which one he/she prefers. As alternatives may be very different, this might be a strong requirement (see [1, 6], for weakenings of this assumption).

Note that Axiom 2 implies reflexivity, that is, $x \succsim x$ for all $x \in X$. When $\succsim$ is reflexive and transitive (e.g., when it satisfies Axioms 1 and 2), following the consumer theory terminology, we call indifference curves the equivalence classes $[x] = \{y \in X : y \sim x\}$ for any $x \in X$. We denote the collection $\{[x] : x \in X\}$ of all indifference curves by $X/\sim$, which is a partition of X . That is, each $x \in X$ belongs to one, and only one, indifference curve.

Axioms 1 and 2 do not depend on any particular structure of the set X . In most applications, however, X is a subset of an ordered vector space $(V, \geq)$, that is, of a space V that has both a vector and an order structure. The space $\mathbb{R}^n$ endowed with the natural order $\geq$ is an important example of an ordered vector space. Given any $x, y \in \mathbb{R}^n$, when the vectors x and y are regarded as consumption bundles, $x \geq y$ means that the bundle x has at least as much of each good than the bundle y , while the convex combination $\alpha x + (1 - \alpha)y$ is interpreted as a mix of the two vectors (implicitly we are assuming that goods are suitably divisible).

The following axioms are based on the order and vector structures of X . For simplicity, we assume $X \subseteq \mathbb{R}^n$, though most of what follows holds in more general ordered vector spaces with units. Here, $x \succ y$ means $x \geq y$ and $x \neq y$ (i.e., $x_i > y_i$ for at least some $i = 1, \dots, n$).

Axiom 3 (Monotonicity). *For any two elements x and y in $X \subseteq \mathbb{R}^n$, if $x \succ y$, then $x \succ y$.*

This axiom connects the order $\geq$ on X and the DM's preference relation $\succsim$. In the context of consumer theory, it says that “the more, the better.” In particular, given two vectors x and y with $x \geq y$, it is enough that x has strictly more of at least some good i to be strictly preferred to y . This means that all goods are “essential” that is, the DM pays attention to each of them. Moreover, observe that, by Axiom 3 and reflexivity, $x \geq y$ implies $x \succsim y$. This is because $x \geq y$ if either $x = y$ or $x > y$.

The following two axioms rely on the vector structure of X .

Axiom 4 (Archimedean). *Suppose that x , y , and z are any three elements of a convex $X \subseteq \mathbb{R}^n$ such that $x > y > z$. Then there exist $\alpha, \beta \in (0, 1)$ such that $\alpha x + (1 - \alpha)z > y > \beta x + (1 - \beta)z$.*

According to this axiom, there are no infinitely preferred or infinitely despised alternatives. That is, given any pairs $x > y$ and $y > z$, alternative x cannot be infinitely better than y , and alternative z cannot be infinitely worse than y . Indeed, we can always mix x and z to get better alternatives, that is, $\alpha x + (1 - \alpha)z$, or worse alternatives, that is, $\beta x + (1 - \beta)z$, than y .

It may be useful to remember the analogous property that holds for real numbers: if x , y , and z are real numbers with $x > y > z$, then there exist $\alpha, \beta \in (0, 1)$ such that $\alpha x + (1 - \alpha)z > y > \beta x + (1 - \beta)z$. This property does not hold any more if we consider ∞ and $-\infty$, that is, the extended real line $\mathbb{R} = [-\infty, \infty]$. Specifically, let $x = \infty$ or $z = -\infty$. In this case, x is infinitely larger than y , z is infinitely smaller than y , and there are no $\alpha, \beta \in (0, 1)$ that satisfy the previous inequality. In fact, $\alpha\infty = \infty$ and $\beta(-\infty) = -\infty$ for all $\alpha, \beta \in (0, 1)$.

Axiom 5 (Convexity). *Given any two elements x and y of a convex set $X \subseteq \mathbb{R}^n$, if $x \sim y$ then $\alpha x + (1 - \alpha)y \succsim x$ for all $\alpha \in [0, 1]$.*

This axiom captures a preference for mixing: given any two indifferent alternatives, the DM always prefers any of their combination to each of the original alternatives. This preference for mixing is often assumed in applications and is a convexity property of indifference curves,^a the modern counterpart of the classic assumption of diminishing marginal utility.

Summing up, we have introduced a few properties that are often assumed on the preference $\succsim$. All these axioms are behavioral, that is, they are expressed

in terms of choice behavior. In particular, their behavioral meaning is transparent and, with the exception of the Archimedean axiom, they are all behaviorally falsifiable by suitable choice patterns. For example, one can show that a DM does not satisfy the transitivity axiom by finding alternatives $x, y, z \in X$ over which his/her choices exhibit the cycle $x \succsim y \succsim z > x$. This choice pattern would be enough to reject the hypothesis that his/her preference over X is transitive.

The use of preference axioms that have a transparent behavioral interpretation and that are falsifiable through choice behavior is the main methodological tenet of modern utility theory, often called the *revealed preference methodology*. In fact, choice behavior data are regarded as the only observable data that economic theories can rely upon.

Another important methodological feature of modern utility theory is that it adopts a weak notion of rationality, which requires only the consistency of choices without any demand on their motives. For example, transitivity is viewed as a rationality requirement in this sense because its violations would entail inconsistent patterns of choices that no DM would consciously follow, regardless of his/her motivations (see [15], for a recent discussion of this methodological issue).

Paretian Utility Functions

Although the preference ordering $\succsim$ is the fundamental notion, for analytical convenience it is often of interest to find a numerical representation of $\succsim$. Such numerical representations are called *utility functions*; formally, a real-valued function $u : X \rightarrow \mathbb{R}$ is a (*Paretian*) *utility function* if, for any pair $x, y \in X$,

$$x \succsim y \quad \text{if and only if} \quad u(x) \geq u(y) \quad (1)$$

In particular, for the derived relations $>$ and $\sim$ it holds, respectively, $x > y$ if and only if $u(x) > u(y)$ and $x \sim y$ if and only if $u(x) = u(y)$. Indifference curves can thus be written in terms of utility functions as $[x] = \{y \in X : u(y) = u(x)\}$.

Utility functions are analytically very convenient, but do not have any intrinsic psychological meaning: what matters is that they numerically rank vectors in the same way as the preference ordering $\succsim$. This implies, *inter alia*, that every monotone

transformation of a utility function is still a utility function, that is, utility functions are invariant under monotone transformations. To see why this is the case, let $u(X) = \{u(x) : x \in X\} \subseteq \mathbb{R}$ be the range of u and $f : u(X) \rightarrow \mathbb{R}$ a (strictly) monotone function, that is, $t > s$ implies $f(t) > f(s)$ for any scalars $t, s \in u(X)$. Clearly, $x \succsim y$ if and only if $(f \circ u)(x) \geq (f \circ u)(y)$ for any pair $x, y \in X$, and this shows that the transformation $f \circ u$ is still a utility function.

Example 1 A classic utility function $u : \mathbb{R}_{++}^2 \rightarrow \mathbb{R}$ is the Cobb–Douglas utility function:

$$u(x, y) = x^a y^{1-a} \quad \text{with } 0 \leq a \leq 1 \quad (2)$$

Suppose a preference $\succsim$ is represented by a Cobb–Douglas utility function. Then, $\succsim$ is also represented by the following monotone transformations of u :

1. $\lg(u(x, y)) = \lg(x^a y^{1-a}) = a \lg x + (1-a) \lg y$;
2. $\sqrt{u(x, y)} = \sqrt{x^a y^{1-a}} = x^{\frac{a}{2}} y^{\frac{1-a}{2}}$; and
3. $u(x, y)^3 = x^{3a} y^{3(1-a)}$.

In view of this invariance under monotone transformations, the utility theory presented here is often called *ordinal utility theory*. Observe that in this ordinal theory, utility differences such as $u(x) - u(y)$ are of no interest. This is because inequalities such as $u(x) - u(y) \geq u(z) - u(w)$ have no meaning in this setup: given any such inequality, it is easy to come up with monotone transformations $f : \mathbb{R} \rightarrow \mathbb{R}$ such that $(f \circ u)(x) - (f \circ u)(y) < (f \circ u)(z) - (f \circ u)(w)$. An important consequence of this observation is that incremental ratios of utility functions defined on subsets of $\mathbb{R}^n$ have no interest, except for their sign. For example, the classic notion of decreasing marginal utility, which is based on properties of the partial derivatives $\partial u(x)/\partial x_k$, is thus meaningless in ordinal utility theory.

In applications, utility functions $u : X \rightarrow \mathbb{R}$ are often used in optimization problems

$$\max_{x \in C} u(x) \quad (3)$$

where C is a suitable subset of the choice set X , determined by possible constraints that limit the DM's choices. For example, in consumer theory, C

is given by the budget set

$$C = \left\{ x \in X : \sum_{i=1}^n p_i x_i \leq w \right\} \quad (4)$$

where w is the consumer's wealth and each p_i is the price per unit of good i .

It is immediately seen that the solutions of the optimization problem (3) are the same, regardless of what monotone transformation of u is selected to make calculations. On the other hand, all these monotone transformations represent the same preference $\succsim$ and the solutions reflect only the DM's basic preference $\succsim$, not the particular utility function used to represent $\succsim$. This further shows that $\succsim$ is the fundamental notion. The choice of which u to use, among all equivalent monotone transformations, is only a matter of analytical convenience (e.g., in the Cobb–Douglas case, it is often convenient to use the logarithmic version $a \lg x + (1-a) \lg y$).

The optimization problems (3), which play a key role in economics, also illustrate the analytical importance of utility functions. In fact, a numerical representation of preferences allows to use the powerful methods of optimization theory to find and characterize the solutions of problem (3), which would be otherwise impossible, if we were to only rely on the preference $\succsim$. In other words, though the study of the preference $\succsim$ is what gives ordinal utility theory its scientific status by making it a behaviorally founded and falsifiable theory, it is its numerical representation provided by utility functions that gives the theory its operational content.

Given the importance of utility functions, the main problem of ordinal utility theory is to establish conditions under which the preference ordering $\succsim$ admits a utility representation. This is not a simple problem. We first state an existence result for the special case when the collection $X/\sim$ of indifference curves is at most countable.

Theorem 1 A preference ordering $\succsim$ defined on a choice set X with $X/\sim$ at most countable satisfies Axioms 1 and 2 if and only if there exists a function $u : X \rightarrow \mathbb{R}$ such that equation (1) holds.

Proof 1 [12] page 14.

Matters are more complicated when the collection $X/\sim$ is uncountable. It is easy to come up with examples of preferences that satisfy Axioms 1 and 2 and

do not admit a utility representation (see Example 2). We refer to [2, 12, 18] for general representation theorems. Here we establish an existence result for the important special case of preferences defined on $\mathbb{R}^n$, based on [3]. It is closely related to Theorems 3.3 and 3.6 of Fishburn (1970). For brevity, we omit its proof.

Write $x \leq \infty$ (respectively, $x \geq -\infty$) when either $x \in \mathbb{R}^n$ or $x_i = \infty$ (respectively, $x_i = -\infty$) for each i . That is, $x \leq \infty$ or $x \geq -\infty$ means that either each x_i is finite or each x_i is infinite. A subset of $\mathbb{R}^n$ is a closed order interval if, given $-\infty \leq y < z \leq \infty$, it has the form $[y, z] = \{x \in \mathbb{R}^n : y \leq x \leq z\}$ and is an open order interval if it has the form $(y, z) = \{x \in \mathbb{R}^n : y_i < x_i < z_i \text{ for each } i\}$. The half-open order intervals $[y, z)$ and $(y, z]$ are similarly defined. For example, $[z, \infty) = \{x \in \mathbb{R}^n : x \geq z\}$, and so $[0, \infty) = \mathbb{R}_+^n$.

A function $u : X \rightarrow \mathbb{R}$ is *monotone* if $x > y$ implies $u(x) > u(y)$ and is *quasiconcave* if its upper sets $\{x : u(x) \geq t\}$ are convex for all $t \in \mathbb{R}$ [16]. Since $\{y : u(y) \geq u(x)\} = \{y : y \succeq x\}$, the quasi-concavity of u implies the convexity of the upper contour sets of indifference curves (cf. End Note a).

Theorem 2 *For a preference ordering $\succeq$ defined on a order interval $X \subseteq \mathbb{R}^n$, the following conditions are equivalent:*

1. $\succeq$ satisfies Axioms 1–4 and
2. *there exists a monotonic and continuous function $u : X \rightarrow \mathbb{R}$ such that equation (1) holds.*

Moreover, Axiom 5 holds if and only if u is quasiconcave.

Theorem 2 is an important result. Almost every work in economics contains a utility function, often defined on order intervals of $\mathbb{R}^n$ and assumed to be monotone and quasi-concave. Theorem 2 shows the behavioral conditions that underlie this key modeling assumption.

By Theorem 2, the convexity axiom 5 is equivalent to the quasi-concavity of the utility function u . This is a substantially weaker property than the concavity of u , which would require $u(\alpha x + (1 - \alpha)y) \geq \alpha u(x) + (1 - \alpha)u(y)$ for all $x, y \in X$ and all $\alpha \in [0, 1]$. For example, any increasing function $u : \mathbb{R} \rightarrow \mathbb{R}$ is automatically quasi-concave.

Since concave utility functions are often used in applications because of their remarkable properties in

optimization problems, a natural question is whether, among all monotone transformations $f \circ u$ of a quasi-concave utility function u , there exists a concave one and this would ensure the existence of a concave representation of a preference $\succeq$ that satisfies Axiom 5. This important question was first studied by de Finetti [11], who showed that there exist quasi-concave functions that do not have any concave monotone transformation. Hence, convex indifference curves are not necessarily determined by a concave utility function (the converse is obviously true) and quasiconcavity in Theorem 2 cannot be improved to concavity. *Inter alia*, the seminal paper of de Finetti started the study of quasi-concave functions, later substantially developed by Fenchel [8], which is arguably the most important generalization of concavity.

Finally, observe that the utility function in Theorem 2 is continuous even though none of the axioms involves any topological notion. This is a remarkable consequence of the order and vector structures that the axioms use.

We close with an example of a preference that does not admit a utility representation.

Example 2 Lexicographic preferences are a classic example of preference orderings that do not admit a utility representation. Set $X = \mathbb{R}^2$ and say that $x \succeq y$ if either $x_1 > y_1$ or $x_1 = y_1$ and $x_2 \geq y_2$. That is, the DM first looks at the first coordinate: if $x_1 > y_1$, then $x \succeq y$. However, if $x_1 = y_1$, then the DM turns his/her attention to the second coordinate: if $x_2 \geq y_2$, then $x \succeq y$. This is how dictionaries order words and this motivates the name of this particular ordering. Although they satisfy Axioms 1–3, it can be proved ([18], pages 24–25) that lexicographic preferences do not admit a utility representation (it is easy to check that they do not satisfy the Archimedean axiom).

Brief Historical Remarks

The early development of utility theory is surveyed in the two 1950 articles of George Stigler [24]. Here it is worth noting that originally utility functions were regarded as a primitive notion whose role was to quantify a Benthamian pain/pleasure calculus. In other words, utility functions were viewed as a measure or a quantification of an underlying physiological phenomenon. This view of utility theory is sometimes called *cardinalism* and utility

functions derived within this approach are called *cardinal utility functions*. A key feature of cardinalism is that utility differences and their ratios are meaningful notions that quantify differences in pain/pleasure that DMs experience among different quantities of the outcomes. In particular, marginal utilities measure the marginal pain/pleasure that results from choices and these played a central role in the early cardinal consumer theory.

However, the difficulty of any reliable scientific measurement of cardinal utility raised serious doubts on the scientific status of cardinalism. At the end of the nineteenth century Pareto revolutionized utility theory by showing that an ordinal approach, based on indifference curves as a primitive notion—unlike Edgeworth [7], who introduced them as level curves of an original cardinal utility function—was enough for consumer theory purposes [20]. In particular, Pareto showed that the classic consumer problem could be solved and characterized by replacing marginal utilities with marginal rates of substitutions along indifference curves. For example, the classic key assumption of diminishing marginal utilities is replaced by the convexity property (Axiom 5) of indifference curves (the latter is actually a stronger property, unless utility functions are separable).

Unlike cardinal utility functions, indifference curves and their properties can be empirically determined and tested. Pareto's insight thus represented a key methodological advance and his ordinal approach, later substantially extended by Hicks and Allen [17, 23], is today the mainstream version of consumer theory. More generally, Pareto's ordinal revolution paved the way to the modern use of preferences as the primitive notion of decision theory. In fact, the use of preferences is the natural conceptual development of Pareto's original insight of considering indifference curves as a primitive notion. The first appearance of preferences as primitive notions seems to be in [9, 13]. They earned their current central theoretical place in decision theory with the classic works [4, 9, 12].

The utility theory under certainty outlined here reached its maturity in the 1960s (see, e.g., [5]). Subsequent work on decision theory has been mainly concerned with choice under uncertainty, extending the scope of the seminal contributions [9, 10, 19, 21, 22]. We refer the reader to [14] for a thorough and updated introduction to these more recent advances.

End Notes

^aObserve that this convexity property of indifference curves is weaker than the convexity of their upper contour sets $\{y \in X : y \succsim x\}$.

References

- [1] Aumann, R. (1962). Utility theory without the completeness axiom, *Econometrica* **30**, 445–462.
- [2] Bridges, D.S. & Mehta, G.B. (1995). *Representations of Preference Orderings*, Springer-Verlag, Berlin.
- [3] Cerreia-Vioglio, S., Maccheroni, F., Marinacci, M. & Montrucchio, L. (2009). *Uncertainty Averse Preferences*, mimeo.
- [4] Debreu, G. (1959). *Theory of Value*, Yale University Press.
- [5] Debreu, G. (1964). Continuity properties of Paretian utility, *International Economic Review* **5**, 285–293.
- [6] Dubra, J., Maccheroni, F. & Ok, E.A. (2004). Expected utility theory without the completeness axiom, *Journal of Economic Theory* **115**, 118–133.
- [7] Edgeworth, F.Y. (1881). *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*, Kegan Paul, London.
- [8] Fenchel, W. (1953). *Convex Cones, Sets, and Functions*, Princeton University Press, Princeton.
- [9] de Finetti, B. (1931). Sul significato soggettivo della probabilità, *Fundamenta Mathematicae* **18**, 298–329.
- [10] de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives, *Annales de l'Institut Henri Poincaré* **7**, 1–68.
- [11] de Finetti, B. (1949). Sulle stratificazioni convesse, *Annali di Matematica Pura ed Applicata* **30**, 173–183.
- [12] Fishburn, P.C. (1970). *Utility Theory for Decision Making*, Wiley, New York.
- [13] Frisch, R. (1926). Sur un problème d'économie pure, *Norsk Matematisk Forenings Skrifter* **1**, 1–40.
- [14] Gilboa, I. (2009). *Theory of Decision under Uncertainty*, Cambridge University Press, Cambridge.
- [15] Gilboa I., Maccheroni, F., Marinacci, M. & Schmeidler, D. (2009). Objective and subjective rationality in a multiple priors model, *Econometrica*, forthcoming.
- [16] Greenberg, H.J. & Pierskalla, W.P. (1971). A review of quasi-convex functions, *Operations Research* **19**, 1553–1570.
- [17] Hicks, J.R. & Allen, R.G.D. (1934). A reconsideration of the theory of value I, II, *Economica* **1**, 52–76, 196–219.
- [18] Kreps, D.M. (1988). *Notes on the Theory of Choice*, Westview Press, London.
- [19] von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, 2nd Edition, Princeton University Press, Princeton.
- [20] Pareto, V. (1906). *Manuale di Economia Politica*, Società Editrice Libreria, Milano.

- [21] Ramsey, F.P. (1931). Truth and probability, in *Foundations of Mathematics and other Essays*, R.B. Braithwaite, ed., Routledge.
- [22] Savage, L.J. (1954). *The Foundations of Statistics*, Wiley, New York.
- [23] Slutsky, E. (1915). Sulla teoria del bilancio del consumatore, *Giornale degli Economisti* **51**, 1–26.
- [24] Stigler, G.J. (1950). The development of utility theory I, II, *Journal of Political Economy* **58**, 307–327, 373–396.

Related Articles

Expected Utility Maximization: Duality Methods; Expected Utility Maximization; Recursive Preferences; Risk Aversion; Utility Indifference Valuation; Utility Theory: Historical Perspectives.

MASSIMO MARINACCI

Recursive Preferences

The standard additive utility model defines time- t utility for a discrete-time consumption process $\{c_t; t = 1, \dots, T\}$ as

$$U_t = E_t \sum_{s=t}^T e^{-\beta(s-t)} u(c_s) = E_t \{u(c_t) + e^{-\beta} U_{t+1}\} \quad (1)$$

where E_t denotes the conditional expectation. The virtue of the model is its simplicity: only discounted probabilities and the function u determine preferences. However, the additive treatment of states and times precludes the model from distinguishing between aversion to variability in consumption across states and across time. In fact, the agent's preferences are entirely determined by preferences over deterministic consumption streams (see [23]). Furthermore, because agents care only about the distribution of future consumption, they do not care about the temporal resolution of uncertainty.

A more flexible preference model is obtained with Krep and Porteus [14] recursive specification (see also [11]):

$$U_t = F(c_t, E_t u(U_{t+1})), \quad U_T = v(c_T) \quad (2)$$

where the aggregator function F models intertemporal substitution and u the aversion to risk in next period's utility. The popular Epstein and Zin [12] model is the special case characterized by scale-invariant preferences (U_t homogeneous in (c_t, U_{t+1})) and $v(c) = c$ and constant elasticity of substitution. The stochastic differential utility (SDU) formulation

$$U_t = E_t \left(\int_{s=t}^T b(c_s, U_s) ds + \frac{1}{2} a(U_s) d[U, U]_s \right) \quad (3)$$

where $[\cdot, \cdot]$ denotes quadratic variation, which was obtained by Duffie and Epstein [8] as the continuous-time limit of recursive utility. Time-additive utility is the special case $b(c, U) = u(c) - \beta U$ and $a = 0$. Skiadas [22] shows that SDU includes the robust control formulations of Anderson *et al.* [1], Hansen *et al.* [13], and Maenhout [17]. It is straightforward to show that SDU also includes the continuous-time limit of Chew [3] and Dekel [7] preferences.

In this paper, we examine the generalized SDU model, given in differential form by equation (5).

This preference model was introduced by Lazrak and Quenez [15] to unify the recursive formulation of Duffie and Epstein [8] and multiple-prior formulation of Chen and Epstein [2]. Schroder and Skiadas [19] and Skiadas [23] (see also [20] for the case with jumps) show that the more flexible form of the aggregator allows preferences to depend on the source of risk (e.g., domestic *versus* foreign), as well as first-order risk aversion (which imposes a higher penalty for small levels of risk) in addition to the standard second-order risk aversion dependence in equation (3).^a

Relative to the time-additive model, the loss of tractability under generalized SDU is surprisingly small and mainly confined to the complete-markets setting. In the case of power utility, for example, once incompleteness or market constraints are imposed, the additive problem is no simpler to solve than a more general class of scale-invariant (homothetic) recursive utility. The tractability of the most popular additive utility models is obtained not from additivity but from the scale or translation invariance property. The second and third sections examine the recursive classes with these invariance properties and show that their solution essentially reduces to solving a single constrained backward stochastic differential equation.

After defining the preferences and markets in the second and third sections, the general solution to the optimal portfolio and consumption problem is presented in the fourth section. The solution is obtained by first characterizing the utility supergradient density (a generalization of marginal utility) and the state-price density. The state-price result is useful in other asset-pricing applications because it characterizes the set of pricing operators consistent with no arbitrage in a general market setting.^b The optimal consumption process is obtained by equating a supergradient density and state-price density (a generalized notion of equating marginal utility and prices). All results in this article are based on [18–20]; these references also develop more specialized and tractable formulations, based on quadratic modeling of risk aversion, and the last introduces jump risk (modeled by marked point processes).

All uncertainty is generated by d -dimensional standard Brownian motion B over the finite time horizon $[0, T]$, supported by a probability space $(\Omega, \mathcal{F}, P)$. All processes dealt with in this article are assumed to be progressively measurable with respect to the augmented filtration $\{\mathcal{F}_t : t \in [0, T]\}$ generated

by B . We define a *cash flow* as a process x such that $E \left[\int_0^T x_t^2 dt + x_T^2 \right] < \infty$. We interpret x_t as a time- t payment rate and x_T as a lump-sum terminal payment. The set of all cash flows is denoted $\mathcal{H}$, which we regard as a Hilbert space under the inner product

$$(x|y) = E \left[\int_0^T x_t y_t dt + x_T y_T \right], \quad x, y \in \mathcal{H} \quad (4)$$

The set of *consumption plans* is the convex cone $\mathcal{C} \subseteq \mathcal{H}$. Finally, we let $\mathcal{S}_p$, $p = 1, 2$, denote the set of cash flows, satisfying $E \left[\left(\text{ess sup}_{t \in [0, T]} |x_t|^p \right) \right] < \infty$. The qualification “almost surely” is omitted throughout. The coefficients of all the stochastic differential equations introduced will be assumed sufficiently integrable so that the equations are well defined.

Recursive Preferences

We define preferences in terms of a *utility aggregator*, $F : \Omega \times [0, T] \times \mathbb{R}^{2+d} \rightarrow \mathbb{R}$. For every consumption plan $c \in \mathcal{C}$, we assume that there is a unique solution (U, Σ) to the backward stochastic differential equation (BSDE):

$$\begin{aligned} dU_t &= -F(t, c_t, U_t, \Sigma_t) dt + \Sigma_t' dB_t, \\ U_T &= F(T, c_T) \end{aligned} \quad (5)$$

(terminal utility depends only on (ω, T, c_T)), and we define $U(c) = U$. Throughout we assume that F is differentiable, $F(\omega, t, \cdot)$ is concave, and that the range of $F_c(\omega, t, \cdot, U, \Sigma)$ is $(0, \infty)$. SDU is the special case $F(\omega, t, c, U, \Sigma) = b(\omega, t, c, U) + \frac{1}{2}a(\omega, t, U)\Sigma'\Sigma$, and standard additive utility corresponds to the linear aggregator $F(\omega, t, c, U, \Sigma) = u(t, c) - \beta(t)U$. A multiple-priors formulation of [2] is given by

$$dU_t = - \left(b(t, c_t, U_t) - \max_{\theta \in \Theta_t} \theta' \Sigma_t \right) dt + \Sigma_t' dB_t \quad (6)$$

for some function Θ from $\Omega \times [0, T]$ to the set of convex compact subsets of $\mathbb{R}^d$.

Markets and the Wealth Equation

The agent is endowed with initial financial wealth w_0 and an endowment process $e \in \mathcal{H}$. The dollar value of

the agent's position in m risky assets is represented by the process $\phi = (\phi^1, \dots, \phi^m)'$. The agent's financial wealth process (not including the present value of the future endowment), W , is defined in terms of the *wealth aggregator* $f : \Omega \times [0, T] \times \mathbb{R}^{m+1} \rightarrow \mathbb{R}$, which represents the instantaneous expected growth of the agent's portfolio. Cuoco and Cvitanic [4] propose a nonlinear wealth aggregator to model the price impact of a large investor or differential borrowing and lending rates. Trading and wealth constraints are modeled by requiring the vector (W_t, ϕ_t) to lie in a convex set $K \subseteq \mathbb{R}^{1+m}$ at all times. The returns diffusion matrix is an $\mathbb{R}^{d \times m}$ -valued process σ^R . The agent's plan (c, W, ϕ) is *feasible* if it satisfies the *budget equation*

$$\begin{aligned} dW_t &= (f(t, W_t, \phi_t) + e_t - c_t)dt + \phi_t' \sigma_t^{R'} dB_t, \\ W_0 &= w_0, \quad c_T = W_T + e_T, \quad (W_t, \phi_t) \in K \end{aligned} \quad (7)$$

and the integrability conditions $\int_0^t (|f(s, W_s, \phi_s)| + \phi_s' \sigma_s^{R'} \sigma_s^R \phi_s) ds < \infty$ and $(-W_t)^+ \in \mathcal{S}_2$ (the latter to rule out doubling-type strategies). A consumption plan c is feasible if it is part of a feasible plan (c, W, ϕ) .

Example 1 Linear Budget Equation. Suppose that a money-market security pays interest at a rate r_t , and the risky assets' instantaneous excess returns relative to r are $dR_t = \mu_t^R dt + \sigma_t^{R'} dB_t$. Then we get the standard case:

$$\begin{aligned} f(\omega, t, w, \alpha) &= r(\omega, t)w + \alpha' \mu^R(\omega, t), \\ (w, \alpha) &\in K \end{aligned} \quad (8)$$

Example 2 Different Borrowing and Lending Rates. Extending Example 1, if b is a strictly positive process and money-market lending and borrowing occur at the rates r_t and $r_t + b_t$, respectively, then

$$\begin{aligned} f(\omega, t, w, \alpha) &= r(\omega, t)w + \mu^R(\omega, t)' \alpha - b(\omega, t) \\ &\quad \times (\mathbf{1}' \alpha - w)^+, \quad (w, \alpha) \in K \end{aligned} \quad (9)$$

For a related analysis, see Appendix B of [6].

General Solution Method

The agent's problem is to choose an optimal consumption plan: a feasible c such that $U(c) \geq U(\tilde{c})$

for all other feasible consumption plans $\tilde{c}$. We first show that optimality of c is essentially equivalent to the utility supergradient density of U at c satisfying the conditions for a state-price density, and then characterize these density equations in terms of the utility and wealth aggregators defined above. The resulting first-order conditions satisfy a constrained forward-backward stochastic differential equation (FBSDE) system.

Given the feasible consumption plan c , the process $\pi \in \mathcal{H}$ is a *state-price density* at c if

$$\begin{aligned} (\pi|x) &\leq 0 \\ \text{for all } x \text{ such that } c+x &\text{ is a} \\ \text{feasible consumption plan} \end{aligned} \quad (10)$$

We can interpret $(\pi|x)$ as the net present value of the cash flow x , which must be nonpositive for any feasible (i.e., affordable) incremental cash flow.

The process $\pi \in \mathcal{H}$ is a *supergradient density* of U_0 at c if

$$\begin{aligned} U_0(c+x) &\leq U_0(c) + (\pi|x) \\ \text{for all } x \text{ such that } c+x &\in \mathcal{C} \end{aligned} \quad (11)$$

and a *utility gradient density* of U_0 at c if

$$\begin{aligned} (\pi|x) &= \lim_{\alpha \downarrow 0} \frac{U_0(c+\alpha x) - U_0(c)}{\alpha} \quad \text{for all } x \text{ such} \\ \text{that } c+\alpha x &\in \mathcal{C} \text{ for some } \alpha > 0 \end{aligned} \quad (12)$$

If π is a supergradient density of U_0 at c and the utility gradient of U_0 at c exists, then the utility gradient density is π .

The general optimality result follows.

Proposition 1 *Suppose that (c, W, ϕ) is a feasible plan. If $\pi \in \mathcal{H}$ is both a supergradient density of U_0 at c and a state-price density at c , then the plan (c, W, ϕ) is optimal. Conversely, if the plan (c, W, ϕ) is optimal and $\pi \in \mathcal{H}$ is a utility gradient density of U_0 at c , then π is a state-price density at c .*

To apply Proposition 1, we obtain the dynamics of the utility supergradient and state-price densities corresponding to the utility and market models, as discussed in the sections Recursive Preferences and Markets and the Wealth Equation. Both depend on the feasible reference plan (c, W, ϕ) and are expressed

in terms of the differential or superdifferential (in the absence of differentiability or in the presence of constraints) of the corresponding aggregator.

The *superdifferential* of $f(t, \cdot)$ at (ω, t, w, α) relative to the constraint set K is the set $\partial f(\omega, t, w, \alpha)$ of all pairs $(d_w, d_\phi) \in \mathbb{R}^{1+m}$ such that

$$\begin{aligned} f(\omega, t, \tilde{w}, \tilde{\alpha}) - f(\omega, t, w, \alpha) &\leq d_w(\tilde{w} - w) \\ + d'_\phi(\tilde{\alpha} - \alpha) &\quad \text{for all } (\tilde{w}, \tilde{\alpha}) \in K \end{aligned} \quad (13)$$

Sufficient conditions for a state-price density follow.^c

Proposition 2 *Suppose that (c, W, ϕ) is feasible and $\pi \in \mathcal{H}_{++}$ satisfies*

$$\frac{d\pi_t}{\pi_t} = -\zeta_t dt - \eta'_t dB_t, \quad (\zeta_t, \sigma_t^{R'} \eta_t) \in \partial f(t, W_t, \phi_t) \quad (14)$$

and $\pi W \in \mathcal{S}_1$. Then π is a state-price density at c .

The process η is often called the *market price of risk*, with η_t^i representing the time- t shadow incremental expected wealth return per unit additional exposure to dB_t^i . The drift term ζ represents the shadow incremental return per unit wealth. In the case of a linear budget equation (8) and $K = \mathbb{R}^{1+m}$ (no constraints but possibly incomplete markets), we obtain the standard result $\zeta_t = r_t$ and $\mu_t^R = \sigma_t^{R'} \eta_t$.

Example 3 Collateral Constraint. Suppose that there is a single risky asset ($m = 1$), and, as in Example 1, $f(\omega, t, w, \alpha) = r(\omega, t)w + \mu^R(\omega, t)\alpha$. We consider an agent who faces the collateral constraint:

$$K = \{(w, \alpha) \in \mathbb{R}^2 : w \geq \varrho|\alpha|\} \quad (15)$$

for some $\varrho \in (0, 1)$. Then condition $(\zeta, \sigma^{R'} \eta) \in \partial f(W, \phi)$ is equivalent to the following restrictions:

$$\begin{aligned} \delta_t &= \zeta_t - r_t \geq 0, \quad \varepsilon_t = \mu_t^R - \Theta_t \in [-\varrho\delta_t, \varrho\delta_t] \\ (\phi_t > 0 &\Rightarrow \varepsilon_t = \varrho\delta_t), \quad (\phi_t < 0 \Rightarrow \varepsilon_t = -\varrho\delta_t), \\ (W_t > \varrho|\phi_t| &\Rightarrow \delta_t = 0) \end{aligned} \quad (16)$$

Papers analyzing collateral constraints in a Brownian setting and additive utility include [5, 16].

Assuming differentiability of the utility aggregator F (nondifferentiability is accommodated by replacing

the differential with a superdifferential defined as for f), we now provide sufficient conditions for a utility supergradient density.

Proposition 3 *Suppose that $c \in \mathcal{C}$, (U, Σ) solves BSDE (5), $\pi \in \mathcal{H}_{++}$ satisfies*

$$\pi_t = \mathcal{E}_t F_c(t, c_t, U_t, \Sigma_t) \quad (17)$$

where

$$\begin{aligned} \frac{d\mathcal{E}_t}{\mathcal{E}_t} &= F_U(t, c_t, U_t, \Sigma_t)dt + \\ &F_\Sigma(t, c_t, U_t, \Sigma_t)'dB_t, \quad \mathcal{E}_0 = 1 \end{aligned} \quad (18)$$

and $\mathcal{E}U \in \mathcal{S}_1$. Then, π is a utility gradient density of U_0 at c .

The supergradient density expression (17) is consistent with the calculations of Skiadas [2], Duffie and Skiadas [9], Chen and Epstein [10], and El_Karoui *et al.* [21]. All these papers assume Lipschitz-growth conditions that are violated in our setting.

We now apply Proposition 1 to characterize the first-order conditions. A key role in the solution is played by the strictly positive process

$$\lambda_t = F_c(t, c_t, U_t, \Sigma_t) \quad (19)$$

computed at the optimum, which represents the derivative of time- t optimal utility with respect to time- t wealth (as in the familiar envelope result). We solve for μ^λ and σ^λ in the Ito expansion

$$\frac{d\lambda_t}{\lambda_t} = \mu_t^\lambda dt + \sigma_t^{\lambda'} dB_t \quad (20)$$

by applying Ito's lemma to the utility gradient density, $\pi_t = \mathcal{E}_t \lambda_t$, and matching coefficients with those of the state-price density in Proposition 2. Having solved for λ , invert equation (19) to express the consumption plan c as

$$c_t = \mathcal{I}(t, \lambda_t, U_t, \Sigma_t) \quad (21)$$

where the function $\mathcal{I} : \Omega \times [0, T] \times (0, \infty) \times \mathbb{R}^{d+1} \rightarrow \mathbb{R}$ is defined implicitly through the following equation:

$$F_c(t, \mathcal{I}(t, y, U, \Sigma), U, \Sigma) = y, \quad y \in (0, \infty) \quad (22)$$

Combining the dynamics of λ with the utility BSDE (5), the budget equation (7), and the state-pricing restriction of Proposition 2, we obtain the

optimality conditions in the form of a constrained FBSDE system:

$$\begin{aligned} dU &= -F(\mathcal{I}(\lambda, U, \Sigma), U, \Sigma)dt + \Sigma' dB, \\ U_T &= F(T, W_T + e_T) \\ \frac{d\lambda_t}{\lambda_t} &= -(\zeta + F_U + \sigma^{\lambda'} F_\Sigma)dt + \sigma^{\lambda'} dB, \\ \lambda_T &= F_c(T, W_T + e_T) \\ dW &= (f(W, \phi) + e - \mathcal{I}(\lambda, U, \Sigma))dt + \phi' \sigma_t^{R'} dB_t, \\ W_0 &= w_0 \\ (\zeta, -\sigma^{R'}(F_\Sigma + \sigma^\lambda)) &\in \partial f(W, \phi), \\ (\phi, W) &\in K \end{aligned} \quad (23)$$

Given a solution $(U, \Sigma, \lambda, \sigma^\lambda, W, \phi)$ and suitable integrability assumptions (to satisfy Propositions 1–3), then c in equation (21) defines an optimal consumption plan.

Scale and Translation-invariant Solutions

The first-order conditions significantly simplify when utility and wealth dynamics fall into either the scale or translation-invariant classes. The scale-invariant, or homothetic, class exhibits homogeneity of degree one in consumption (when in certainty equivalent form) and includes, as special cases, homothetic Duffie–Epstein utility and additive power and log utility. The translation-invariant class exhibits quasilinearity with respect to a reference consumption stream and generalizes additive exponential utility. In both cases, the FBSDE of the first-order conditions uncouples into a single pure backward equation for λ and a pure forward equation for wealth.

Scale-invariant Class

We assume that consumption is strictly positive, and the aggregator $F(\omega, t, \cdot)$ is homogeneous of degree one, allowing the representation

$$\begin{aligned} F(\omega, t, c, U, \Sigma) &= UG\left(\omega, t, \frac{c}{U}, \frac{\Sigma}{U}\right), \\ F(T, c) &= c \end{aligned} \quad (24)$$

It is easy to confirm that utility is therefore homogeneous of degree one in consumption:

$$U(kc) = kU(c) \quad \text{for all } k \in \mathbb{R}_+ \text{ and } c \in \mathcal{C} \quad (25)$$

Defining $\sigma_t^U = \Sigma_t/U_t$, the BSDE (5) is equivalent to

$$\frac{dU_t}{U_t} = -G(t, c_t/U_t, \sigma_t^U)dt + \sigma_t^{U'}dB_t, \quad U_T = c_T \quad (26)$$

Example 4 Schroder and Skiadas [19] show that the quasi-quadratic aggregator

$$G(\omega, t, x, \sigma) = g(\omega, t, x) - \frac{1}{2}\Sigma'Q(\omega, t)\Sigma \quad (27)$$

where Q is positive definite for all (ω, t) , is particularly tractable, while allowing the modeling of source-dependent second-order risk aversion through Q . The continuous-time version of Epstein and Zin [11] is the special case with $Q = \gamma I$, for some constant $\gamma > 0$, and

$$g(\omega, t, x) = \alpha + \beta \frac{x^{1-\eta} - 1}{1-\eta} \quad (28)$$

for constants α and $\eta, \beta > 0$. Additive utility corresponds to $\gamma = \eta$ (the coefficient of relative risk aversion is equal to the inverse of the elasticity of intertemporal substitution).

On the markets side, we assume that wealth is strictly positive; the endowment process, e , is zero; constraints are on investment proportions, $\phi_t/W_t \in K^1$, for some convex set $K^1 \subseteq \mathbb{R}^m$; and the wealth aggregator $f(\omega, t, \cdot)$ is homogeneous of degree one. Letting $\psi_t = \phi_t/W_t$ denote the vector of investment proportions, and defining $f^1(t, \cdot) = f(t, 1, \cdot)$, the budget equation (7) becomes

$$\begin{aligned} \frac{dW_t}{W_t} &= \left(f^1(t, \psi_t) - \frac{c_t}{W_t} \right) dt + \psi_t' \sigma_t^{R'} dB_t, \\ W_0 &= w_0, \quad c_T = W_T, \quad \psi_t \in K^1 \end{aligned} \quad (29)$$

(in the linear budget constraint case of Example 1, we have $f^1(t, \psi_t) = r_t + \psi_t' \mu_t^R$).

The state-price density condition $(\zeta_t, \sigma_t^{R'} \eta_t) \in \partial f(t, W_t, \phi_t)$ is then equivalent to

$$\zeta_t = f^1(t, \psi_t) - \psi_t' \sigma_t^{R'} \eta_t \quad \text{and} \quad \sigma_t^{R'} \eta_t \in \partial f^1(t, \psi_t) \quad (30)$$

The scale-invariance properties imply that at the optimum utility is proportional to wealth:

$$U_t = \lambda_t W_t \quad (31)$$

and therefore $\sigma_t^U = \sigma_t^\lambda + \sigma_t^R \psi_t$. Recalling $\lambda_t = G_c(t, c_t/U_t, \sigma_t^U)$, we define the inverse function $\mathcal{I}^G(\cdot)$ analogously to equation (22) to obtain $c_t/U_t = \mathcal{I}^G(t, \lambda_t, \sigma_t^U)$. Defining the dual function of G^*

$$G^*(t, \lambda, \sigma) = G(t, \mathcal{I}^G(t, \lambda, \sigma), \sigma) - \mathcal{I}^G(t, \lambda, \sigma)\lambda \quad (32)$$

we obtain the first-order conditions (necessary and sufficient) as a constrained backward equation for λ (independent of wealth):

$$\begin{aligned} \frac{d\lambda}{\lambda_-} &= - \left(G^*(\lambda, \sigma^\lambda + \sigma^R \psi) + f^1(\psi) \right. \\ &\quad \left. + \psi' \sigma^{R'} \sigma^\lambda \right) dt + \sigma_t^{\lambda'} dB_t, \quad \lambda_T = 1, \\ &\quad - \sigma^R (G_\sigma + \sigma^\lambda) \in \partial f^1(\psi) \end{aligned} \quad (33)$$

Given a solution $(\lambda, \sigma^\lambda, \psi)$ and sufficient regularity, then $c_t/W_t = \lambda_t \mathcal{I}^G(t, \lambda, \sigma)$ is substituted into the wealth equation to complete the solution.

Translation-invariant Class

We allow consumption to take any value in $\mathbb{R}$ and fix some strictly positive and bounded reference consumption plan $\gamma \in \mathcal{H}$. The aggregator is assumed to satisfy

$$\begin{aligned} F(\omega, t, c, U, \Sigma) &= G\left(\omega, t, \frac{c}{\gamma(\omega, t)} - U, \Sigma\right), \\ F(\omega, T, c) &= \frac{c}{\gamma(\omega, T)} \end{aligned} \quad (34)$$

which implies that U is quasilinear with respect to γ :

$$U(c + k\gamma) = U(c) + k \quad \text{for all } k \in \mathbb{R} \text{ and } c \in \mathcal{C} \quad (35)$$

Example 5 Additive exponential utility corresponds to

$$G(\omega, t, x, \Sigma) = \beta(\omega, t) - \exp(-x) - \frac{1}{2}\Sigma' \Sigma \quad (36)$$

This follows because the ordinally equivalent utility $V_t = -\exp(-U_t)$ satisfies (under suitable integrability restrictions)

$$V_t = E_t \left[\int_t^T -\exp\left(-\int_t^s \beta_u du - \frac{c_s}{\gamma_s}\right) ds \right]$$

$$- \exp \left(- \int_t^T \beta_u du - \frac{c_T}{\gamma_T} \right) \Big] \quad (37)$$

On the markets side, we assume that the reference consumption stream γ is part of the feasible plan $(\gamma, \Gamma, \Gamma\kappa)$:

$$\frac{d\Gamma_t}{\Gamma_t} = \left(\mu_t^\kappa - \frac{\gamma_t}{\Gamma_t} \right) dt + \kappa' \sigma_t^{R'} dB_t, \quad \Gamma_T = \gamma_T \quad (38)$$

That is, Γ is the price of a fund paying dividend process γ ; $\kappa \in \mathbb{R}^m$ represents the investment proportions of the fund; and μ^κ is the fund's instantaneous expected return process.

For any $(w, \alpha) \in K$, we assume $(w + v, \alpha + v\kappa) \in K$ and $f(\omega, t, w + v, \alpha + v\kappa) = f(\omega, t, w, \alpha) + v\mu^\kappa(\omega, t)$ for all $v \in \mathbb{R}$. That is, trading in the portfolio κ is unrestricted and earns instantaneous expected return μ^κ regardless of the agent's plan. For example, under the linear budget equation (Example 1), we have $\mu^\kappa = r + \kappa' \mu^R$.

Defining the zero-wealth constraint set, aggregator, and portfolio and consumption processes

$$\begin{aligned} K^0 &= \{ \alpha : (0, \alpha) \in K \}, \\ f^0(\omega, t, \alpha) &= f(\omega, t, 0, \alpha), \\ \phi_t^0 &= \phi_t - W_t \kappa, \\ c_t^0 &= c_t - W_t \frac{\gamma_t}{\Gamma_t} \end{aligned} \quad (39)$$

the budget equation (7) is equivalent to

$$\begin{aligned} dW_t &= W_t \frac{d\Gamma_t}{\Gamma_t} + (f^0(t, \phi_t^0) + e_t - c_t^0) dt \\ &\quad + \phi_t^{0'} \sigma_t^{R'} dB_t, \\ c_T^0 &= e_T, \quad \phi_t^0 \in K^0 \end{aligned} \quad (40)$$

At the optimum, the quasi-linearity of utility and markets implies that there are two components to consumption and trading. The pair (c^0, ϕ^0) depends on the investment opportunity set and the endowment, but is independent of W . All incremental financial wealth is invested in the portfolio κ , and the resulting dividend stream rate γ is consumed; therefore, $(c - c^0, \phi - \phi^0)$ depend only on W and the dividend yield

γ/Γ . Utility and marginal utility of wealth processes satisfy

$$U_t = \frac{1}{\Gamma_t} (Y_t + W_t), \quad \lambda_t = \frac{1}{\Gamma_t} \quad (41)$$

where (Y, ϕ^0) is determined by a constrained backward SDE, given below, that is independent of financial wealth.

Defining the superdifferential notation ∂f^0 analogously to ∂f , the state-price density condition $(\zeta, \sigma^{R'} \eta) \in \partial f(W, \phi)$ is equivalent to $\zeta = \mu^\kappa - \kappa' \sigma^{R'} \eta$ and $\sigma^{R'} \eta \in \partial f^0(\phi^0)$. Defining the inverse and dual functions $\mathcal{X}, G^* : \Omega \times [0, T] \times \mathbb{R}^{d+1} \rightarrow \mathbb{R}$ by

$$\begin{aligned} G_x(\omega, t, \mathcal{X}(\omega, t, y, \Sigma), \Sigma) &= y, \\ G^*(\omega, t, y, \Sigma) &= G(\omega, t, \mathcal{X}(\omega, t, y, \Sigma), \Sigma) \\ &\quad - y \mathcal{X}(\omega, t, y, \Sigma) \end{aligned} \quad (42)$$

the processes (Y, σ^Y, ϕ^0) satisfy

$$\begin{aligned} dY &= - (e - Y\mu^\kappa + f^0(\phi^0) + \Gamma G^*(\delta, \Sigma) \\ &\quad - \Gamma \kappa' \sigma^{R'} \Sigma) dt + \sigma^{Y'} dB, \quad Y_T = e_T \\ \Sigma &= \frac{\sigma^Y + (\phi^0 - \kappa Y)' \sigma^R}{\Gamma}, \\ -\sigma^{R'} (G_\Sigma - \sigma^R \kappa) &\in \partial f^0(\phi^0), \quad \phi^0 \in K^0 \end{aligned} \quad (43)$$

Given the solution (Y, σ^Y, ϕ^0) and sufficient regularity, the optimal wealth-independent component of consumption is

$$c_t^0 = \frac{\gamma_t}{\Gamma_t} Y_t + \gamma_t \mathcal{X} \left(t, \frac{\gamma_t}{\Gamma_t}, \Sigma_t \right) \quad \text{and} \quad c_T^0 = e_T \quad (44)$$

Substituting (c^0, ϕ^0) into the budget equation (40), the optimal plan is $(c^0 + W\gamma/\Gamma, W, \phi^0 + W\kappa)$.

Acknowledgments

I am grateful to Costis Skiadas for many fruitful years of joint research, on which this article is based.

End Notes

^aSee also [25], which develops the discrete-time counterpart of (5), and [24], which develops the continuous-time formulations of other notions of ambiguity aversion.

^bWe define arbitrage in the constrained case as a feasible incremental cash flow (given the current portfolio of the agent) that is nonnegative and nonzero.

^cWith some additional mild technical conditions, Schroder and Skiadas [20] show the necessity of the state-price characterization for the market settings in the scale and translation-invariant classes, which are discussed below in the text.

References

- [1] Anderson, E., Hansen, L. & Sargent, T. (2000). *Robustness, Detection and the Price of Risk*, working paper, Department of Economics, University of Chicago.
- [2] Chen, Z. & Epstein, L. (2002). Ambiguity, risk, and asset returns in continuous time, *Econometrica* **70**, 1403–1443.
- [3] Chew, S.H. (1983). A generalization of the Quasi-linear mean with applications to the measurement of inequality and decision theory resolving the allais paradox, *Econometrica* **51**, 1065–1092.
- [4] Cuoco, D. & Cvitanic, J. (1998). Optimal consumption choices for a large investor, *Journal of Economic Dynamics and Control* **22**, 401–436.
- [5] Cuoco, D. & Liu, H. (2000). A martingale characterization of consumption choices and hedging costs with margin requirements, *Mathematical Finance* **10**, 355–385.
- [6] Cvitanic, J. & Karatzas, I. (1992). Convex duality in constrained portfolio optimization, *The Annals of Applied Probability* **2**, 767–818.
- [7] Dekel, E. (1986). An axiomatic characterization of preferences under uncertainty: weakening the independence axiom, *Journal of Economic Theory* **40**, 304–318.
- [8] Duffie, D. & Epstein, L.G. (1992). Stochastic differential utility, *Econometrica* **60**, 353–394.
- [9] Duffie, D. & Skiadas, C. (1994). Continuous-time security pricing: a utility gradient approach, *Journal of Mathematical Economics* **23**, 107–131.
- [10] El Karoui, N., Peng, S. & Quenez, M.-C. (2001). A dynamic maximum principle for the optimization of recursive utilities under constraints, *Annals of Applied Probability* **11**, 664–693.
- [11] Epstein, L.G. & Zin, S.E. (1989). Substitution, risk aversion, and the temporal behavior of consumption and asset returns: a theoretical framework, *Econometrica* **57**, 937–969.
- [12] Epstein, L.G. & Zin, S.E. (1991). Substitution, risk aversion, and the temporal behavior of consumption and asset returns: an empirical analysis, *The Journal of Political Economy* **99**, 263–286.
- [13] Hansen, L., Sargent, T., Turmuhambetova, G. & Williams, N. (2001). *Robustness and Uncertainty Aversion*, working paper, Department of Economics, University of Chicago.
- [14] Kreps, D. & Porteus, E. (1978). Temporal resolution of uncertainty and dynamic choice theory, *Econometrica* **46**, 185–200.
- [15] Lazrak, A. & Quenez, M.C. (2003). A generalized stochastic differential utility, *Mathematics of Operations Research* **28**, 154–180.
- [16] Liu, J. & Longstaff, F.A. (2004). Losing money on arbitrage: optimal dynamic portfolio choice in markets with arbitrage opportunities, *Review of Financial Studies* **17**(3), 611–641.
- [17] Maenhout, P. (1999). *Robust Portfolio Rules and Asset Pricing*, working paper, INSEAD.
- [18] Schroder, M. & Skiadas, C. (2003). Optimal lifetime consumption-portfolio strategies under trading constraints and generalized recursive preferences, *Stochastic Processes and Their Applications* **108**, 155–202.
- [19] Schroder, M. & Skiadas, C. (2005). Lifetime consumption-portfolio choice under trading constraints and non-tradeable income, *Stochastic Processes and their Applications* **115**, 1–30.
- [20] Schroder, M. & Skiadas, C. (2008). Optimality and state pricing in constrained financial markets with recursive utility under continuous and discontinuous information, *Mathematical Finance* **18**, 199–238.
- [21] Skiadas, C. (1992). *Advances in the Theory of Choice and Asset Pricing*, Ph.D. Thesis, Stanford University.
- [22] Skiadas, C. (2003). Robust control and recursive utility, *Finance and Stochastics* **7**, 475–489.
- [23] Skiadas, C. (2008). Dynamic portfolio choice and risk aversion, in *Handbooks in OR & MS*, J.R. Birge & V. Linetsky, eds, Elsevier, Vol. 15, Chapter 19, pp. 789–843.
- [24] Skiadas, C. (2008). *Smooth Ambiguity Aversion Toward Small Risks and Continuous-Time Recursive Utility*, working paper, Kellogg School of Management, Northwestern University.
- [25] Skiadas, C. (2009). *Asset Pricing Theory*, Princeton University Press, Princeton, NJ.

Related Articles

Backward Stochastic Differential Equations; Utility Function; Utility Theory: Historical Perspectives.

MARK SCHRODER

Risk Aversion

An agent is risk averse if he/she dislikes the actions whose outcomes are not certain. In the following, only actions with one-dimensional final outcomes, for example, sums of money, are taken into consideration. To define risk aversion, it is necessary that a probability is associated to every possible consequence, that is, that the actions can be represented as lotteries. A lottery is *simple* if all possible consequences are final outcomes (sums of money) and it is *compound* if other lotteries are included among its consequences. An outcome coincides with a *degenerate* lottery, that is, the lottery that generates it with probability one.

Formally, a decision-making situation under risk is represented by the quintuple $\langle S; 2^S; p; X; \mathcal{L} \rangle$, where S is a set of states of the nature; 2^S is its power set (i.e., the set of all subsets of S , the empty set included); p is a probability distribution on 2^S ; X is a set of outcomes (with $X \subseteq \mathbb{R}$ if they are one-dimensional); and $\mathcal{L}$ is the set of lotteries. A simple lottery is represented by $\ell = (x(E), p(E))_{E \in \text{Part}(S)}$, where outcomes and probabilities are associated with the events $E \subseteq S$ that form a partition $\text{Part}(S)$ of S ; a compound lottery by $\ell = (\ell(E), p(E))_{E \in \text{Part}(S)}$, with $\ell(E) = (\ell'(E'), p(E'))_{E' \in \text{Part}(E)}$; and a degenerate lottery by $\ell = (x, 1)$. A simple lottery is also represented by the cumulative probability function $F : X \rightarrow [0, 1]$, where $F(\cdot)$ is a nondecreasing function with range $[0, 1]$, and, if S is finite, that is, $S = \{s_1, \dots, s_m\}$, by $\ell = (x_i, p_i)_{i=1}^n$, where $p_i = p(E_i)$ with $E_i = \{s_h \in S : x(s_h) = x_i\}$. An agent in a risky situation is a system of preferences $\langle \mathcal{L}, \succsim \rangle$ over the set of lotteries. Let $\langle \mathcal{L}, \succsim \rangle$ be regular (i.e., complete and transitive) and continuous. Moreover, let it be strongly monotone with respect to degenerate lotteries, that is, $(x, 1) \succ (x', 1)$ if $x > x'$. Then, preferences can be represented by a utility function $U : \mathcal{L} \rightarrow \mathbb{R}$, that is, such that $U(\ell) \geq U(\ell')$ if and only if $\ell \succsim \ell'$. This function is not necessarily the expected utility function. However, if the expected utility model is introduced, then every lottery is equivalent to a simple lottery because of the *compound lottery principle* (implied by expected utility), according to which any compound lottery is indifferent to the simple (or reduced) lottery that associates to each final outcome its compound probability, and preferences are represented by the expected utility function,

which is represented by $EU(\ell) = \int_{x \in X} u(x) dF(x)$ if $F(\cdot)$ is differentiable or $EU(\ell) = \sum_{i=1}^n p_i u(x_i)$ if the lottery is finite. The von Neumann–Morgenstern (or Bernoulli) utility function $u : X \rightarrow \mathbb{R}$ represents the preferences over the set of degenerate lotteries, that is, over the set of outcomes.

Definition 1 (*Expected Value*). *The expected value of a lottery $\ell \in \mathcal{L}$ is $EV(\ell) = \int_{x \in X} x dF(x)$ or, if the lottery is finite, $EV(\ell) = \sum_{i=1}^n p_i x_i$ and the function $EV : \mathcal{L} \rightarrow X$ is the expected value function.*

Definition 2 (*Certainty Equivalent*). *The certainty equivalent $CE(\ell)$ of a lottery $\ell \in \mathcal{L}$ is the outcome for which the individual is indifferent between this outcome and the lottery, that is, $(CE(\ell), 1) \sim \ell$, where $(CE(\ell), 1)$ is the degenerate lottery with outcome $CE(\ell)$. Having $U(\ell) = u(CE(\ell))$, the certainty equivalent function is $CE(\ell) = u^{-1}(U(\ell))$. If the system of preferences $\langle \mathcal{L}, \succsim \rangle$ can be represented by an expected utility function, then $CE(\ell) = u^{-1}(EU(\ell))$.*

Proposition 1 (*Existence and Uniqueness of the Certainty Equivalent*). *Let us assume that the set of outcomes is compact, that is, $X = [\underline{x}, \bar{x}]$ and the system of preferences $\langle \mathcal{L}, \succsim \rangle$ is regular (i.e., complete and transitive), continuous and such that $(\bar{x}, 1) \succsim \ell \succsim (\underline{x}, 1)$ for every $\ell \in \mathcal{L}$. Then, there exists one and only one certainty equivalent $CE(\ell) \in X$ for every $\ell \in \mathcal{L}$.*

The following discussion on the notion of risk aversion refers, for the sake of simplicity, to finite simple lotteries on a compact set of outcomes, where not specified differently.

Global Risk Aversion

Definition 3 (*Risk Premium and Global Risk Aversion*). *The risk premium $RP(\ell)$ of a lottery is the maximum sum of money that the agent is willing to pay to get the expected value of the lottery in place of the lottery. Therefore,*

$$RP(\ell) = EV(\ell) - CE(\ell) \quad (1)$$

since the conditions $(EV(\ell) - RP(\ell), 1) \sim \ell$ and $\ell \sim (CE(\ell), 1)$ imply $EV(\ell) - RP(\ell) = CE(\ell)$. The agent denotes (global) *risk aversion* if his/her

system of preferences $\langle \mathcal{L}, \succsim \rangle$ requires $CE(\ell) \leq EV(\ell)$, so $RP(\ell) \geq 0$, for every $\ell \in \mathcal{L}$. The agent is risk loving if $RP(\ell) \leq 0$ and risk neutral if $RP(\ell) = 0$. He/she is strictly risk averse if $RP(\ell) > 0$ for every nondegenerate $\ell \in \mathcal{L}$ (strictly risk loving if $RP(\ell) < 0$). An agent is globally neither risk averse nor risk loving if there is a pair $\ell, \ell' \in \mathcal{L}$ for which $RP(\ell) > 0$ and $RP(\ell') < 0$.

Proposition 2 [5]. *Let us introduce the set of the lotteries that are not preferred to the certain outcome x and the set of lotteries that have an expected value not higher than x , that is,*

$$\begin{aligned} G(x) &= \{\ell \in \mathcal{L} : CE(\ell) \leq x\}, \\ H(x) &= \{\ell \in \mathcal{L} : EV(\ell) \leq x\} \end{aligned} \quad (2)$$

The agent is risk averse if and only if $H(x) \subseteq G(x)$ for every $x \in X$, risk loving if and only if $H(x) \supseteq G(x)$, and risk neutral if and only if $H(x) = G(x)$.

In the Hirshleifer–Yaari diagram, where simple lotteries with only two possible outcomes with given probabilities are represented, the certainty equivalent of a lottery $\ell^* = (x_1^*, p; x_2^*, 1 - p)$ is, by definition, determined as the intersection of the 45° line and the corresponding indifference curve. Therefore, the certainty equivalent is equal to the coordinates of this point. Moreover, the expected value of the same lottery is equal to the coordinates of the point where the 45° line intersects the expected value line (described by the equation $px_1 + (1 - p)x_2 = EV(\ell^*) = px_1^* + (1 - p)x_2^*$). This line passes through ℓ^* and has the slope equal to $-\frac{p}{1-p}$. Thus, the agent is risk averse if the first intersection point is not above the second one, as shown in Figures 1 and 2.

Proposition 2 indicates that risk aversion implies $H(CE(\ell^*)) \subseteq G(CE(\ell^*))$. It means that the indifference curve and the expected value line passing through the same point on the 45° line do not cross and that the indifference curve is to the north-east with respect to the expected value line.

Proposition 3 *If the expected utility model applies, then the agent is risk averse if and only if his/her von Neumann–Morgenstern utility function $u : X \rightarrow \mathbb{R}$ is concave, risk loving if and only if it is convex, and risk neutral if and only if it is a linear.*

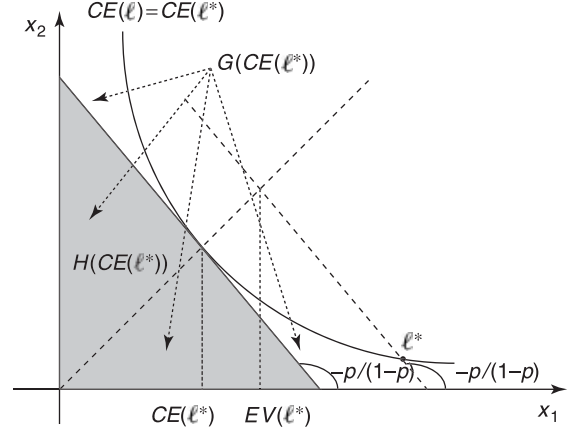


Figure 1 Indifference curve of a risk averse expected utility agent

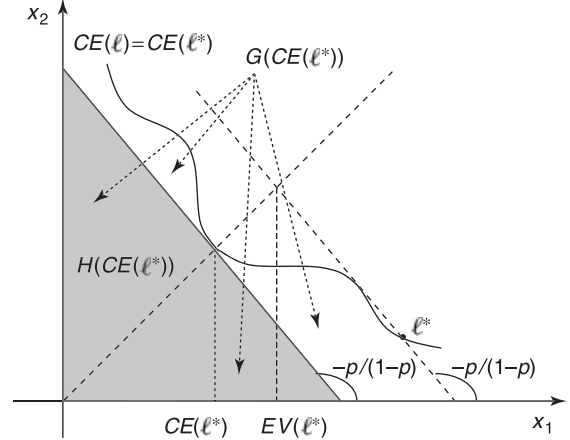


Figure 2 Indifference curve of a risk averse nonexpected utility agent

The inequality $\sum_{i=1}^n p_i u(x_i) \leq u(\sum_{i=1}^n p_i x_i)$, which is called the *Jensen inequality*, is a definition of concavity and is equivalent to $EU(\ell) \leq u(EV(\ell))$.

In Figure 3 we can see how the concavity of the function $u(\cdot)$ implies risk aversion. The expected value, expected utility, and certainty equivalent are represented for the lottery $\ell = (x_1, 0.5; x_2, 0.5)$.

The concavity of the von Neumann–Morgenstern function $u(\cdot)$ implies the concavity of the expected utility function with respect to the outcomes. That is, if $u(\lambda x_i + (1 - \lambda)x_i') \geq \lambda u(x_i) + (1 - \lambda)u(x_i')$ for every pair $x_i, x_i' \in X$ and every $\lambda \in [0, 1]$,

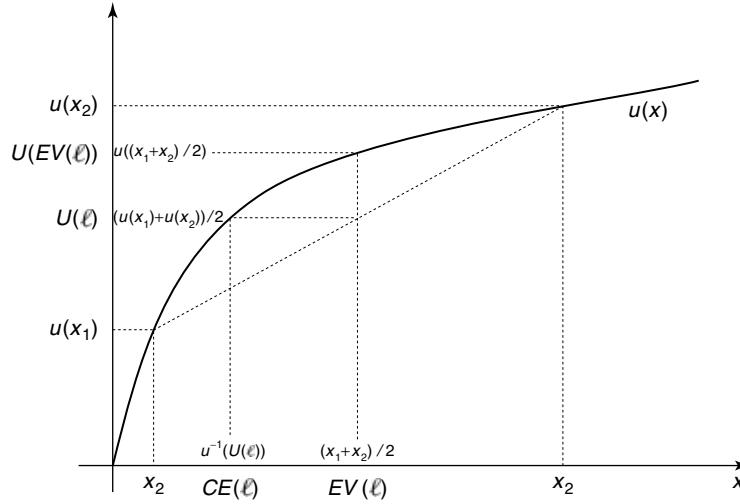


Figure 3 Risk aversion and concavity of utility function

then $EU(\ell'') \geq \lambda EU(\ell) + (1 - \lambda)EU(\ell')$ for every $\lambda \in [0, 1]$ and every triplet $\ell, \ell', \ell'' \in \mathcal{L}$, with $\ell = (x_i, p_i)_{i=1}^n$, $\ell' = (x'_i, p_i)_{i=1}^n$ and $\ell'' = (x''_i, p_i)_{i=1}^n$. Thus, if the agent is risk averse and the expected utility theory holds, the function $EU(\cdot)$ is concave (and, all the more so, quasiconcave) with respect to the outcomes. Consequently, the indifference curves in the Hirshleifer–Yaari diagram are convex (as described in Figure 1, but not in Figure 2, which can represent an agent who is risk averse but does not maximize expected utility).

Proposition 4 [5]. *The agent is risk averse if the certainty equivalent function $CE : \mathcal{L} \rightarrow X$ is convex with respect to the probabilities. The agent is risk loving if it is concave and risk neutral if it is linear.*

The condition stated in Proposition 4 for risk aversion is sufficient, but not necessary, nor is it necessary that the certainty equivalent function $CE(\cdot)$ is quasi-convex with respect to the probabilities. However, if the expected utility theory holds and there is risk aversion, then the certainty equivalent function is convex with respect to the probabilities: in fact, in such a case, we have $CE(\ell) = u^{-1}(\sum_{i=1}^n p_i u(x_i))$, where function $u(\cdot)$ is increasing and concave and function $u^{-1}(\cdot)$ is increasing and convex.

Definition 4 (Comparison of Risk Aversion across Agents). *An agent A is more risk averse than agent*

B if their systems of preferences $\langle \mathcal{L}, \succsim^A \rangle$ and $\langle \mathcal{L}, \succsim^B \rangle$ give $CE_A(\ell) \leq CE_B(\ell)$ for every $\ell \in \mathcal{L}$.

In the Hirshleifer–Yaari diagram, this definition implies that the indifference curves of the agents that go through the same point on the 45° line do not cross and that the indifference curve of the more risk averse agent is to the north-east with respect to the indifference curve of the less risk averse agent, as shown in Figure 4.

Proposition 5 [7]. *If agent A is more risk averse than agent B and the expected utility model applies,*

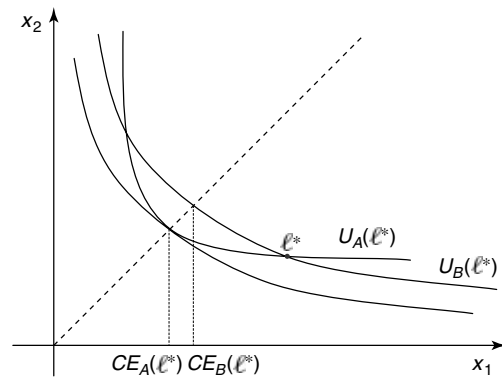


Figure 4 Indifference curves of two agents of whom one is more risk averse than the other

then the von Neumann–Morgenstern utility function $u_A(\cdot)$ is a concave transformation of $u_B(\cdot)$. That is, there exists an increasing and concave function $g : \mathbb{R} \rightarrow \mathbb{R}$ such that $u_A(x) = g(u_B(x))$ for every $x \in X$.

Local Risk Aversion

Till now, we considered *global risk aversion*, that is, the relationship $CE(\ell) \leq EV(\ell)$ was introduced for every lottery $\ell \in \mathcal{L}$. Now, let us consider *local risk aversion*, by taking into account only small lotteries, that is, the lotteries that have only little differences in consequences. For this purpose, we denote the lottery $(x + tx_i, p_i)_{i=1}^n$ with $x + t\ell$, where $\ell = (x_i, p_i)_{i=1}^n$.

Definition 5 (*Local Risk Aversion*). An agent is locally risk averse, if, for every $x \in X$ and $\ell \in \mathcal{L}$, there exists a $t^* > 0$ such that $CE(x + t\ell) \leq EV(x + t\ell)$ for all $t \in [0, t^*]$. Thus, if the certainty equivalent function can be derived, then the agent is locally risk averse if $\lim_{t \rightarrow 0} \frac{d}{dt} (EV(x + t\ell) - CE(x + t\ell)) > 0$ and only if $\lim_{t \rightarrow 0} \frac{d}{dt} (EV(x + t\ell) - CE(x + t\ell)) \geq 0$ for every $x \in X$ and $\ell \in \mathcal{L}$. By analogy, the definition holds with reversed inequality signs for the local risk loving.

Although the global risk aversion requires that in the Hirshleifer–Yaari diagram the indifference curve and the expected value line passing through some point on the 45° line do not cross and that the indifference curve is to the north-east with respect to the expected value line, this condition needs to be satisfied only in the vicinity of the 45° line for the local risk aversion.

Proposition 6 If the expected utility theory holds, then the agent is locally risk averse if and only if his/her von Neumann–Morgenstern utility function $u : X \rightarrow \mathbb{R}$ is concave. In other words, if the expected utility theory holds, then the conditions for local and global risk aversion (risk loving or neutrality) are the same.

Measure of the Risk Aversion

If the expected utility theory holds, then the local risk aversion can be measured by the concavity of the von Neumann–Morgenstern utility function $u(\cdot)$. However, the second derivative of the utility function

$u''(\cdot)$, which is a measure of its concavity, is not invariant to increasing linear transformations of $u(\cdot)$. An invariant measure is the *de Finetti–Arrow–Pratt coefficient of risk aversion* (due to de Finetti [3], Pratt [7], and Arrow [1]). This measure of (*absolute*) risk aversion is defined as

$$r(x) = -\frac{u''(x)}{u'(x)} \quad (3)$$

There also exists a measure of *relative risk aversion* $r_r(x) = -x \frac{u''(x)}{u'(x)}$, which is important in the case of multiplicative lotteries $\ell = (\alpha_i W, p_i)_{i=1}^n$.

The de Finetti–Arrow–Pratt measure can be justified in relation to the local risk premium, which is (by Definition 3)

$$\begin{aligned} RP(x + t\ell) &= EV(x + t\ell) - CE(x + t\ell) \\ &= x + tEV(\ell) \\ &\quad - u^{-1} \left(\sum_{i=1}^n p_i u(x + tx_i) \right) \end{aligned} \quad (4)$$

Then, assuming that this function is differentiable with respect to t , we get $RP(x) = 0$, $\left. \frac{\partial RP(x + t\ell)}{\partial t} \right|_{t=0} = 0$ and $\left. \frac{\partial^2 RP(x + t\ell)}{\partial t^2} \right|_{t=0} = -\frac{u''(x)}{u'(x)} \sigma^2(\ell)$. Therefore, in the neighborhood of the certain outcome x , the risk premium is proportional to the de Finetti–Arrow–Pratt measure. Nevertheless, the fact that only the second derivative of the risk premium can be different from zero at $t = 0$, while the first derivative is always equal to zero, means that the expected utility theory allows only for local risk aversion of the second order, while that of the first order is zero. Other theories (e.g., *rank-dependent expected utility*, which is discussed later) also allow for the risk aversion of the first order and can, as a result, describe the preferences that indicate more relevant types of aversion to risk (like the one presented in Allais paradox) than the risk aversion admitted by the expected utility theory and measured by the de Finetti–Arrow–Pratt index.

Local risk aversion in the Hirshleifer–Yaari diagram is linked to the curvature of indifference curves at the point where they intersect the 45° line. In other words, it is linked to the value of the second derivative $x_2''(x_1)$ at $x_1 = x$, where

the function $x_2(x_1)$ that represents the indifference curve is implicitly defined by the condition $CE(x_1, x_2) = x$. Then, if the expected utility theory holds, we get $x_2(x) = x$, $x_2'(x) = -\frac{p}{1-p}$ and $x_2''(x) = -\frac{p}{(1-p)^2} \frac{u''(x)}{u'(x)}$, that is, the curvature of the indifference curves along the 45° line is proportional to the de Finetti–Arrow–Pratt measure of risk aversion.

The dependence of the de Finetti–Arrow–Pratt index $r(x)$ on x defines the decreasing absolute risk aversion if $r'(x) < 0$ (increasing if $r'(x) > 0$), as well as, with regard to $r_r(x)$, the decreasing relative risk aversion if $r_r'(x) < 0$ (increasing if $r_r'(x) > 0$).

Aversion Toward Increases in Risk

Risk aversion can also be analyzed taking into account the riskiness of lotteries, that is, considering preference for less risky lotteries. However, there does not exist a unique definition of riskiness according to which lotteries can be ordered. In the following, only two definitions of riskiness are examined. Both introduce a partial ordering criterion.

1. The first definition refers to *mean preserving spreads* (introduced by Rothschild and Stiglitz [10]). A lottery $\ell = (x_i, p_i)_{i=1}^n$ is not less risky than lottery $\ell^* = (x_i^*, p_i^*)_{i=1}^n$ if ℓ can be obtained from ℓ^* by mean preserving spreads. That is, if $EV(\ell) = EV(\ell^*)$, $x_i = x_i^*$ for every $i = 1, \dots, n$ and $p_i = p_i^*$ for every $i = 1, \dots, n$ except for three outcomes $x_a > x_b > x_c$, for which we have $p_a \geq p_a^*$, $p_b \leq p_b^*$, and $p_c \geq p_c^*$. For example, $\ell = (x_1, p_1; x_2, p_2; x_3, p_3)$ is not less risky than $\ell^* = (x_1, p_1^*; x_2, p_2^*; x_3, p_3^*)$ if $p_2 \leq p_2^*$, $p_1 = p_1^* + \frac{x_2 - x_3}{x_1 - x_3}(p_2^* - p_2)$, $p_3 = p_3^* + \frac{x_1 - x_2}{x_1 - x_3}(p_2^* - p_2)$, and $x_1 > x_2 > x_3$.

Definition 6 (*Aversion to Mean Preserving Spreads Increases in Risk*). An agent is averse to the increases in risk if $CE(\ell) \leq CE(\ell^*)$ for every pair of lotteries $\ell, \ell^* \in \mathcal{L}$ with ℓ not less risky than ℓ^* (according to mean preserving spreads).

Proposition 7 If an agent is averse to mean preserving spreads increases in risk, then he/she is also risk averse (for this reason, sometimes the aversion to mean preserving spreads increases in risk is

called *strong risk aversion* and the risk aversion as introduced in Definition 3 is called *weak risk aversion* [2]). To be precise, if $CE(\ell) \leq CE(\ell^*)$ for every pair $\ell, \ell^* \in \mathcal{L}$ with ℓ not less risky than ℓ^* (according to mean preserving spreads), then $CE(\ell) \leq EV(\ell)$ for every $\ell \in \mathcal{L}$.

Proposition 8 If the expected utility model applies, then there is aversion toward mean preserving spreads increases in risk if and only if the von Neumann–Morgenstern utility function $u : X \rightarrow \mathbb{R}$ is concave.

Note that the concavity of the utility function is a necessary and sufficient condition for both risk aversion and aversion to increases in risk (determined by mean preserving spreads). The equality of this condition holds in the case of expected utility theory. For other theories, we will generally have two different conditions (one for risk aversion and the other for the aversion to increases in risk).

An ordering of the lotteries according to their riskiness that is equivalent to the mean preserving spreads concept (for the lotteries that have equal expected value) is provided by the notion of the *second-order stochastic dominance*.

Definition 7 (*First-order Stochastic Dominance*). A lottery $\ell = (x_i, p_i)_{i=1}^n$, where $x_i > x_{i+1}$ for every $i = 1, \dots, n-1$, first order stochastically dominates lottery $\ell' = (x_i, p_i')_{i=1}^n$ if $\sum_{h=1}^i p_h \geq \sum_{h=1}^i p_h'$ (or, equivalently, $\sum_{h=i+1}^n p_h \leq \sum_{h=i+1}^n p_h'$) for every $i = 1, \dots, n-1$, that is, with respect to the cumulative probability functions (introduced earlier), if $F(x) \leq F'(x)$ for every $x \in X$.

First-order stochastic dominance means that probabilities of the better (worse) outcomes are higher (lower) in the dominant lottery than in the dominated lottery. It implies that $EV(\ell) \geq EV(\ell')$ and, also, $CE(\ell) \geq CE(\ell')$ for a rational agent.

Definition 8 (*Second-order Stochastic Dominance*). A lottery $\ell = (x_i, p_i)_{i=1}^n$, where $x_i > x_{i+1}$ for every $i = 1, \dots, n-1$, second order stochastically dominates lottery $\ell' = (x_i, p_i')_{i=1}^n$ if $D_j(\ell, \ell') = \sum_{i=j}^{n-1} (x_i - x_{i+1}) \sum_{h=1}^j (p_h - p_h') \geq 0$ for every $j = 1, \dots, n-1$, that is, with respect to the cumulative probability functions in the continuous case, if $\int_{\underline{x}}^x (F(t) - F'(t))dt \leq 0$ for every $x \in X = [\underline{x}, \bar{x}]$. The first-order

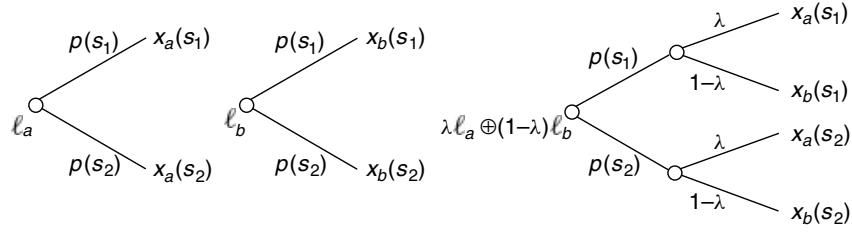


Figure 5 Probability mixture of two lotteries

stochastic dominance implies second-order stochastic dominance, but not vice versa.

Proposition 9 *Let two lotteries ℓ and ℓ' have the same expected value, so that $\sum_{i=1}^{n-1} (x_i - x_{i+1}) \sum_{h=1}^j (p_h - p'_h) = 0$. If the lottery ℓ' is more risky than ℓ (according to the mean preserving spreads criterion), then ℓ second-order stochastically dominates ℓ' . Conversely, if ℓ second-order stochastically dominates ℓ' , then ℓ' can be obtained from ℓ by a sequence of mean preserving spreads.*

The equivalence of the second-order stochastic dominance and mean preserving spreads for the lotteries with the same expected value implies that the same conditions that determine the aversion to the increases in risk (introduced by mean preserving spreads) also determine the aversion for the lotteries that are second-order stochastically dominated (in comparison between lotteries of the same expected value).

- The second definition of riskiness refers to *probability mixtures* [11]. According to this definition, a compound lottery is, *ceteris paribus*, more risky than a simple lottery. More precisely, let us define as a probability mixture of two simple lotteries $\ell_a = (x_a(s_j), p(s_j))_{j=1}^m$ and $\ell_b = (x_b(s_j), p(s_j))_{j=1}^m$, where $S = \{s_1, \dots, s_m\}$ is the set of the states of the nature, the two-stages lottery $\lambda \ell_a \oplus (1-\lambda) \ell_b = (((x_a(s_j), \lambda), (x_b(s_j), (1-\lambda))), p(s_j))_{j=1}^m$, where $\lambda \in [0, 1]$. Figure 5 represents the simplest case of a probability mixture.

Definition 9 (*Aversion to Probability Mixture Increases in Risk*). *An agent is averse to the increases in risk if $CE(\lambda \ell_a \oplus (1-\lambda) \ell_b) \leq \max\{CE(\ell_a), CE(\ell_b)\}$ for every pair of lotteries $\ell_a, \ell_b \in \mathcal{L}$ and $\lambda \in [0, 1]$.*

Note that the expected utility model implies neutrality toward probability mixture increases in risk, since this model satisfies the compound lottery principle, according to which $EU(\lambda \ell_a \oplus (1-\lambda) \ell_b) = \lambda EU(\ell_a) + (1-\lambda) EU(\ell_b)$.

Risk Aversion and Aversion to Increasing Risk with Regard to Rank-dependent Expected Utility

Let us take into consideration a generalization of expected utility theory in order to show some aspects of risk aversion and aversion to increasing risk, which appear very different from the case of expected utility.

Definition 10 (*Rank-dependent Expected Utility* [8, 4]). *The system of preferences $\langle \mathcal{L}, \succsim \rangle$ is represented by rank-dependent expected utility $U : \mathcal{L} \rightarrow \mathbb{R}$ if, for every lottery $\ell \in \mathcal{L}$ with $\ell = (x_i, p_i)_{i=1}^n$ and $x_i > x_{i+1}$ for every $i = 1, \dots, n-1$, where $x_i \in X$ with $X = [\underline{x}, \bar{x}] \subset \mathbb{R}$, we have*

$$U(\ell) = u(x_n) + \sum_{i=1}^{n-1} (u(x_i) - u(x_{i+1})) \varphi \left(\sum_{h=1}^i p_h \right) \quad (5)$$

where function $u : X \rightarrow \mathbb{R}$ represents the system of preferences $\langle X, \succsim \rangle$ over the set of outcomes and function $\varphi : [0, 1] \rightarrow [0, 1]$, which is increasing, with $\varphi(0) = 0$ and $\varphi(1) = 1$, distorts the decumulative probability function.

Thus, the rank-dependent expected utility model describes the agent's system of preferences by means of a utility function on outcomes and a probability distortion function (while the expected utility model requires only the first function). Note that, when

the probability distortion function is the identity function, that is, when $\varphi(p) = p$ for every $p \in [0, 1]$, then rank-dependent expected utility coincides with expected utility.

Recalling that an agent is risk averse if $CE(\ell) \leq EV(\ell)$ for every $\ell \in \mathcal{L}$, that is, if the risk premium $RP(\ell) = EV(\ell) - CE(\ell)$ is nonnegative for every $\ell \in \mathcal{L}$, let us split the risk premium $RP(\ell)$ into two parts: first-order risk premium $RP_1(\ell) = CE_{EU}(\ell) - CE(\ell)$ (with $CE_{EU}(\ell) = u^{-1}(\sum_{i=1}^n p_i u(x_i))$) and second-order risk premium $RP_2(\ell) = EV(\ell) - CE_{EU}(\ell)$.

Proposition 10 [6]. Let $\langle \mathcal{L}, \succsim \rangle$ be represented by rank-dependent expected utility. There is first-order risk aversion, that is, $RP_1(\ell) = CE_{EU}(\ell) - CE(\ell) \geq 0$ for every $\ell \in \mathcal{L}$, if and only if probability distortion function $\varphi : [0, 1] \rightarrow [0, 1]$ is such that $\varphi(p) \leq p$ for every $p \in [0, 1]$. The agent exhibits second-order risk aversion, that is, $RP_2(\ell) = EV(\ell) - CE_{EU}(\ell) \geq 0$ for every $\ell \in \mathcal{L}$, if and only if the utility function $u : X \rightarrow \mathbb{R}$ is concave. As a consequence, an agent is risk averse, that is, $RP(\ell) = EV(\ell) - CE(\ell) \geq 0$ for every $\ell \in \mathcal{L}$, if $\varphi(p) \leq p$ for every $p \in [0, 1]$ and $u : X \rightarrow \mathbb{R}$ is concave. In essence, the condition $\varphi(p) \leq p$ means that the agent overstates the probabilities of some bad outcomes and understates the probabilities of some better outcomes. Because of the probability distortion, rank-dependent expected utility admits first-order risk aversion, therefore allowing for a significant risk aversion even when stakes are small, contrary to expected utility [9]. This may be relevant in finance applications when the agent's choice concerns lotteries in which a small amount of wealth is involved.

Proposition 11 Let $\langle \mathcal{L}, \succsim \rangle$ be represented by rank-dependent expected utility. The agent is locally risk averse if the probability distortion function $\varphi : [0, 1] \rightarrow [0, 1]$ is such that $\varphi(p) < p$ for every $p \in (0, 1)$ and only if $\varphi(p) \leq p$.

In other words, if the rank-dependent expected utility theory holds, then the condition for local risk aversion concerns only the probability distortion function. (As a consequence, the de Finetti–Arrow–Pratt coefficient of risk aversion, which has as its object the utility function $u(\cdot)$, is of no importance in the case of rank-dependent expected utility.)

Another interesting point is that the first-order derivative of risk premium $RP(x + t\ell)$ with

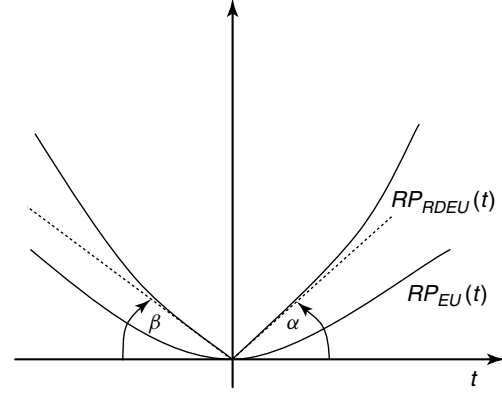


Figure 6 Risk premium function of a risk averse agent

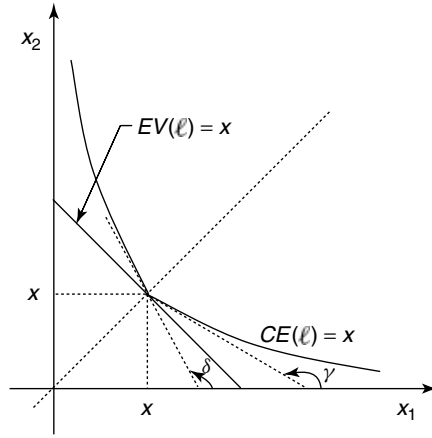


Figure 7 Indifference curve of a risk averse rank-dependent expected utility agent

respect to t is generally nonzero and discontinuous at $t = 0$. For example, if $n = 2$ and $x_1 > x_2$, we get $\lim_{t \rightarrow 0^+} \frac{\partial RP(x + t\ell)}{\partial t} = (x_1 - x_2)(p_1 - \varphi(p_1))$ and $\lim_{t \rightarrow 0^-} \frac{\partial RP(x + t\ell)}{\partial t} = (x_1 - x_2)(p_2 - \varphi(p_2))$. (However, the expected utility theory would yield $\lim_{t \rightarrow 0} \frac{\partial RP(x + t\ell)}{\partial t} = 0$.) In Figure 6, the curve $RP(t)$ represents function $RP(x + t\ell)$ of a risk averse agent, where $\tan \alpha = (x_1 - x_2)(p_1 - \varphi(p_1))$ and $\tan \beta = (x_1 - x_2)(p_2 - \varphi(p_2))$. The curve $RP_{EU}(t)$ represents the same function when the expected utility theory holds. In the Hirshleifer–Yaari diagram (Figure 7), the indifference curves have a kink at $x_1 = x_2$, with $\lim_{x_1 - x_2 \rightarrow 0^+} \frac{dx_2(x_1)}{dx_1} =$

$$\frac{\varphi(p_1)}{1 - \varphi(p_1)} = \tan \gamma \quad \text{and} \quad \lim_{x_1 - x_2 \rightarrow 0^-} \frac{dx_2(x_1)}{dx_1} = \frac{\varphi(p_2)}{1 - \varphi(p_2)} = \tan \delta.$$

If the expected utility theory is valid, then both risk aversion and aversion toward increases in risk (introduced with mean preserving spreads) come from the same condition, which is concavity of the von Neumann–Morgenstern utility function (Propositions 3 and 8). These conditions are different when the rank-dependent expected utility theory holds. Moreover, a rank-dependent expected utility agent may exhibit aversion to probability mixture increases in risk, while the expected utility agent is always neutral.

Proposition 12 *Let $\langle \mathcal{L}, \succsim \rangle$ be represented by rank-dependent expected utility. Then, an agent is averse toward (mean preserving spreads) increases in risk if the function $\varphi : [0, 1] \rightarrow [0, 1]$ is convex and the function $u : X \rightarrow \mathbb{R}$ is concave. He/she is averse toward (probability mixtures) increases in risk if and only if the function $\varphi : [0, 1] \rightarrow [0, 1]$ is convex.*

References

- [1] Arrow, K.J. (1965). *Aspects of the Theory of Risk-Bearing*, Yrjö Jahnssonin Säätiö, Helsinki.
- [2] Cohen, M.D. (1995). Risk-aversion concepts in expected- and non-expected-utility models, *Geneva Papers on Risk and Insurance Theory* **20**, 73–91.
- [3] de Finetti, B. (1952). Sulla preferibilità, *Giornale degli Economisti NS* **11**, 685–709.
- [4] Machina, M.J. (1987). Choice under uncertainty: problems solved and unsolved, *Economic Perspectives* **1**, 121–154.
- [5] Montesano, A. (1999). Risk and uncertainty aversion on certainty equivalent functions, in *Beliefs, Interactions and Preferences in Decision Making*, M.J. Machina & B. Munier, eds, Kluwer, Dordrecht, pp. 23–52.
- [6] Montesano, A. (1999). Risk and uncertainty aversion with reference to the theories of expected utility, rank dependent expected utility, and Choquet expected utility, in *Uncertain Decisions: Bridging Theory and Experiments*, L. Luini, ed, Kluwer, Boston, pp. 3–37.
- [7] Pratt, J.W. (1964). Risk aversion in the small and in the large, *Econometrica* **32**, 122–136.
- [8] Quiggin, J. (1982). A theory of anticipated utility, *Journal of Economic Behavior and Organization* **3**, 323–343.
- [9] Rabin, M. (2000). Risk aversion and expected utility theory: a calibration theorem, *Econometrica* **68**, 1281–1292.
- [10] Rothschild, M. & Stiglitz, J.E. (1970). Increasing risk: I. A definition, *Journal of Economic Theory* **2**, 225–243.
- [11] Wakker, P.P. (1994). Separating marginal utility and probabilistic risk aversion, *Theory and Decision* **36**, 1–44.

Related Articles

Ambiguity; Behavioral Portfolio Selection; Expected Utility Maximization; Risk–Return Analysis; Utility Function.

ALDO MONTESANO

Ambiguity

In the literature on decision making under uncertainty, *ambiguity* is now consistently used to define those decision settings in which an economic agent perceives “[...] uncertainty about probability, created by missing information that is relevant and could be known” [17]. Other terms have been used interchangeably, notably “Knightian uncertainty,” based on Knight’s [32] distinction between “risk” (a context in which all the relevant “odds” are known and unanimously agreed upon) and “uncertainty” (a context in which some “odds” are not known). The term *ambiguity*, which avoids charging uncertainty with too many meanings, was introduced in [12], the paper that first showed how ambiguity represents a normative criticism to Savage’s [38] subjective expected utility (SEU) model.

Ellsberg proposed two famous thought experiments involving choices on urns in which the exact distribution of ball colors is unknown (one of which was anticipated in both [29] and [32]). A variant of Ellsberg’s so-called two-urn paradox is the following example, due to David Schmeidler. “Suppose that I ask you to make bets on two coins, one taken out of your pocket—a coin, which you have flipped countless times—the other taken out of my pocket. If asked to bet on ‘heads’ or on ‘tails’ on one of the two coins, would you rather bet on your coin or mine?” Most people, when posed this question, announce a mild but strict preference for betting on their own coin rather than on somebody else’s, *both for heads and for tails*. The rationale is precisely that their coin has a well-understood stochastic behavior, while the other person’s coin does not; that is, its behavior is *ambiguous*. The possibility that the coin be biased, although remote, cannot be dismissed altogether. This pattern of preference is called *ambiguity aversion*, and is, as suggested, very common ([6, p. 646] e.g., references many experimental replications of the “paradox”). It is easy to see that it is not compatible with the SEU model. For, suppose that a decision maker has a probabilistic prior P over the state space $S = \{HH, HT, TH, TT\}$ (where HT is the state in which the familiar coin lands heads up and the unfamiliar coin lands tails up, etc.). Then, by saying that he/she prefers a bet that pays off €1 if the familiar coin lands heads up—that is, a bet on the event $A = \{HH, HT\}$ —to the bet that pays €1 if

the unfamiliar coin lands heads up—that is, a bet on the event $B = \{HH, TH\}$ —an SEU decision maker reveals that

$$u(1) P(A) + u(0) (1 - P(A)) > u(1) P(B) + u(0) (1 - P(B)) \quad (1)$$

that is, $P(A) > P(B)$. Analogously, by preferring the bet on tails on the familiar coin to the bet on tails on the unfamiliar coin, an SEU decision maker reveals that

$$P(\{TH, TT\}) = P(A^c) = 1 - P(A) > 1 - P(B) = P(B^c) = P(\{HT, TT\}) \quad (2)$$

that is, $P(A) < P(B)$: a contradiction. Yet, few people would immediately describe these preferences as being an example of irrationality. Ellsberg reports that Savage himself chose in the manner described above, and did not feel that his choices were clearly wrong [12, p. 656]. (Indeed, Savage was aware of the issue well before Ellsberg proposed his thought experiments, for Savage wrote in the *Foundations of Statistics* (pp. 57–58) that “there seem to be some probability relations about which we feel relatively ‘sure’ as compared to others,” adding that he did not know how to make such notion of comparatively “sure” less vague.)

Ellsberg’s paper generated quite a bit of debate immediately after its publication (most of which is discussed in Ellsberg’s PhD dissertation [13]), but the lack of axiomatically founded models that could encompass a concern for ambiguity while retaining most of the compelling features of the SEU model worked to douse the flames. Moreover, the so-called *Allais paradox* [2], another descriptive failure of expected utility, which predated Ellsberg’s by a few years, monopolized the attention of decision theorists until the early 1980s. However, statisticians such as Good [23] and Arthur Dempster [9] did lay the foundations of statistics with sets of probabilities, providing analysis and technical results, which eventually made it into the toolbox of decision theorists.

Models of Ambiguity-sensitive Preferences

The interest in ambiguity as a reason for departure from the SEU model was revived by David Schmeidler, who proposed and characterized axiomatically

two of the most successful models of decision making in the presence of ambiguity, the *Choquet expected utility* (CEU) and the *maxmin expected utility* (MEU) models.

CEU [39] “resolves” the Ellsberg paradox by allowing a decision maker’s willingness to bet on an event to be represented by a set-function that is not necessarily additive; that is, a v , which, to disjoint events A and B , may assign $v(A \cup B) \neq v(A) + v(B)$. More precisely, call a *capacity* any function v defined on a σ -algebra Σ of subsets of a state space S , which satisfies the following properties: (i) $v(\emptyset) = 0$, (ii) $v(S) = 1$, (iii) for any $A, B \in \Sigma$ such that $A \subseteq B$, $v(A) \leq v(B)$. (Note that a *probability (charge)* is v , which satisfies instead of (iii) the property $v(A \cup B) = v(A) + v(B) - v(A \cap B)$ for any $A, B \in \Sigma$.) It is simple to see that if v represents a decision maker’s beliefs, we may observe the preferences described above in the two-coin example. Just substitute P in equations (1) and (2) with v satisfying $v(A) = v(A^c) = 1/2$ and $v(B) = v(B^c) = 1/4$. The obvious question is that of defining expectations for a notion of “belief”, which is not a measure. As the model’s name suggests, Schmeidler used the notion of integral for capacities, which was developed by Choquet [8]. Formally, given a capacity space (S, Σ, v) and a Σ -measurable function $a : S \rightarrow \mathbb{R}$, the *Choquet integral of a with reference to (w.r.t.) v* is given by the following formula:

$$\int_S a(s) dv(s) \equiv \int_0^\infty v(\{s \in S : a(s) \geq \alpha\}) d\alpha + \int_{-\infty}^0 [v(\{s \in S : a(s) \geq \alpha\}) - 1] d\alpha \quad (3)$$

This is shown to correspond to Lebesgue integration when the capacity v is a probability. Schmeidler provided axioms on a decision maker’s preference relation $\succsim$, which guarantee that the latter is represented by the Choquet expectation w.r.t. v of a real-valued utility function u (on final prizes $x \in X$). Precisely, given choice options (acts) $f, g : S \rightarrow X$,

$$f \succsim g \iff \int_S u(f(s)) dv(s) \geq \int_S u(g(s)) dv(s) \quad (4)$$

That is, the decision maker prefers f to g whenever the Choquet integral of $u \circ f$ is greater than that

of $u \circ g$. The interested reader is referred to Schmeidler’s paper for details of the axiomatization. For our purpose, it suffices to observe that, not too surprisingly, the key axiomatic departure from SEU (in the variant due to [3]) is a relaxation of the independence axiom—or what Savage calls the *sure-thing principle*—which is the property of preferences that the Ellsberg-like preferences above violate.

Not all capacities give rise to behavior which is averse to ambiguity, as in the above example. Schmeidler proposed the following behavioral notion of aversion to ambiguity. Assuming that the payoffs x can themselves be (objective and additive) lotteries over a set of certain prizes, define for any $\alpha \in [0, 1]$ the α -mixture of acts f and g as follows: for any $s \in S$,

$$(\alpha f + (1 - \alpha)g)(s) \equiv \alpha f(s) + (1 - \alpha)g(s) \quad (5)$$

where the object on the right-hand side is the lottery that pays off prize $f(s)$ with probability α and prize $g(s)$ with probability $(1 - \alpha)$. Now, say that a preference satisfies *ambiguity hedging* (Schmeidler calls this property *uncertainty aversion*) if for any f and g such that $f \sim g$ we have

$$\alpha f + (1 - \alpha)g \succsim f \quad (6)$$

for any α . That is, the decision maker *may* prefer to “hedge” the ambiguous returns of two indifferent acts by mixing them appropriately. This makes sense if we consider two acts whose payoff profiles are negatively correlated (over S), so that the mixture has a payoff profile, which is flatter, hence less sensitive to the information on S , than the original acts. (Ghirardato and Marinacci [20] discuss ambiguity hedging, arguing that it captures more than just the ambiguity aversion of equations 1 and 2.) Schmeidler shows that a CEU decision maker satisfies ambiguity hedging if and only if her capacity v is *supermodular*; that is, for any $A, B \in \Sigma$,

$$v(A \cup B) \geq v(A) + v(B) - v(A \cap B) \quad (7)$$

Ambiguity hedging also plays a key role in the second model of ambiguity-sensitive preferences proposed by Schmeidler, the MEU model introduced alongside that of Itzhak Gilboa [21]. In MEU, the decision maker’s preferences are represented by (a utility function u and) a *set* C of probability charges

on (S, Σ) —which is nonempty, (weak*)-closed and convex—as follows:

$$\begin{aligned} f \succsim g &\iff \min_{P \in C} \int_S u(f(s)) dP(s) \\ &\geq \min_{P \in C} \int_S u(g(s)) dP(s) \end{aligned} \quad (8)$$

Thus, the presence of ambiguity is reflected by the nonuniqueness of the prior probabilities over the set of states. In the authors’ words, “the subject has too little information to form a prior. Hence, (s)he considers a *set* of priors as possible” [21, p. 142]. In the two-coin example, let S be the product space $\{H, T\} \times \{H, T\}$ and consider the set of priors

$$C \equiv \cup_{a \in [1/4, 3/4]} \{1/2, 1/2\} \times \{a, 1-a\} \quad (9)$$

It is easy to see that a decision maker with such a C will “assign” to events A and A^c the weight $\min_{P \in C} P(A) = 1/2 = \min_{P \in C} P(A^c)$, and to events B and B^c the weight $\min_{P \in C} P(B) = 1/4 = \min_{P \in C} P(B^c)$, thus displaying the classical Ellsberg preferences. Gilboa and Schmeidler showed that MEU is axiomatically very close to CEU. While ambiguity hedging is required (being single-handedly responsible for the “min” in the representation; see [19]), a weaker version of independence is used.

Ambiguity hedging characterizes the intersection of the CEU and MEU models. Schmeidler [39] shows that a decision maker’s preferences have *both* CEU and MEU representations if and only if (i) the v in the CEU representation is supermodular, and (ii) the lower envelope of the set C in the MEU representation, $\underline{C}(\cdot) \equiv \min_{P \in C} P(\cdot)$, is a supermodular capacity and C is the set of *all* the probability charges that dominate $\underline{C}$ (the *core* of $\underline{C}$). On the other hand, there are CEU preferences that are not MEU (take a capacity v which is not supermodular), and MEU preferences that are not CEU (see [30, Example 1]).

The CEU and MEU models brought ambiguity back to the forefront of decision theoretic research, and in due course, as “applications” of such theoretical models started to appear, they were key in attracting the attention of mainstream economics and finance.

On the theoretical front, a number of alternative axiomatic models have been developed. First, there are generalizations of CEU and MEU. For instance, Maccheroni *et al.* [33] presented a model that they

called *variational preferences*, which relaxes the independence condition used in MEU while retaining the ambiguity hedging condition. An important special case of variational preferences is the so-called *multiplier model* of Hansen and Sargent [25], a key model in the applications literature to be discussed later. Siniscalchi [42] proposed a model that he called *vector expected utility*, in which an act is evaluated by modifying its expectation (w.r.t. a “baseline probability”) by an adjustment function capturing ambiguity attitudes. Such a model is also built with applications in mind, as it (potentially) employs a smaller number of parameters than CEU and MEU.

Second, Bewley [4] (originally circulated in 1986) suggested that ambiguity might result in incompleteness of preferences, rather than in violation of independence. Under such assumptions, he found a representation in which a set of priors C appears in a “unanimity” sense as follows:

$$\begin{aligned} f \succsim g &\iff \int_S u(f(s)) dP(s) \\ &\geq \int_S u(g(s)) dP(s) \text{ for all } P \in C \end{aligned} \quad (10)$$

That is, the decision maker prefers f over g whenever f dominates g according to every “possible scenario” in C . Preferences are undecided otherwise, and Bewley suggested completing them by following an “inertia” rule: the *status quo* is retained if undominated by any available act. In a model that joins the two research strands just described, Ghirardato *et al.* [19] showed that if we drop ambiguity hedging from the MEU axioms, we can still obtain the set of priors C as a “unanimous” representation of a suitably defined incomplete subset of the decision maker’s preference relation, which they interpreted as “unambiguous” preference (i.e., a preference that is not affected by the presence of ambiguity). This yields a model—of which both CEU and MEU are special cases—in which the decision maker evaluates act f *via* the functional

$$\begin{aligned} V(f) &= a(f) \min_{P \in C} \int_S u(f(s)) dP(s) \\ &\quad + (1 - a(f)) \max_{P \in C} \int_S u(f(s)) dP(s) \end{aligned} \quad (11)$$

where $a(f) \in [0, 1]$ is the decision maker's ambiguity aversion in evaluating f (a generalization of the decision rule suggested by Hurwicz [27]).

A third modeling approach relaxes the “reduction of compound lotteries” property that is built within the expected utility model. The basic idea is that the decision maker forms a “second-order” probability μ over the set of possible priors over S , and that he/she does not reduce the resulting compound probability. That is, he/she could evaluate act f by first calculating its expectation $E_P(u \circ f) \equiv \int u(f(s)) dP(s)$ with respect to each prior P that he/she deems possible, and then computing

$$\int_{\Delta} \phi(E_P(u \circ f)) d\mu(P) \quad (12)$$

where Δ denotes the set of all possible probability charges on (S, Σ) , and $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is a function, which is not necessarily affine. This is the reasoning adopted by Segal [40], followed by Ergin and Gul [16], Klibanoff *et al.* [31], Nau [37], and Seo [41]. The case of SEU corresponds to ϕ being affine, while Klibanoff *et al.* [31] show that ϕ being concave corresponds intuitively to ambiguity averse preferences. That is, the “external” utility function describes ambiguity attitude, while the “internal” one describes risk attitude. An important feature of such a model is that its representation is smooth (in utility space), whereas those of MEU and CEU are generally not. For this reason, this is called the *smooth ambiguity model*.

In concluding this brief survey of decision models, it is important to stress that, owing to space constraint, the focus is on *static* models. The literature on *intertemporal* models is more recent and less developed, in part, because of the fact that non-SEU preferences often violate a property called *dynamic consistency* [18], making it hard to use the traditional dynamic programming tools. Important contributions in this area are found in [14, 22] (characterizing the so-called *recursive MEU* model) and [24, 34].

Applications

As mentioned above, the CEU and MEU models were finally successful in introducing ambiguity into mainstream research in economics and finance. Many papers have been written, which assume that (some) agents have CEU or MEU preferences. The interested reader is referred to [36] for an extensive survey of

such applications, while some applications to finance are briefly discussed here.

In a seminal contribution, Dow and Werlang [10] showed that a CEU agent with supermodular capacity may display a nontrivial bid–ask spread on the price of an (ambiguous) Arrow security, even without frictions. If the price of the security falls within such an interval, the agent will not want to trade the security at all (given an initial riskless position). Epstein and Wang [15] employed the recursive MEU model to study the equilibrium of a representative agent economy *à la Lucas*. They showed that price indeterminacy can arise in equilibrium for reasons that are closely related to Dow and Werlang's observation. Other contributions followed along this line; for example, see [7, 35, 43]. More recently, the smooth ambiguity model has also been receiving attention; see, for example, [28].

Though originally not motivated by the Ellsberg paradox and ambiguity, the “model uncertainty” literature due to Hansen *et al.* ([26], but more comprehensively found in [25]) falls squarely within the scope of the applications of ambiguity. Moreover, both decision models they employ are special cases of the models described above: the “multiplier model” is a special case of variational preferences, and the “constraint model” is a special case of MEU.

Most of the applications of ambiguity to finance—an exception being [11]—are cast in a representative agent environment, with the preferences of the representative agent satisfying in one case MEU, in another CEU, and so on. Recent work on experimental finance by Bossaerts *et al.* [5] and Ahn *et al.* [1] finds that experimental subjects, when making portfolio choices with ambiguous Arrow securities, display substantial heterogeneity in ambiguity attitudes. Because Bossaerts *et al.* [5] show that such heterogeneity may easily result in a breakdown of the representative agent result, such findings cast some doubt on the generality of a representative agent approach to financial markets equilibrium.

References

- [1] Ahn, D., Choi, S., Gale, D. & Shachar, K. (2007). *Estimating Ambiguity Aversion in a Portfolio Choice Experiment*, UC Berkeley, Mimeo.
- [2] Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats

- et axiomes de l'école américaine, *Econometrica* **21**, 503–546.
- [3] Anscombe, F.J. & Aumann, R.J. (1963). A definition of subjective probability, *Annals of Mathematical Statistics* **34**, 199–205.
- [4] Bewley, T. (2002). Knightian decision theory: part I, *Decisions in Economics and Finance* **25**(2), 79–110. (First version 1986).
- [5] Bossaerts, P., Ghirardato, P., Guarnaschelli, S. & Zame, W.R. (2006). Ambiguity and asset markets: theory and experiment, *Review of Financial Studies*, forthcoming, Notebook 27, Collegio Carlo Alberto.
- [6] Camerer, C. (1995). Individual decision making, in *The Handbook of Experimental Economics*, J.H. Kagel & A.E. Roth, eds, Princeton University Press, Princeton, NJ, pp. 587–703.
- [7] Chen, Z. & Epstein, L.G. (1999). *Ambiguity, Risk and Asset Returns in Continuous Time*, University of Rochester, Mimeo.
- [8] Choquet, G. (1953). Theory of capacities, *Annales de l'Institut Fourier (Grenoble)* **5**, 131–295.
- [9] Dempster, A.P. (1967). Upper and lower probabilities induced by a multi-valued mapping, *Annals of Mathematical Statistics* **38**, 325–339.
- [10] Dow, J. & Werlang, S. (1992). Uncertainty aversion, risk aversion, and the optimal choice of portfolio, *Econometrica* **60**, 197–204.
- [11] Easley, D. & O'Hara, M. Ambiguity and nonparticipation: the role of regulation, *Review of Financial Studies* **22**(5), 1817–1843.
- [12] Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms, *Quarterly Journal of Economics* **75**, 643–669.
- [13] Ellsberg, D. (2001). *Risk, Ambiguity and Decision*. PhD thesis, Harvard University, 1962. Published by Garland Publishing Inc., New York.
- [14] Epstein, L.G. & Schneider, M. (2003). Recursive multiple-priors, *Journal of Economic Theory* **113**, 1–31.
- [15] Epstein, L.G. & Wang, T. (1994). Intertemporal asset pricing under Knightian uncertainty, *Econometrica* **62**, 283–322.
- [16] Ergin, H. & Gul, F. (2004). *A Subjective Theory of Compound Lotteries*. February.
- [17] Frisch, D. & Baron, J. (1988). Ambiguity and rationality, *Journal of Behavioral Decision Making* **1**, 149–157.
- [18] Ghirardato, P. (2002). Revisiting Savage in a conditional world, *Economic Theory* **20**, 83–92.
- [19] Ghirardato, P., Maccheroni, F. & Marinacci, M. (2004). Differentiating ambiguity and ambiguity attitude, *Journal of Economic Theory* **118**(2), 133–173.
- [20] Ghirardato, P. & Marinacci, M. (2002). Ambiguity made precise: a comparative foundation, *Journal of Economic Theory* **102**, 251–289.
- [21] Gilboa, I. & Schmeidler, D. (1989). Maxmin expected utility with a non-unique prior, *Journal of Mathematical Economics* **18**, 141–153.
- [22] Gilboa, I. & Schmeidler, D. (1993). Updating ambiguous beliefs, *Journal of Economic Theory* **59**, 33–49.
- [23] Good, I.J. (1962). Subjective probability as the measure of a nonmeasurable set, in *Logic, Methodology and Philosophy of Science*, E. Nagel, P. Suppes & A. Tarski, eds, Stanford University Press, Stanford, pp. 319–329.
- [24] Hanany, E. & Klibanoff, P. (2007). Updating preferences with multiple priors, *Theoretical Economics* **2**(3), 261–298.
- [25] Hansen, L.P. & Sargent, T.J. (2007). *Robustness*, Princeton University Press, Princeton, NJ.
- [26] Hansen, L.P., Sargent, T.J. & Tallarini, T.D. (1999). Robust permanent income and pricing, *Review of Economic Studies* **66**, 873–907.
- [27] Hurwicz, L. (1951). *Optimality Criteria for Decision Making under Ignorance*. Statistics 370, Cowles Commission Discussion Paper.
- [28] Izhakian, Y. & Benninga, S. (2008). *The Uncertainty Premium in an Ambiguous Economy*. Technical report, Recanat School of Business, Tel-Aviv University.
- [29] Keynes, J.M. (1921). A treatise on probability, *The Collected Writings of John Maynard Keynes*, Macmillan, London and Basingstoke, paperback 1988 edition, Vol. VIII.
- [30] Klibanoff, P. (2001). Characterizing uncertainty aversion through preference for mixtures, *Social Choice and Welfare* **18**, 289–301.
- [31] Klibanoff, P., Marinacci, M. & Mukerji, S. (2005). A smooth model of decision making under ambiguity, *Econometrica* **73**(6), 1849–1892.
- [32] Knight, F.H. (1921). *Risk, Uncertainty and Profit*, Houghton Mifflin, Boston.
- [33] Maccheroni, F., Marinacci, M. & Rustichini, A. (2006). Ambiguity aversion, robustness, and the variational representation of preferences, *Econometrica* **74**(6), 1447–1498.
- [34] Maccheroni, F., Marinacci, M. & Rustichini, A. (2006). Dynamic variational preferences, *Journal of Economic Theory* **128**(1), 4–44.
- [35] Mukerji, S. & Tallon, J.-M. (2001). Ambiguity aversion and incompleteness of financial markets, *Review of Economic Studies* **68**(4), 883–904.
- [36] Mukerji, S. & Tallon, J.-M. (2004). An overview of economic applications of David Schmeidler's models of decision making under uncertainty, in *Uncertainty in Economic Theory: A Collection of Essays in Honor of David Schmeidler's 65th Birthday*, I. Gilboa, ed., Routledge, Chapter 13, pp. 283–302.
- [37] Nau, R.F. (2006). Uncertainty aversion with second-order utilities and probabilities, *Management Science* **52**(1), 136.
- [38] Savage, L.J. (1954). *The Foundations of Statistics*, Wiley, New York.
- [39] Schmeidler, D. (1989). Subjective probability and expected utility without additivity, *Econometrica* **57**, 571–587.
- [40] Segal, U. (1987). The Ellsberg paradox and risk aversion: an anticipated utility approach, *International Economic Review* **28**, 175–202.

- [41] Seo, K. (2006). *Ambiguity and Second-order Belief*, University of Rochester, Mimeo.
- [42] Siniscalchi, M. Vector expected utility and attitudes toward variation, *Econometrica* **77**(3), 801–855.
- [43] Uppal, R. & Wang, T. (2003). Model misspecification and under-diversification, *Journal of Finance* **58**(6), 2465–2486.

Related Articles

Behavioral Portfolio Selection; Convex Risk Measures; Expected Utility Maximization; Expected Utility Maximization: Duality Methods; Risk Aversion; Utility Function; Utility Theory: Historical Perspectives.

PAOLO GHIRARDATO

Risk Premia

Risk premia are the expected excess returns that compensate investors for taking on aggregate risk. The first section of this article defines risk premia analytically. The second section surveys empirical evidence on equity, bond, and currency excess returns. The third section reviews the models that explain these risk premia.

Theoretical Definition

Risk premia are derived analytically from Euler equations that link returns to stochastic discount factors (SDFs). These Euler equations can be derived under three different assumptions: complete markets, the law of one price, or the existence of investors' preferences. These three assumptions are reviewed here, followed by the analytical definition of risk premia.

Euler Equations

Utility-based Asset Pricing. Assume that the investor derives some utility u from consumption C now and in the next period. This setup can be easily generalized to many periods. Let us find the price P_t at time t of a payoff X_{t+1} at time $t + 1$. Let Q be the original consumption level in the absence of any asset purchase and let ξ be the amount of the asset the investor chooses to buy. The constant subjective discount factor is β . The maximization problem of this investor is

$$\begin{aligned} & \text{Max}_{\xi} \quad u(C_t) + E_t[\beta u(C_{t+1})] \\ & \text{subject to: } C_t = Q_t - P_t \xi, \\ & \quad C_{t+1} = Q_{t+1} + X_{t+1} \xi \end{aligned} \quad (1)$$

Substituting the constraints into the objective and setting the derivative with respect to ξ to zero yields

$$P_t u'(C_t) = E_t[\beta u'(C_{t+1}) X_{t+1}] \quad (2)$$

where $P_t u'(C_t)$ is the loss in utility if the investor buys another unit of the asset, and $E_t[\beta u'(C_{t+1}) X_{t+1}]$ is the expected and discounted increase in utility he/she obtains from the extra payoff X_{t+1} . The

investor continues to buy or sell the asset until the marginal loss equals the marginal gain. The Euler equation is thus

$$P_t = E_t \left[\beta \frac{u'(C_{t+1})}{u'(C_t)} X_{t+1} \right] = E_t[M_{t+1} X_{t+1}] \quad (3)$$

where the SDF M_{t+1} is defined as $M_{t+1} \equiv \beta u'(C_{t+1})/u'(C_t)$.

Complete Markets. Let us now abstract from utilities and assume that markets are *complete*. There are S states of nature tomorrow, and s denotes an individual state. A contingent claim is a security that pays one dollar (or one unit of the consumption good) in one state s only tomorrow. The price today of this contingent claim is $P_c(s)$. In complete markets, investors can buy any contingent claim (or synthesize all contingent claims). Let $\underline{X}$ be the payoff space and $X(s) \in \underline{X}$ denote an asset's payoff in state of nature s . Let $\pi(s)$ be the probability that state s occurs. Then the price of this asset is

$$P(X) = \sum_{s=1}^S P_c(s) X(s) = \sum_{s=1}^S \pi(s) \frac{P_c(s)}{\pi(s)} X(s) \quad (4)$$

We define M as the ratio of the contingent claim's price to the corresponding state's probability $M(s) \equiv P_c(s)/\pi(s)$ to obtain the Euler equation in complete markets:

$$P(X) = \sum_{s=1}^S \pi(s) M(s) X(s) = E(MX) \quad (5)$$

Law of One Price and the Absence of Arbitrage.

Finally, assume now that markets are *incomplete* and that we simply observe a set of prices P and payoffs X . Under a minimal set of assumptions, some discount factor exists that represents the observed prices by the same equation $P = E(MX)$. These assumptions are defined below:

Definition 1 *Free portfolio formation:* $X_1, X_2 \in \underline{X} \Rightarrow aX_1 + bX_2 \in \underline{X}$ for any real a and b .

Definition 2 *Law of one price:* $P(aX_1 + bX_2) = aP(X_1) + bP(X_2)$.

Note that free portfolio formation rules out short sales constraints, bid/ask spreads, leverage limitations, and so on. The law of one price says

that investors cannot make instantaneous profits by repackaging portfolios. These assumptions lead to the following theorem:

Theorem 1 *Given free portfolio formation and the law of one price, there exists a unique payoff $X^* \in \underline{X}$ such that $P(X) = E(X^*X)$ for all $X \in \underline{X}$.*

As a result, there exists an SDF M such that $P(X) = E(MX)$. Note that the existence of a discount factor implies the law of one price $E[M(X + Y)] = E[MX] + E[MY]$. The theorem reverses this logic. Cochrane [7] offers a geometric and an arithmetic proof. With a stronger assumption, the absence of arbitrage, the SDF is strictly positive and thus represents some—potentially unknown—preferences. Let us first review the definition of the absence of arbitrage and then turn to this new theorem.

Definition 3 *Absence of arbitrage: A payoff space $\underline{X}$ and pricing function $P(X)$ leave no arbitrage opportunities if every payoff X that is always nonnegative ($X \geq 0$ almost surely) and strictly positive ($X > 0$) with some positive probability has some strictly positive price $P(X) > 0$.*

In other words, no arbitrage says that one cannot get for free a portfolio that might pay off positively but will certainly never cost one anything. This assumption leads to the next theorem:

Theorem 2 *No arbitrage and the law of one price imply the existence of a strictly positive discount factor $M > 0$ such that $P = E(MX)$, $\forall X \in \underline{X}$.*

We have seen three ways to derive the Euler equation that links any asset's price to the SDF. Before we exploit the Euler equation to define risk premia, note that only aggregate risk matters for asset prices.

Aggregate and Idiosyncratic Risk. Only the component of payoffs that is correlated with the SDF shows up in the asset's price. Idiosyncratic risk, uncorrelated with the SDF, generates no premium. To see this, let us project X on M and decompose the payoff as follows:

$$X = \text{proj}(X|M) + \varepsilon \quad (6)$$

Projecting X on M is like regressing X on M without a constant:

$$\text{proj}(X|M) = \frac{E(MX)}{E(M^2)}M \quad (7)$$

The residuals ε are orthogonal to the right-hand side variable M : $E(M\varepsilon) = 0$, which means that the price of ε is zero. The price of the projection of X on M is the price of X :

$$P(\text{proj}(X|M)) = E\left(M \frac{E(MX)}{E(M^2)}M\right) = E(MX) \quad (8)$$

Payoffs and Returns. We have reviewed three frameworks that lead to the Euler equation. This equation defines the asset price P for any asset. For stocks, the payoff X_{t+1} is the price next period P_{t+1} and the dividend D_{t+1} . For a one-period bond, the payoff is 1: one buys a bond at price P_t and receive 1 dollar next period. Alternatively, we can write the Euler equation in terms of returns. For stocks, returns are payoffs divided by prices: $R_{t+1} = X_{t+1}/P_{t+1}$. For bonds, one pays 1 dollar today and receives R_{t+1} dollars tomorrow. In any case, the Euler equation in terms of returns is thus

$$E_t[M_{t+1}R_{t+1}] = 1 \quad (9)$$

The Euler equation naturally applies to a risk-free asset. If one pays 1 dollar today and receives R_t^f dollars tomorrow for sure, the risk-free rate R_t^f satisfies

$$R_t^f = 1/E_t[M_{t+1}] \quad (10)$$

Expected Excess Returns

Definition of Risk Premia. Applying the definition of the covariance to the Euler equation (9) for the asset return R^i leads to $E_t(M_{t+1})E_t(R_{t+1}^i) + \text{cov}_t[M_{t+1}, R_{t+1}^i] = 1$. Using the definition of the risk-free rate in equation (10), we obtain

$$E_t(R_{t+1}^i) - R_t^f = -R_t^f \text{cov}_t[M_{t+1}, R_{t+1}^i] \quad (11)$$

The left-hand side of equation (11) defines the expected excess return. The right-hand side of equation (11) defines the risk premium. When the asset return R^i is negatively correlated to the SDF, the investor expects a positive excess return on asset

i. All assets have an expected return equal to the risk-free rate, plus a risk adjustment that is positive or negative.

To gain some intuition on the definition above, let us consider the case of preference-based SDFs. Assume that utility increases, and marginal utility decreases with consumption; this is the consumption-capital asset pricing model (consumption-CAPM). Here, the SDF—also known as *intertemporal marginal rate of substitution*—is the ratio of marginal utility of consumption tomorrow divided by the marginal utility of consumption today. Substituting the SDF into equation (11), we obtain

$$E_t(R_{t+1}^i) - R_t^f = -R_t^f \frac{\text{Cov}_t[\beta u'(C_{t+1}), R_{t+1}^i]}{u'(C_t)} \quad (12)$$

Marginal utility $u'(C)$ declines as consumption C rises. Thus, an asset's expected excess return is positive if its return covaries positively with consumption. The reason for this can be explained as follows. Our assumption on the investors' utility function implies that investors dislike uncertainty about consumption. An asset whose return covaries positively with consumption pays off well when the investor is already feeling wealthy and it pays off badly when he/she is already feeling poor. Thus, such an asset will make the investor's consumption stream more volatile. As a result, assets whose returns covary positively with consumption make consumption more volatile, and so must promise higher expected returns to induce investors to hold them.

Beta-representation and Market Price of Risk. We can rewrite the right-hand side of equation (11) as

$$\underbrace{\left(-\frac{\text{Cov}_t[M_{t+1}, R_{t+1}^i]}{\text{Var}_t[M_{t+1}]} \right)}_{\beta_{i,M}} \underbrace{\left(\frac{\text{Var}_t[M_{t+1}]}{E_t[M_{t+1}]} \right)}_{\lambda_M} \quad (13)$$

$E(R_{t+1}^i) - R_t^f = \beta_{i,M} \lambda_M$ is then a beta-representation of the Euler equation. Note that λ_M is independent of the asset i . It is called the *market price of risk*. $\beta_{i,M}$ is the quantity of risk. The expected excess return on asset i is equal to the quantity of risk of this asset times the price of risk.

Euler Equation with Log Returns and Log SDF. To interpret risk premia, it is often easier to rewrite

the previous results in terms of the log SDF m_{t+1} and log return r_{t+1}^i . Assuming that SDF and returns are lognormal, equation (9) leads to

$$E_t(m_{t+1}) + E_t(r_{t+1}^i) + \frac{1}{2} \text{Var}_t(m_{t+1}) + \frac{1}{2} \text{Var}_t(r_{t+1}^i) + \text{Cov}_t(m_{t+1}, r_{t+1}^i) = 0 \quad (14)$$

where lowercase letters denote logs. The same equation holds for the risk-free rate r_t^f . Let $\tilde{r}_{t+1}^{e,i}$ be the excess return corrected for the Jensen term: $\tilde{r}_{t+1}^{e,i} = r_{t+1}^i - r_t^f + \frac{1}{2} \text{Var}_t(r_{t+1}^i)$. Then, the expected log excess return is equal to

$$E_t(\tilde{r}_{t+1}^{e,i}) = -\text{Cov}_t(m_{t+1}, \tilde{r}_{t+1}^{e,i}) \quad (15)$$

For the consumption-CAPM, the utility each period is $u(C) = C^{1-\gamma}/(1-\gamma)$. The log SDF depends only on consumption growth and is equal to $m_{t+1} = \log \beta - \gamma g - \gamma(\Delta c_{t+1} - g)$, where g is the average consumption growth. In this case, the expected excess return is equal to

$$E_t(\tilde{r}_{t+1}^{e,i}) = \gamma \text{Cov}_t(\Delta c_{t+1} - g, \tilde{r}_{t+1}^{e,i}) \quad (16)$$

Again, assets whose returns covary positively with consumption must promise positive expected returns to induce investors to hold them.

Empirical Evidence

Now the empirical stylized facts on risk premia are discussed. A large literature shows that, in many asset markets, expected excess returns are sizable and time-varying. The equity, bond, and currency markets are considered (*see Predictability of Asset Prices*).

Stock Markets

Evidence of large risk premia abound on equity markets. The size of the average excess return on the stock market is actually puzzling from a consumption-based asset pricing perspective; it constitutes the equity premium puzzle. Moreover, expected equity returns appear time-varying.

Equity Premium Puzzle. To understand the equity premium puzzle, let us first define the Sharpe ratio.

Definition 4 The Sharpe ratio SR measures how much return the investor receives per unit of volatility:

$$SR = \frac{E(R^i) - R^f}{\sigma(R^i)} \quad (17)$$

where $\sigma(R^i)$ denotes the standard deviation of the return R^i .

Over the period 1927–2006 in the United States, real excess returns on the New York Stock Exchange (NYSE) stock index have averaged 8%, with a standard deviation of 20%, and thus the Sharpe ratio has been about 0.4. Starting from equation (11) and using the fact that correlations are below unity, the Sharpe ratio is linked to the first and second moments of SDFs:

$$\left| \frac{E(R^i) - R^f}{\sigma(R^i)} \right| \leq \left| \frac{\sigma(M)}{E(M)} \right| \quad (18)$$

Now, recall the consumption-CAPM and assume that consumption is lognormal. Then, the right-hand side is approximately

$$\frac{\sigma(M)}{E(M)} = \sqrt{e^{\gamma^2 \sigma_{\Delta c}^2} - 1} \approx \gamma \sigma_{\Delta c} \quad (19)$$

Aggregate nondurable and services consumption growth has a mean of 2% and a standard deviation of 1%, implying a risk-aversion coefficient of 40! If we take into account the low correlation between consumption growth rates and market returns, the implied risk aversion is even higher. This is the equity premium puzzle of Mehra and Prescott [16]. Such a high risk-aversion coefficient implies implausibly high risk-free rates. This is the risk-free rate puzzle of Weil [19]. The above-mentioned evidence is based on *realized* excess returns. Yet similar results are obtained with *expected* excess returns, which turn out to be large and time-varying.

Time-varying Expected Excess Returns. The Campbell and Shiller [5] decomposition of stock returns frames the evidence on stock market predictability. To see this, start from $1 = R_{t+1}^{-1} R_{t+1} = R_{t+1}^{-1} (P_{t+1} + D_{t+1}) / P_t$, and multiply both sides by the price-dividend ratio P_t / D_t to obtain

$$\frac{P_t}{D_t} = R_{t+1}^{-1} \left(1 + \frac{P_{t+1}}{D_{t+1}} \right) \frac{D_{t+1}}{D_t} \quad (20)$$

Taking logs leads to

$$p_t - d_t = -r_{t+1} + \Delta d_{t+1} + \log(1 + e^{p_{t+1} - d_{t+1}}) \quad (21)$$

A first-order Taylor approximation of the last term around the mean price-dividend ratio P/D gives

$$p_t - d_t = -r_{t+1} + \Delta d_{t+1} + k + \rho(p_{t+1} - d_{t+1}) \quad (22)$$

where $k = \log(1 + P/D)$ and $\rho = (P/D)/(1 + P/D)$. Iterating forward and assuming that $\lim_{j \rightarrow \infty} \rho^j(p_{t+j} - d_{t+j}) = 0$, one obtains

$$p_t - d_t = \text{Constant} + \sum_{j=1}^{\infty} \rho^{j-1} (\Delta d_{t+j} - r_{t+j}) \quad (23)$$

This equation holds *ex-post*, and thus also *ex-ante*:

$$p_t - d_t = \text{Constant} + E_t \sum_{j=1}^{\infty} \rho^{j-1} (\Delta d_{t+j} - r_{t+j}) \quad (24)$$

Now multiply both sides by $p_t - d_t - E(p_t - d_t)$. Then the variance of the log price-dividend ratio is

$$\begin{aligned} & \text{cov} \left(p_t - d_t, \sum_{j=1}^{\infty} \rho^{j-1} \Delta d_{t+j} \right) \\ & - \text{cov} \left(p_t - d_t, \sum_{j=1}^{\infty} \rho^{j-1} r_{t+j} \right) \end{aligned} \quad (25)$$

The fact that the price-dividend ratio varies means that either dividend growth rates or returns must be forecastable. The question is: which one is forecastable? Long-horizon regressions show little predictability in dividend growth rates and some predictability in returns and excess returns (Table 1).

We have seen that the aggregate stock market offers evidence of sizable and time-varying risk premia. Many subsets of the market offer comparable results. For example, Fama and French [12] sort stocks along different dimensions (e.g., their market size, book-to-market ratios, or past returns), build the corresponding portfolios and obtain large cross sections of returns. Buying the stocks in the last portfolio and selling the ones in the first portfolio lead to large and predictable excess returns, and thus evidence of equity risk premia.

Table 1 Long-horizon stock market predictability tests

Horizon h	Excess returns			Dividend growth		
	α	s.e.	R^2	α	s.e.	R^2
1	3.77	1.38	0.07	-0.11	1.00	0.00
2	7.46	2.36	0.12	-0.76	0.86	0.01
3	12.07	3.70	0.18	0.12	0.98	0.00
4	17.62	5.27	0.24	0.41	1.26	0.00
5	22.01	5.66	0.29	0.03	0.89	0.00

This table reports slope coefficients α , standard errors s.e. and R^2 from in-sample predictability tests. In the left panel, the univariate regressions are $R_{t,t+h}^e = C + \alpha D_t/P_t + \varepsilon_{t+h}$, where $R_{t,t+h}^e$ denotes the h -year ahead stock market excess return and D_t/P_t the dividend-price ratio. In the right panel, the regressions are $D_{t+h}/D_t = C + \alpha D_t/P_t + \varepsilon_{t+h}$, where D_{t+h}/D_t denotes the h -year ahead dividend growth rate. The sample relates to the period 1927–2006. Data are annual

Bond Markets

Equivalent results are obtained on bond markets, where expected excess returns exist and are time-varying. These results contradict the usual “expectation hypothesis of the term structure”. It is reviewed, followed by the empirical evidence on bond excess returns (see **Expectations Hypothesis**).

The expectation hypothesis can be defined in three equivalent ways:

- The yield y_t^n of a bond with maturity n is equal to the average of the expected yields of future one-year bonds, up to a constant risk premium:

$$y_t^n = \frac{1}{n} E_t(y_t^1 + y_{t+1}^1 + \cdots + y_{t+n-1}^1) \quad (26)$$

- The forward rate equals the expected future spot rate, up to a constant risk premium:

$$f_t^{n \rightarrow n+1} = E_t(y_{t+n}^1) \quad (27)$$

- The expected holding-period return (defined as the return on buying a bond of a given maturity n and selling it in the next period) is the same for any maturity n , up to a constant risk premium:

$$E_t(hpr_{t+1}^n) = y_t^1, \forall n \quad (28)$$

where $hpr_{t+1}^n = p_{t+1}^{n-1} - p_t^n$ denotes the log holding-period return and p_t^n the log price of a bond of maturity n .^a

Following Campbell and Shiller [6], the expectation hypothesis is often tested with the following

Table 2 Expectation hypothesis tests

$n = 2$	$n = 3$	$n = 4$	$n = 5$
-0.88 [0.47]	-1.46 [0.48]	-1.62 [0.53]	-1.70 [0.61]

This table reports slope coefficients β_n and associated standard errors from the following univariate regressions: $y_{t+1}^{n-1} - y_t^n = \alpha + \beta_n \left(\frac{y_t^n - y_t^1}{n-1} \right) + \varepsilon_{t+1}$, where y_t^n denotes the n -year bond yield. The sample relates to the period 1952–2006. Data are annual

equation:

$$y_{t+1}^{n-1} - y_t^n = \alpha + \beta_n \left(\frac{y_t^n - y_t^1}{n-1} \right) + \varepsilon_{t+1} \quad (29)$$

The expectation hypothesis implies that $\beta_n = 1$. In the data, the slope coefficient β_n is significantly below 1, often negative, and decreasing with the horizon n (Table 2). The rejection of the expectation hypothesis implies that bond markets offer time-varying, expected excess returns.

Currency Markets

Risk premia are also prevalent on currency markets. Currency excess returns correspond to the following investment strategy: borrowing in the domestic currency, exchanging this amount for some foreign currency, lending abroad, and converting back the earnings into the domestic currency. According to the standard uncovered interest rate parity (UIP) condition, the expected change in exchange rate should be equal to the interest rate differential between foreign

and domestic risk-free bonds. In this case, expected currency excess returns should be zero. However, the UIP condition is clearly rejected in the data. In a simple regression of exchange rate changes on interest rate differentials, UIP predicts a slope coefficient of 1. Instead, empirical work following Hansen and Hodrick [13] and Fama [11] consistently reveals a regression coefficient that is smaller than 1 and very often negative. The international economics literature refers to these negative UIP slope coefficients as the UIP puzzle or forward premium anomaly. Negative slope coefficients mean that currencies with higher than average interest rates actually tend to appreciate. Investors in foreign one-period discount bonds thus earn the interest rate spread, which is known at the time of their investment, plus the bonus from the appreciation of the currency during the holding period. As a result, the failure of the UIP condition implies positive predictable excess returns when investing in high interest rate currencies and negative excess returns for investing in low interest rate currencies. Lustig and Verdelhan [15] build portfolios of currency excess returns by sorting currencies on their interest rate differentials with the United States. They obtain a large cross section of currency excess returns and show that these excess returns compensate the US investor for bearing US aggregate macroeconomic risk because high interest rate currencies tend to depreciate in bad times. As a result, currency excess returns are also evidence of risk premia.

To summarize this section, equity, bond, and currency markets offer predictable excess returns, and are thus characterized by risk premia. Now the potential theoretical explanations of these risk premia are discussed.

Theoretical Interpretations

As observed above, the consumption-CAPM (also known as *power utility*) can replicate average equity excess returns only with implausibly high risk-aversion coefficients. Moreover, if consumption growth shocks are close to independent and identically distributed (i.i.d.)—as they are in the data—this model does not explain time variations in expected excess returns. A large literature seeks to address these shortcomings and offers different interpretations of the observed risk premia. Now the three most

successful classes of models in this literature, namely, habit preferences, long-run risk, and disaster risk, are reviewed.

Habit Preferences

Habit preferences assume that the agent does not care about the absolute level of his/her consumption, but cares about its relative level compared to a habit level that can be interpreted as a subsistence level, past consumption, or the neighbors' consumption. Hence, preferences over habits H are defined using ratios or differences (C/H or $C - H$), where H depends on past consumption: $H_t = f(C_{t-1}, C_{t-2}, \dots)$. Major examples of habit preferences are found in Abel [1], Campbell and Cochrane [4], Constantinides [8] and Sundaresan [18]. Preferences defined using differences between consumption and habit (e.g., $u(C) = (C - H)^{-\gamma}$) imply time-varying risk-aversion coefficient if the percentage gap between consumption and habit changes through time:

$$\gamma_t = -\frac{CU_{CC}}{U_C} = \frac{\gamma H_t}{C_t - H_t} \quad (30)$$

Campbell and Cochrane [4] propose a model along these lines. In their model, the habit level is slow moving; in bad times, consumption falls close to the habit level, and the investor is very risk averse. This model offers a new interpretation to risk premia: investors fear bad returns and wealth loss because they tend to happen in recessions, when consumption falls relative to its recent past. These preferences generate many interesting asset pricing features: pro-cyclical variations of stock prices, long-horizon predictability, countercyclical variation of stock market volatility, countercyclicity of the Sharpe ratio, and the short- and long-run equity premium.

Long-run Risk

The long-run risk literature works off the class of preferences due to Epstein and Zin [9, 10] and Kreps and Porteus [14]. These preferences impute a concern for the timing of the resolution of uncertainty to agents, and the risk-aversion coefficient is no longer the inverse of the intertemporal elasticity of substitution as it is with the consumption-CAPM (*see Recursive Preferences*). Building on these preferences,

Bansal and Yaron [2] propose a model where the consumption and dividend growth processes contain a low-frequency component and are heteroscedastic. These two features capture time-varying growth rates and time-varying economic uncertainty. Because this low-frequency component is persistent, a high value today signals high expected consumption growth in the future. If the intertemporal elasticity of substitution is above 1, then, in response to higher expected growth, agents buy more assets, and the price to consumption ratio rises: the intertemporal substitution effect dominates the wealth effect. In this case, asset prices are high in good times and low in bad times; thus, investors require risk premia. In this model, agents have preference for early resolution of uncertainty, which increases the risk compensation for long-run growth and uncertainty risks.

Disaster Risk

In the disaster risk literature, the agent is characterized by the usual constant relative risk-aversion preferences. Rietz [17] assumes that in each period a small-probability disaster may occur, and in this case, consumption and dividends drop sharply. Barro [3] calibrates disaster probabilities from the twentieth-century global history and shows that they are consistent with the high equity premium, low risk-free rate, and volatile stock returns. In this model, risk premia exist because investors fear rare economic disasters.

Conclusion

Under a minimal set of assumptions, any return satisfies a simple Euler equation. This equation implies that expected returns in excess of the risk-free rate, that is, risk premia, exist because returns comove with aggregate factors that matter for the investor. Empirical evidence from the equity, bond, and currency markets points to large and time-varying predictable excess returns. A recent literature tries to replicate and interpret these risk premia as compensations for recession, long-run, or disaster risks.

Acknowledgments

I owe a great part of my knowledge on risk premia to John Cochrane and to his book on “Asset Pricing”, which has inspired large parts of this article.

End Notes

^aRecall that the yield y_t^n of an n -year bond is a fraction of the log price p_t^n of the bond: $y_t^n = -\frac{1}{n}p_t^n$.

References

- [1] Abel, A.B. (1990). Asset prices under habit formation and catching up with the Joneses, *American Economic Review* **80**(2), 38–42.
- [2] Bansal, R. & Yaron, A. (2004). Risks for the long run: a potential resolution of asset pricing puzzles, *The Journal of Finance* **59**, 1481–1509.
- [3] Barro, R. (2006). Rare disasters and asset markets in the twentieth century, *Quarterly Journal of Economics* **121**, 823–866.
- [4] Campbell, J.Y. & Cochrane, J.H. (1999). By force of habit: a consumption-based explanation of aggregate stock market behavior, *Journal of Political Economy* **107**(2), 205–251.
- [5] Campbell, J.Y. & Shiller, R.J. (1988). The dividend-price ratio and expectations of future dividends and discount factors, *Review of Financial Studies* **1**, 195–228.
- [6] Campbell, J.Y. & Shiller, R.J. (1991). Yield spreads and interest rates: a bird’s eye view, *Review of Economic Studies* **58**, 495–514.
- [7] Cochrane, J.H. (2001). *Asset Pricing*. Princeton University Press, Princeton, NJ.
- [8] Constantinides, G.M. (1990). Habit formation: a resolution of the equity premium puzzle, *The Journal of Political Economy* **98**, 519–543.
- [9] Epstein, L.G. & Zin, S. (1989). Substitution, risk aversion and the temporal behavior of consumption and asset returns: a theoretical framework, *Econometrica* **57**, 937–969.
- [10] Epstein, L.G. & Zin, S. (1991). Substitution, risk aversion and the temporal behavior of consumption and asset returns, *Journal of Political Economy* **99**(6), 263–286.
- [11] Fama, E. (1984). Forward and spot exchange rates, *Journal of Monetary Economics* **14**, 319–338.
- [12] Fama, E.F. & French, K.R. (1992). The cross-section of expected stock returns, *Journal of Finance* **47**(2), 427–465.
- [13] Hansen, L.P. & Hodrick, R.J. (1980). Forward exchange rates as optimal predictors of future spot rates: an econometric analysis, *Journal of Political Economy* **88**(5), 829–853.

- [14] Kreps, D. & Porteus, E.L. (1978). Temporal resolution of uncertainty and dynamic choice theory, *Econometrica* **46**, 185–200.
- [15] Lustig, H. & Verdelhan, A. (2007). The cross-section of foreign currency risk premia and consumption growth risk, *American Economic Review* **97**(1), 89–117.
- [16] Mehra, R. & Prescott, E. (1985). The equity premium: a puzzle, *Journal of Monetary Economics* **15**(2), 145–161.
- [17] Rietz, T.A. (1988). The equity risk premium: a solution, *Journal of Monetary Economics* **22**, 117–131.
- [18] Sundareshan, S. (1989). Intertemporal dependent preferences and the volatility of consumption and wealth, *The Review of Financial Studies* **2**(1), 73–88.
- [19] Weil, P. (1989). The equity premium puzzle and the risk-free rate puzzle, *Journal of Monetary Economics* **24**, 401–424.

Related Articles

Arbitrage Pricing Theory; Capital Asset Pricing Model; Stochastic Discount Factors; Utility Function.

ADRIEN VERDELHAN

Predictability of Asset Prices

Predictability can be interpreted in many ways in finance. The fundamental issue in asset pricing is to determine the relationship between risk and reward. To quantify such a relationship, an economic model is built to “predict” how the expected asset returns should vary with their risk measures. In this case, predictability means contemporaneous association between the expected return of an asset and the expected returns of different risk factors. For example, the capital asset pricing model (CAPM) predicts that a security’s expected risk premium is proportional to the expected return from the market factor, where the proportionality reflects the systematic risk measure. This type of predictability is not the focus of this article. Instead, the focus is on whether future security returns can be predicted from current known information.

One important assumption used to build a rational asset pricing model is the market efficiency (*see Efficient Market Hypothesis*), in which security prices reflect all available information quickly and fairly. This was interpreted literally in the 1950s and 1960s as saying that any lagged variables possess no power in predicting current or future security prices or returns. The modern finance theory, however, has a different interpretation for the evidence of return predictability. In fact, researchers have recognized since 1980s that the expected returns can vary over time due to changes in investors’ risk tolerance and/or investment opportunities [30] over business cycles. If business cycles are predictable to some degree, returns can also be predictable, which poses no challenge to the efficient market hypothesis (EMH). Under this view, one should not rely solely on the historical average returns to estimate expected returns in assisting our investment decisions. In other words, the task of estimating the expected returns precisely largely depends on our ability to predict future stock returns.

Given the fact that the serial correlations for aggregate stock returns are weak especially in the recent decade, the quest for additional predictors goes on. Many financial variables have been shown to possess predictive power for stock returns. A partial list of these variables can be characterized as variables that

are related to interest rates: relative interest rate [7], term spread and the default spread [7, 16, 23], inflation rate [14, 18]; variables that are related to “one over the price”: dividend yield [10], payout yield [4], earning–price ratio and dividend–earnings (payout) ratio [26], book-to-market ratio [25, 32]; and other variables including aggregate net issuing activity [2] and consumption–wealth–income ratio [27].

Although the focus is on the rational explanation for predictability, the evidence has also been interpreted differently under different views. Their differences are illustrated by the following story. Once there were four students walking on a street with their professor. A dollar bill lying on the sidewalk quickly caught the professor’s eyes. The professor asked the four students why nobody was picking up the dollar bill. The first student answered although the dollar bill was real, people just pretended not seeing it. The second student argued that the dollar bill was just an illusion (or a statistical illusion). The third student said that, even though the dollar bill was real, no one would bother to pick it up because it was too costly to pick it up (or transactions costs). The last student’s answer was that the dollar bill was real. Someone left it there for a needy person. Generally speaking, the first student is a behaviorist; the second and third students hold the traditional efficient market view; and the last student holds the modern view on the EMH. No matter which student’s answer represents your view, predictability cannot be too large. There is an old saying: if you can predict the market, why aren’t you rich!

The existence of predictability is crucial in testing the conditional asset pricing models [19], in return decomposition [8], in asset allocation [22], and so on. Because of the theoretical foundation for predictability, this article focuses primarily on aggregate market returns. Predictability is also related to anomalies. An anomaly is defined as the deviation from an asset pricing model. In most empirical studies, anomalies are tied to a specific part of the market, such as small firms, firms with low book-to-market ratios, and so on, or particular sample periods, such as January, weekends, and so on. A detailed review on anomalies can be found in [35].

This article intends to offer a perspective on both the evidence and the reasons for return predictability. A detailed discussion about the economic reasons for predictability is given in the section Economic Interpretation of Predictability. Recent empirical studies

have uncovered many useful predictors, which are summarized in the section Understanding Some Useful Predictors. Predictability is not without controversy. Many of the statistical issues in testing the predictability are discussed in the section Statistical Issues, followed by conclusion in the last section.

Evidence on Predictability

The most simple form of predictability is the return autocorrelation. To gain a perspective on the magnitude of the serial correlation, returns of different frequencies and over different sample periods are examined. Owing to the availability of daily returns, the whole sample period is from 1962 to 2006. The summary statistics is listed in Table 1 for both value-weighted and equal-weighted NYSE/AMEX/NASDAQ composite index returns.

For the whole sample period, the average value-weighted index daily return is 0.044% with a volatility of 0.859%. Such a large difference between average return and volatility implies a very low Sharpe ratio of 5%. If returns are autocorrelated, the “true” Sharpe ratio should be larger.^a For the value-weighted index returns, the autocorrelation is about 13%. Such a large autocorrelation further increases to 31% when an equal-weighted index is used. If we fit an $AR(1)$ model to the equal-weighted index returns, we see an R^2 of 9.61%! The autocorrelation difference in the two types of index returns suggests that small stocks are more predictable than large stocks. To see whether such a predictability is stable over time, the whole sample period is split into two. From Table 1,

it is clear that most of the predictability from past returns concentrates in the early sample period from 1962 to 1984, with autocorrelations as high as 22.4 and 38.5% for value-weighted and equal-weighted indices, respectively.

Predictability in daily returns might be subject to market microstructure effects discussed in the section The Economic Interpretation of Predictability. One way to alleviate such effects is to examine the behavior of monthly returns. For both value- and equal-weighted index returns, the autocorrelations have been substantially attenuated. For example, over the whole sample period, autocorrelation for value-weighted index returns is only 4.3%, almost negligible. For the equal-weighted index, however, the autocorrelation is still as large as 17.6% for the whole sample period and is stable over the two subsample periods. Therefore, it can be concluded that return serial correlations are more likely to occur in small stocks. Given there are still substantial serial correlations in low-frequency small stock return data, market microstructure effects cannot be the only factor.

If future returns can only be weakly predicted by past returns, are there other variables that help to predict returns? In Table 2, we further study return predictability using three other variables—the dividend yield, the repurchasing yield, and the relative interest rate. Our sample starts in 1952 after a major shift in the interest rate regime by the Federal Reserve. To be representative, we focus on the value-weighted index returns. During the first 17 years from 1952 to 1978, both the dividend yield and the relative interest rate

Table 1 Autocorrelations in index returns

Sample period	Value weighted			Equal weighted		
	Mean	SD	Corr.	Mean	SD	Corr.
Panel A: daily returns						
1962–2006	0.044	0.859	13.2	0.069	0.744	31.0
1962–1984	0.035	0.794	22.4	0.068	0.787	38.5
1985–2006	0.053	0.922	6.1	0.071	0.696	21.0
Panel A: monthly returns						
1962–2006	0.929	4.216	4.3	1.186	5.345	17.6
1962–1984	0.772	4.422	6.0	1.285	6.252	16.4
1985–2006	1.079	4.010	1.9	1.092	4.312	20.0

This table reports the characteristics of NYSE/AMEX/NASDAQ composite index returns over different samples periods and for different frequencies. “Corr.” stands for the first-order autocorrelation; “SD” is the standard deviation

Table 2 VAR results for index returns

Dependent variable	r_t	$(D/P)_t$	$(F/P)_t$	$rrel_t$	Adjusted R^2
Panel A: sample period 1952–1978					
r_{t+1}	0.061	10.90	0.675	−11.67	0.062
$(D/P)_{t+1}$	−0.000	0.966	0.003	0.042	0.956
$(F/P)_{t+1}$	−0.001	0.038	0.943	0.034	0.898
$rrel_{t+1}$	0.000	0.032	0.005	0.731	0.529
Panel A: sample period 1979–2005					
r_{t+1}	0.030	0.461	3.508	−0.801	0.009
$(D/P)_{t+1}$	−0.000	0.994	−0.009	0.005	0.985
$(F/P)_{t+1}$	−0.001	0.029	0.971	0.071	0.960
$rrel_{t+1}$	−0.000	0.009	−0.010	0.751	0.560

This table reports the VAR results for the four variables including the value-weighted NYSE/AMEX/NASDAQ composite index return, dividend yield, repurchasing yield, and the relative interest rate over different sample periods. The bold face number indicates that the estimate is statistically significant at a 5% level

have helped to predict returns, with an adjusted R^2 of 6.2%. In contrast, the repurchasing yield becomes more important over the next 17 years from 1979 to 2005, with an adjusted R^2 of 0.9%. The evidence suggests that returns are predictable even if not by their past returns. Despite large persistence of all three predictors as shown in Table 2, statistical adjustment for estimates will not likely take away the predictive power of the three variables (see the section Statistical Issues).

Predictability and Market Efficiency

Historically, predictability has been associated with market inefficiency. According to the fundamental law of valuation, a security price should reflect its expected fundamental value for risk-neutral investors with zero interest rate:

$$P_t = E[V^*|\mathcal{I}_t], \quad P_{t+1} = E[V^*|\mathcal{I}_{t+1}] \quad (1)$$

where V^* is the fundamental value and $\mathcal{I}_t$ is the information set at time t . Since the information set $\mathcal{I}_t$ is included in the information set $\mathcal{I}_{t+1}$, the following result is obtained by the law of iterated expectations:

$$P_t = E[V^*|\mathcal{I}_t] = E[E(V^*|\mathcal{I}_{t+1})|\mathcal{I}_t] = E[P_{t+1}|\mathcal{I}_t] \quad (2)$$

Equation (2) suggests that security prices should follow a Martingale process.^b The best predictor for

future prices is the current price. In other words, we have

$$\text{Cov}[(P_{t+j} - P_{t+i}), (P_{t+l} - P_{t+k})|\mathcal{I}_t] = 0 \quad (3)$$

where $i < j < k < l$. In other words, the nonoverlapping price changes are uncorrelated at all leads and lags. If we interpret the price difference as a return, it means that returns should be unpredictable.

This analysis defines the notion of EMH. Financial markets are said to be efficient if security prices *rapidly* reflect *all* relevant information about asset values, and all securities are *fairly priced* in light of the available information. In other words, the EMH describes how security prices should react to available information and how prices should evolve over time. Under this framework, return predictability serves as evidence against the EMH.

Does the EMH indeed exclude predictability? To answer this question, we focus on a stronger version of the Martingale process, which is the random walk process, and assume that investors are risk averse. The random walk process was first used by Bachelier (1900) to model stock prices in his dissertation, and was rekindled by Merton in the late 1960s. For convenience, we use log price p_t

$$p_{t+1} = \mu + p_t + \epsilon_{t+1} \quad (4)$$

where μ is the expected price change. If we define return as $r_{t+1} = p_{t+1} - p_t$, equation (4) can be expressed as

$$r_{t+1} = \mu + \epsilon_{t+1} \quad (5)$$

Strictly speaking, the EMH only puts a restriction on the residual ϵ_{t+1} to satisfy the condition of $E[\epsilon_{t+1}|I_t] = 0$ at any time t in either equation (4) or (5). Since μ is determined by an asset pricing model, such as the CAPM, the traditional view on the *EMH* implicitly assumes that μ is constant. The modern finance theory, however, has offered a different view on μ . For example, Fama and French [17] have suggested that the risk premium might be higher in the economic downturn than in the peak of a business cycle. This evidence suggests that the expected return might be time varying. In fact, many asset pricing models since Merton have emphasized the idea of changing investment opportunities, which requires additional risk compensation over time. Alternatively, investors' risk tolerance might change over time, which will cause the investors to demand different levels of risk premium. No matter which scenario is more likely, one should allow μ to be time varying:

$$r_{t+1} = \mu_{t+1} + \epsilon_{t+1} \quad (6)$$

Although under the EMH, we still have the condition of $E[\epsilon_{t+1}|I_t] = 0$, $E[\mu_{t+1}|I_t]$ is not necessarily constant. For example, if risk premia changes with the business cycle and the business cycle is predictable, return should also be predictable. This analysis opens a channel for the predictability to coexist with the EMH.

Returns from a buy-and-hold strategy on the market portfolio correspond to returns for a representative investor. Predictability means that someone can implement a trading strategy that requires a full investment in some periods and a zero or a short position in other periods in order to earn higher returns than those from a buy-and-hold strategy. Clearly, this investment strategy cannot be implemented by the representative investor since he/she has to fully invest in the equity market. Although such a strategy will pay off in a long run, it is not without risk in short term. The success of this strategy depends on the degree of predictability. Therefore, predictability cannot be too large in order to prevent too many investors defecting from being representative investors.

The Economic Interpretation of Predictability

Without assuming irrationality and market inefficiency, how can we interpret predictability under the

traditional framework? Most explanations focus on market microstructure effects and transactions costs. This section reviews the bid-ask bounce, nonsynchronous trading, and transactions costs in explaining the return autocorrelation.

Bid-ask Bounce

Returns tend to be negatively autocorrelated in a short-run. One possible explanation is offered by Roll [34] from the perspective of bid and ask price differences. In the absence of information, sell orders and buy orders arrive with the same probabilities. In other words, a buy order is likely to follow a sell order, which results in a negative autocorrelation. In particular, let P_t^* be the fundamental value:

$$P_t = P_t^* + I_t(s/2) \quad (7)$$

$$I_t = \begin{cases} +1 & \text{if buy order with prob} = 0.5 \\ -1 & \text{if sell order with prob} = 0.5 \end{cases} \quad (8)$$

where s is the bid-ask spread. This implies a price change of $\Delta P_t = \Delta P_t^* + (I_t - I_{t-1})s/2$. In other words, autocorrelation is related to the spread s in the following way:

$$\text{Cov}(\Delta P_{t-1}, \Delta P_t) = -s^2/4 \quad (9)$$

Since the bid-ask spreads tend to be larger for small company stocks than for large stocks, autocorrelation will be stronger for small firms than for large stocks, other things being equal. Equation (9) can also be used to back out the implied bid-ask spread.

If the autocorrelation is due to differences in the bid and ask prices, the effect should be smaller if the average bid and ask prices are used to compute returns instead of the actual closing prices. Similarly, low-frequency returns, such as monthly returns, should have weaker autocorrelation than high-frequency returns, such as daily returns, which is true in general. We should also see a drop in autocorrelations over time when the average bid-ask spread shrinks, especially after decimalization. This is confirmed in Table 1. In general, investors cannot design a trading strategy to obtain excess returns in this case, since the bid and ask effect is due to the market friction.

Nonsynchronous Trading

Although individual stock returns might exhibit negative serial correlation, portfolio returns tend to be positively autocorrelated. Lo and MacKinlay [29] have offered “nonsynchronous trading” as a mechanism in generating such a positive autocorrelation. In practice, not all stocks, especially small stocks, are traded at any given moment. On the arrival of market wide news, these stocks that are not traded currently will have similar returns to those traded today when trading resumes next period, which will make the portfolio with all stocks look like autocorrelated.

To illustrate the idea, suppose that there are two stocks A and B following random walk processes, implying no autocorrelation in their own returns. At the release of a market wide news at time $t = 0$, we would have observed returns of R_1^A and R_1^B for stocks A and B , respectively. Owing to the commonality of the news, we assume $\text{Cov}(R_1^A, R_1^B) > 0$. If stock A is not traded and stock B is traded, however, we will only observe $\hat{R}_1^A = 0$ and $\hat{R}_1^B = R_1^B$. Similarly, there is a common news released at $t = 1$. Both stocks are traded this time, resulting in returns of R_2^A and R_2^B for the two stocks. Owing to the random walk assumption on individual stocks, we have $\text{Cov}(R_1^A, R_2^A) = 0$ and $\text{Cov}(R_1^B, R_2^B) = 0$. This structure can be summarized as follows:

$$\begin{array}{ccccccc} \text{Stock A :} & | & \hat{R}_1^A = 0 & | & \hat{R}_2^A = R_1^A + R_2^A & | & \\ \text{Stock B :} & | & \hat{R}_1^B = R_1^B & | & \hat{R}_2^B = R_2^B & | & \\ & & t = 0 & & t = 1 & & t = 2 \end{array} \quad (10)$$

Now, consider an equal-weighted portfolio of the two stocks. The portfolio returns in the two periods are $\hat{R}_1^P = \frac{1}{2}R_1^B$ and $\hat{R}_2^P = \frac{1}{2}(R_2^B + R_1^A + R_2^A)$. It is easy to see that $\text{Cov}(\hat{R}_1^P, \hat{R}_2^P) = \frac{1}{2}\text{Cov}(R_1^A, R_1^B) > 0$. The same idea applies to the case when different stocks are traded at different times. Using daily returns from 1962 to 1985 to form 20 size portfolios, Lo and MacKinlay reported the following first-order autocorrelations and the probability of nontrading (Table 3):

Clearly, a large autocorrelation of 35% in small stock portfolio returns can be supported by a 29% likelihood of nontrading in small stocks.^c However, it is difficult to justify the 17% autocorrelation in large stock portfolio returns by the corresponding likelihood of nontrading. In addition, the cross autocorrelation of 33% from the large stock portfolio to the small stock portfolio is also consistent with

Table 3 The probability of nontrading; adopted from Lo and Mackinlay [29]

$t t + 1$	Small	Medium	Large	Probability nontrading
Small	0.35	0.21	0.02	0.291
Medium	0.39	0.31	0.09	0.025
Large	0.33	0.36	0.17	0.008

the nontrading in small stocks. Under this view, no money can be made even with a positive return autocorrelation.

Transactions Costs

Nontrading is a bit of an econometrics device. Instead, security prices could be slow to update in the arrival of information due to transactions costs. In other words, transactions costs put a wedge in how prices might change over time. Let P_t^* be investors' correct valuation. They will trade only when the price cover the transactions costs. In other words, there will be a bound around the current price. Only when P_t^* accumulates to the degree that overcomes the bound, we will see a price change. Otherwise, there will be zero excess demand within the bound. Such a slow adjustment will create a positive autocorrelation in security returns.

Cross Autocorrelation

The evidence of large stock returns predicting small stock returns seems to be persuasive since it existed in both daily and monthly stock returns. Although one might attribute the phenomenon to the nonsynchronous trading story on the daily frequency, nontrading is less likely for monthly returns. In response, Boudoukh *et al.* [5] offered an alternative explanation that utilizes the serial correlation and the contemporaneous correlation to explain the cross correlation.

Suppose that security i 's return follows an $AR(1)$ process of the following form:

$$r_{i,t+1} = \mu + \theta r_{i,t} + \epsilon_{i,t+1} \quad (11)$$

It is easy to see that $\theta = \text{Corr}(r_{i,t+1}, r_{i,t})$. Multiplying both sides of equation (11) by $r_{j,t}$ and assuming that $\text{Cov}(\epsilon_{i,t+1}, r_{j,t}) = 0$, we have the following relation:

$$\text{Corr}(r_{i,t+1}, r_{j,t}) = \text{Corr}(r_{i,t+1}, r_{i,t})\text{Corr}(r_{i,t}, r_{j,t}) \quad (12)$$

Table 4 Portfolio correlations adopted from Boudoukh *et al.* [5]

Portfolio	Small _{t+1}	Medium _{t+1}	Large _{t+1}	Small _t	Large _t
Small _t	0.36	0.19	0.03		
Medium _t	0.35	0.22	0.06	0.89	
Large _t	0.28	0.21	0.07	0.72	0.91

This table reports cross- and auto-correlations among size portfolios

As seen from equation (12), the cross autocorrelation is essentially the self-autocorrelation acted on contemporaneous correlation. Using a different sample period, Boudoukh *et al.* [5] found that results are consistent with equation (12) (Table 4).

Applying equation (12), we can compute the predicted cross autocorrelations as

$$\begin{aligned}
& \text{Corr}^*(r_{\text{small},t+1}, r_{\text{large},t}) \\
&= \text{Corr}(r_{\text{small},t+1}, r_{\text{small},t}) \text{Corr}(r_{\text{small},t}, r_{\text{large},t}) \\
&= 0.36 \times 0.72 = 0.26
\end{aligned} \tag{13}$$

$$\begin{aligned}
& \text{Corr}^*(r_{\text{large},t+1}, r_{\text{small},t}) \\
&= \text{Corr}(r_{\text{large},t+1}, r_{\text{large},t}) \text{Corr}(r_{\text{large},t}, r_{\text{small},t}) \\
&= 0.07 \times 0.72 = 0.05
\end{aligned} \tag{14}$$

These numbers are very close to the actual cross autocorrelations shown in the table. Therefore, we do not need frequent nontrading to justify the observed cross autocorrelation. However, we still need to understand the serial correlation.

Time-varying Expected Returns

The mechanism for the observed autocorrelation, discussed in the previous sections, largely relies on market frictions. As discussed in the section Predictability and Market Efficiency, an alternative rational explanation for predictability is the time-varying expected return. Given the unobservability nature of expected returns, Conrad and Kaul [13] proposed to characterize the movement in expected returns as following a simple $AR(1)$ process of

$$r_{t+1} = E_t(r_{t+1}) + \epsilon_{t+1} \tag{15}$$

$$E_t(r_{t+1}) = \bar{r} + \phi E_{t-1}(r_t) + u_t \tag{16}$$

Note that the coefficients in equations (15) and (16) can be estimated using the Kalman filter procedure. Testing the hypothesis of time-varying expected return is equivalent to test whether $\phi = 0$ in the above models. Using the 10 size-sorted weekly (Wednesday to Tuesday) portfolio returns from 1962 to 1985, Conrad and Kaul [13] found that the autocorrelations coefficients are 41 and 9% for the small and the large decile portfolios, respectively, which are both statistically significant when compared to the confidence bound of $1/\sqrt{T} = 0.03$. Although the persistence parameter estimates ($\hat{\phi}$) of 0.589 and 0.087 for small and large portfolios, respectively, are very different, they are statistically significant at a 1% level.

It is important to understand why expected returns change over time. In the CAPM world, it is implicitly assumed that a firm will continue to produce the same widgets and face the same uncertainty when selling these widgets in the market. In other words, the risk structure in future cash flows (CFs) is fixed. Thus, the comovement with the overall market is fixed. At the same time, investors' attitude toward risk does not change, which implies constant expected returns. Such a model structure may reasonably describe the real world over a short period of time.

Over a longer horizon, however, investment opportunities can change due to either technological advances or changes in consumers' preference toward goods and services. For example, Apple used to be in the business of making personal computers and software 10 years ago. Today, a significant portion of Apple's business is in the consumer electronics including music players and cell phones. Under this view, both the risk environment of a firm and the risk tolerance of investors could change over time. Therefore, the observed predictability may simply provide compensation for investors' exposure to the risk of change in investment opportunities or reflect the differences in the required risk compensation due to change in the risk tolerance over different economic conditions. In this case, a representative investor will not try to utilize the predictability to alter his/her asset allocations. For example, if he/she knows that the next period stock return will likely be high, he/she should allocate more assets to stocks. However, if he/she understands that the high return is associated with high expected return due to his/her increased risk aversion next period,

he/she would not increase his/her holding of the risky stocks.

Understanding Some Useful Predictors

In an interesting paper by Boudoukh *et al.* [5], it is argued that the observed autocorrelation in returns neither is due to market inefficiency nor can be attributed to time-varying expected returns. If autocorrelation in the returns of an index, such as the *S&P500*, is due to market inefficiency or time-varying expected returns, the same autocorrelation should be observed in the *S&P500* future contract returns too, but they did not find supportive evidence.^d This result seems to kill the predictability associated with autocorrelation, but does not necessarily provide evidence against other form of predictability. Moreover, autocorrelation in returns is a sufficient condition for predictability. It is not a necessary condition. There could exist nonreturn-based predictors. Suppose that the return generating process is as follows:

$$r_{t+1} = \beta z_t + \epsilon_{t+1} \quad (17)$$

where z_t is a predictor with $\text{Cov}(z_t, \epsilon_{t+1}) = 0$ and $\text{Cov}(\epsilon_t, \epsilon_{t+1}) = 0$. Since $\text{Cov}(r_{t+1}, r_t) = \beta \text{Cov}(z_t, r_t) + \text{Cov}(\epsilon_{t+1}, r_t) \approx \beta \text{Cov}(z_t, r_t)$, autocorrelation could be close to zero as long as $\text{Cov}(z_t, r_t)$ is small, which is usually the case.

Therefore, recent literature has focused its attention on predictors other than past returns. For example, an incomplete list includes short-term interest rate, term spread, default spread, inflation rate, dividend yield, book-to-market ratio, consumption–wealth–income ratio, repurchasing yield, and so on. It is important to know why these variables predict returns in the first place. Without theory, a variable found to be useful in predicting stock returns could be a result of data mining.

A closer look at these predictors reveals that they are either related to business cycles or associated with stock prices. Since expected returns could vary with business cycles, variables that predict business cycles such as the term spread or the default spread should be useful predictors. Many significant predictors, such as the dividend yield, book-to-market ratio, and repurchasing yield, contain the element of one over price. This common feature comes from the fact

that security prices reflect investors' expectations, and expectations are good predictors of future values. To further illustrate this rationale, we can use mathematical models to relate returns to prices or other variables.

The Dividend–price Ratio—Log-linearization

Perhaps the most frequently used predictor is the dividend–price ratio or the dividend yield. This is also the variable that has been scrutinized the most [1]. Despite many statistical issues discussed in the following section, it is important to understand why the dividend–price ratio should predict future returns. We start from the following return identity:

$$R_{t+1} = \frac{P_{t+1} + D_{t+1}}{P_t} = \frac{P_{t+1}}{D_{t+1}} \frac{D_t}{P_t} \frac{D_{t+1}}{D_t} \left[1 + \frac{D_{t+1}}{P_{t+1}} \right] \quad (18)$$

It is difficult to allow time-varying expected return due to the nonlinearity in equation (18). We, thus, take natural log on both sides of equation (18) and apply Taylor series expansion around the steady state. After simplifying [8], we obtain the following equation:

$$d_t - p_t = \text{const} + \rho(d_{t+1} - p_{t+1}) + r_{t+1} - \Delta d_{t+1} \quad (19)$$

where $\rho = 1/(1 + D/P) = 1/1.04 = 0.96$ (with D/P being the steady-state dividend–price ratio), $\Delta d_{t+1} = d_{t+1} - d_t$, and lowercase variables represent log of the corresponding uppercase variables. Under the assumption of stationary dividend–price ratio, we can solve equation (19) forward,

$$d_t - p_t = \text{const} + \sum_{j=0}^{\infty} \rho^j (-\Delta d_{t+1+j} + r_{t+1+j}) \quad (20)$$

Equation (20) implies that a high dividend–price ratio must mean either a low future dividend growth or a high future return. In addition, dividends and returns that are closer to the present are more influential than dividends and returns far in the future due to the fact that ρ is less than 1.

From an empirical perspective, we can compute the volatility of dividend–price ratio by multiplying both sides of equation (20) by $(d_t - p_t)$

$$\begin{aligned} \text{Var}(d_t - p_t) = & -\text{Cov}\left(d_t - p_t, \sum_{j=0}^{\infty} \rho^j \Delta d_{t+1+j}\right) \\ & + \text{Cov}\left(d_t - p_t, \sum_{j=0}^{\infty} \rho^j r_{t+1+j}\right) \quad (21) \end{aligned}$$

Since the volatility of $(d_t - p_t)$ is positive, it is clear that the dividend–price ratio will forecast either dividend growth or future returns. Empirical evidence suggests that the $(d - p)$ variable does not forecast future dividend growth. Therefore, the $(d - p)$ ratio must forecast future returns. Again, such a predictability does not imply market inefficiency in predicting ϵ_{t+1} in equation (6). A testing strategy based on equation (21) is to regress the sum of future returns on the dividend–price ratio:

$$r_{t+1} + \dots + r_{t+\tau} = \alpha + \beta(\tau)(d_t - p_t) + \epsilon_{t+1,t+\tau} \quad (22)$$

where τ is the number of future periods. Table 5 is adapted from Campbell *et al.* [9] on monthly

NYSE/AMEX/NASDAQ composite index returns. Clearly, the degree of predictability measured by R^2 increases monotonically with the return horizon. For example, R^2 s are 0.7, 8.6, 21.7, and 41.9% over 1 month, 1 year, 2 years, and 4 years, respectively, in the early sample period of 1927–1951. Similarly, R^2 s continue to be impressive with 1.8, 18.8, 32.2, and 41.7% over 1 month, 1 year, 2 years, and 4 years, respectively, in the later sample period from 1952 to 1994.

Issues with Long-horizon Regressions

Long-horizon regressions, such as equation (22), were first advocated by Fama and French [15]. Despite the impressive magnitude of R^2 s, the power of the associated long-horizon tests is doubtful. The issue involves the use of overlapping observations due to the availability of data [33]. In general, overlapping samples could lead to large efficiency gains when the independent variables in a predictive regression are serially uncorrelated. However, most predictors are highly autocorrelated, which implies limited efficiency gains when using overlapping observations. For example, for the 60 years of returns used in the Fama and French [15] study, although the nominal sample size is large using overlapping returns, the effective sample size is not much larger than 12

Table 5 Long-horizon results for index returns

Forecast horizon	1 Month	3 Months	12 Months	24 Months	48 Months
Panel A: sample period 1927–1994					
$\beta(\tau)$	0.016	0.043	0.200	0.386	0.654
$t(\tau)$	1.553	1.420	2.257	4.115	3.870
$R^2(\tau)$	0.007	0.014	0.073	0.143	0.261
Panel A: sample period 1927–1951					
$\beta(\tau)$	0.024	0.054	0.304	0.667	1.085
$t(\tau)$	0.980	0.793	1.915	3.841	3.693
$R^2(\tau)$	0.007	0.011	0.086	0.217	0.419
Panel A: sample period 1951–1994					
$\beta(\tau)$	0.027	0.080	0.327	0.579	0.843
$t(\tau)$	3.118	3.152	3.181	3.072	3.508
$R^2(\tau)$	0.018	0.049	0.188	0.322	0.417

This table reports results from regressing future τ months returns on current dividend–price ratio for the value-weighted NYSE/AMEX/NASDAQ composite index return. The regression model is given as follows:

$$r_{t+1} + \dots + r_{t+\tau} = \alpha + \beta(\tau)(d_t - p_t) + \epsilon_{t+1,t+\tau}$$

(the nonoverlapping 5-year sample) due to the highly persistent regressor. As pointed out by Boudoukh *et al.* [6], if an innovation in an independent variable happens to coincide with the next period return, this relationship will be repeated many times in the long-horizon regression since the shock will not die out for many periods and the particular return will appear many times in the overlapping return series.

Under the null hypothesis of no autocorrelation in returns, that is, $\beta(\tau) = 0, \forall \tau$ in equation (22), Kirby [24] has shown the following asymptotic result for the R^2 from a predictive regression:

$$T \times R^2 \xrightarrow{d} \chi^2(K) \quad (23)$$

where T is the number of observations and K is the number of independent variables in the regression. Let us use a numerical example to illustrate the point. For a univariate regression with $K = 1$ and $T = 12$, what would we expect to see?

- Since the mean of a $\chi^2(1)$ random variable is 1, we have $E(12R^2) = 1$, which implies that $E(R^2) = 8.3\%$.
- The 95% cutoff for R^2 is expected to be 32% since the critical value for a $\chi^2(1)$ distribution under the same confidence level is 3.84. In other words, we can expect to see R^2 s as high as 32% even though there is no predictability.

Therefore, long-horizon regression results are avoided in this article.

The Payout Ratio or the Repurchasing Yield

Recently, Boudoukh *et al.* [4] have proposed to use the total payout ratio as a predictor for stock returns. The importance of the new predictor can be illustrated by its impressive R^2 of 26% using annual data over the sample period from 1926 to 2003. The use of payout ratio can be justified since investors' total wealth could also be affected by share repurchasing. In fact, representative investors should care about the total distribution, which includes both the direct dividend distributions and repurchases. If there are rational reasons to believe that dividend yield predicts stock returns, the payout yield should play a similar role. In fact, the implementation of the SEC rule 10b-18 in 1982 gives firms an incentive to rely more on repurchases due to the tax advantages for investors.

Payout Ratio. Repurchasing could be used either to reduce the effect of stock option exercise or to substitute for dividends. To construct a measure that reflects the latter, Boudoukh *et al.* [4] used change in treasury stock adjusted for potential asynchronicity between the repurchase and option exercise as a measure of repurchasing (TS). They also used total repurchasing from the CF statement. Results are summarized in Table 6. Clearly, the predictive power of the dividend-price ratio (D/P) has gone down when comparing R^2 s from the two sample periods. In particular, R^2 has decreased from 13 to 8% when including the recent sample period.

Table 6 Dividend yield and payout ratio

	$\ln(D/P)$	$\ln(CF/M)$	$\ln(TS/M)$	$\ln(0.1 + \text{Net payout})$
1926–2003				
Coef	0.116	0.209	0.172	0.759
t -ratio	2.240	3.396	2.854	5.311
R^2	0.055	0.091	0.080	0.262
Sim p -value	0.083	0.011	0.020	0.000
1926–1984				
Coef	0.296	0.280	0.300	0.794
t -ratio	3.666	3.688	3.741	5.342
R^2	0.130	0.121	0.135	0.300
Sim p -value	0.044	0.054	0.043	0.001

This table reports results from predictive regressions using various predictors. “CF/M” is the total repurchasing from cash flow statement (CF) over the total market value, while “TS/M” is change in treasury stock adjusted for potential asynchronicity between the repurchase and option exercise as a measure of repurchasing (TS) over the total market value. “ D/P ” is the usual dividend-price ratio for the value-weighted NYSE/AMEX/NASDAQ composite index

Table 7 The adjusted R^2 s for predictive regression using the dividend yield (D/P) and/or the repurchasing yield (F/P) over different sample periods

Frequency	1952–2005		1952–1978		1979–2005	
	D/P	F/P	D/P	F/P	D/P	F/P
Monthly	0.5	1.0	1.9	0.0	0.0	1.7
Quarterly	1.9	2.5	5.6	0.0	0.7	5.6

In contrast, the repurchasing yield (measured as the ratio between repurchasing and market capitalization) is impressive. No matter how it is measured, the explanatory power is much larger than the pure dividend–price ratio. Moreover, when using the net payout yield (measured as the ratio between repurchasing minus new issuing plus dividend and the market capitalization), R^2 is as high as 26%. Although the payout yield is important empirically, its significance is overstated in Boudoukh *et al.* [4]. The most significant contributor to the predictive power of the payout yield is the new issuing yield when examining their accounting-based measures separately. Furthermore, the predictive power of new issuing yield largely comes from the two outliers of 1929 and 1930. In other words, the new issuing yield offers no explanatory power once the sample period starts from 1931 instead of 1926.

An Alternative Approach to Construct the Repurchasing Yield. The conventional approach of computing returns ignores changes in the market capitalization associated with changes in the total number of share outstanding. When the number of shares changes over time due to either repurchasing or seasonal offering, capital gains do not purely reflect growth potential. From an asset pricing perspective, it is more important to consider different components of returns from a representative investor’s perspective. In other words, we can decompose returns from the stand point of a representative investor instead of a buy-and-hold investor. In particular, we can rewrite the return identity as the following:

$$R_{t+1} \equiv \frac{S_{t+1}D_{t+1}}{S_t P_t} + \frac{(S_t - S_{t+1})(P_{t+1} + D_{t+1})}{S_t P_t} + \frac{S_{t+1}P_{t+1}}{S_t P_t} \quad (24)$$

where S_t is the number of share outstanding at time t . Equation (24) can be interpreted in the following way:

- *first term*: dividend yield (D/P) for a representative shareholder;
- *second term*: net repurchasing yield (F/P) at the before ex-dividend day price; and
- *third term*: change in market capitalization, which reflects growth.

Using NYSE/AMEX/NASDAQ index returns, we can construct both the smoothed dividend yield (D/P) and the repurchasing yield (F/P) over the past 12 months. Table 7 reports the adjusted R^2 s for various predictive regressions.

For the whole sample period from 1952 to 2005, quarterly returns are more predictable than monthly returns. Overall, the repurchasing yield has higher predictive power than the dividend yield. When we split the sample into two, it becomes clear that the two predictors have played very different roles. Almost all the predictive power in the first half of the sample comes from the dividend yield, whereas majority predictive power in the second half of the sample is due to the repurchasing yield. This evidence is consistent with the observation of a decreasing trend in the dividend yield and the increasing role played by repurchases.

Statistical Issues

The use of many predictors can be controversial. In many cases, the issue lies in the statistical inference due to the persistence in predictors. These issues include spurious regression, biased estimates due to correlations between innovations to predictors and stock returns, and error in variables when using imperfect predictors.

Spurious Regression

When regressing one nonstationary random variable on another independent nonstationary random variable, we often observe a significant relationship between the two variables. This is because, in a finite sample, both variables are likely to be perceived trending. Spurious regression is first discussed by Granger and Newbold [21]. At a first glance, spurious regression may not seem to be likely for a predictive regression since stock returns on the left-hand side of a regression are not persistent at all. However, if we consider stock returns as containing a persistent expected return component, the predictive regression could be spurious [20]. This problem can even be more severe when researchers are mining for predictors because highly persistent series are more likely to be found significant in the search for predictors. Their simulation results suggest that many of the useful predictors found in the literature could be subject to this criticism.

Predictive Regression

Owing to persistence in the predictors and a correlation between innovations to predictors and stock returns, Stambaugh [36] has suggested that both the coefficient estimate and the t -ratio in a predictive regression are biased. For example, when the current stock price is high, the current return will also be high, whereas the current dividend-price ratios will be low since the D/P ratio has a price in the denominator. Such an association implies a negative relationship between innovations to D/P ratios and innovations to returns. Such a negative correlation will couple with the typical downside bias in the persistence parameter estimate of the D/P ratio to make the predictive regression coefficient bias upward. More specifically, suppose that we have the following system:

$$r_{t+1} = \beta z_t + \epsilon_{t+1} \quad (25)$$

$$z_{t+1} = \psi z_t + u_{t+1} \quad (26)$$

where z_t is the predictor. It can be shown [28] that

$$E(\hat{\beta} - \beta) = \gamma(\hat{\psi} - \psi) \quad (27)$$

where γ is from the regression of $\epsilon_t = \gamma u_t + v_t$. Since γ is negative in the case of dividend-price

ratio and $\hat{\psi}$ is typically biased downward, equation (27) suggests that the beta estimate is biased upward. Therefore, Stambaugh concluded that the predictive power of dividend yield is exaggerated.

While Stambaugh's bias adjustment is based on the well-known bias result of $\hat{\psi}$ being $-(1 + 3\psi)/T$, Lewellen [28] observes that such a bias typically occurs in the data that appear to mean-revert more strongly than they truly do. However, predictors such as the dividend yield are hardly mean-reverting. They contain roots very close to unity. Instead, Lewellen [28] proposed to use $\psi = 1$ as the true value in equation (27) in order to derive a conservative adjustment. For example, using NYSE returns and log dividend-price ratio over the ample period from 1946 to 2000 in equation (25), the least-squares estimate of β is 0.92, with a standard error of 0.48. When applying Stambaugh [36] bias correction, the estimate becomes 0.20 with a one-sided p -value of 0.308. In contrast, using Lewellen's conservative bias adjustment, the estimate becomes 0.66 with a t -ratio of 4.67.

Implied Constraint

To push the idea of incorporating prior knowledge as in Lewellen [28], Cochrane [12] argued that the coefficients from predictive regressions of returns, dividend growth, and dividend-price ratio using the lagged dividend-price ratio should be constrained. In particular, if we run the following regressions,

$$r_{t+1} = a_r + b_r(d_t - p_t) + \epsilon_{r,t+1} \quad (28)$$

$$\Delta d_{t+1} = a_d + b_d(d_t - p_t) + \epsilon_{d,t+1} \quad (29)$$

$$d_{t+1} - p_{t+1} = a_{dp} + \psi(d_t - p_t) + \epsilon_{dp,t+1} \quad (30)$$

the regression coefficients b_r , b_d , and ψ should be related. In fact, by substituting equations (28) through (30) into equation (19), we have the following results:

$$b_r = 1 - \rho\psi + b_d \quad (31)$$

$$\epsilon_{r,t+1} = \epsilon_{d,t+1} - \epsilon_{dp,t+1} \quad (32)$$

Since $\rho < 1$ and $\psi < 1$, equation (31) implies that one cannot test the joint hypothesis of $b_r = 0$ and $b_d = 0$ at the same time. In other words, if we fail to reject the hypothesis of $b_d = 0$, we cannot ignore the evidence that b_r is positive in the predictive

Table 8 The actual parameter estimates and the implied parameters; adopted from Cochrane [12]

	Estimates	$\sigma(\hat{b})$	Implied value	Correlation		
				ϵ_r	ϵ_d	ϵ_{dp}
$\hat{b}_r$	0.097	0.050	0.101	ϵ_r	19.6	66.0 -70.0
$\hat{b}_d$	0.008	0.044	0.004	ϵ_d	66.0	14.0 7.5
$\hat{\psi}$	0.941	0.047	0.945	ϵ_{dp}	-70.0	7.5 15.3

regression of equation (28). As shown in Table 8 (adapted from [12]), the $\hat{b}_d$ estimate is very close to zero with a large standard error. Therefore, b_r is probably close to 0.101 as implied by equation (31) using $\rho = 0.9638$, which is close to the actual estimate of 0.097.

At the same time, equation (32) suggests that shocks to returns and the dividend-price ratio are highly correlated with each other, which are indeed true from Table 8. For example, the negative correlation is as high as 70%. Table 8 also shows that the estimated coefficients b_r , b_d , and ψ and their corresponding implied values from equation (31) are amazingly close.

Error in Variables

If predictability is driven by time-varying expected returns, predictors should predict the expected return. In other words, the conventional predictive regression implicitly assumes that the expected return is a linear function of the predictors. However, from the small magnitude of predictability, it is clear that predictors can only be noisy estimates of the true expected returns. In other words, the predictive regressions are subject to error-in-variables problem, which will bias estimates. To overcome the problem, Pastor and Stambaugh [31] proposed to model the expected return as an unobservable component and to allow its innovation being correlated with innovations in predictors. Specifically, they propose the following model:

$$r_{t+1} = \mu_t + \epsilon_{t+1} \quad (33)$$

$$\mu_{t+1} = \mu_0 + \phi\mu_t + u_{t+1} \quad (34)$$

$$z_{t+1} = \psi z_t + v_{t+1} \quad (35)$$

where μ_t is the expected return and z_t is the predictor. In this system, predictors affect the expected return

through a correlation between u_{t+1} and v_{t+1} . In other words, information in the predictors helps to improve the “quality” of the expected return estimates in the spirit of the classical SUR (seemingly unrelated) regression. Since the system of equations (33)–(35) can be reduced to a predictive regression when $\mu_t = z_t$, it should perform at least as good as the predictive regression. Additional constraints can be imposed to improve the estimation efficiency. For example, when there is a positive shock to the expected return, future expected returns will be high due to the persistence, which will result in a low price, or equivalently a low return. Therefore, we can incorporate this prior constraint of the negative correlation between u_{t+1} and ϵ_{t+1} . Using quarterly data and imposed economic prior in a Bayesian framework, Pastor and Stambaugh [31] found that the dividend yield is a very useful predictor.

In Pastor and Stambaugh [31], a predictor affects the expected return through an indirect channel by improving the precision of the expected return estimate as in the same spirit of the SUR regression. To push the idea further, Baranchuk and Xu [3] studied both the direct and indirect effects of predictors on the expected return. In particular, equations (34) and (35) are replaced by

$$\mu_{t+1} = \mu_0 + \phi\mu_t + \delta m_t + u_{t+1} \quad (36)$$

$$z_{t+1} = m_t + \eta_{t+1} \quad (37)$$

$$m_{t+1} = \alpha m_t + v_{t+1} \quad (38)$$

where m_t is the expected predictor. In this framework, the expected predictor directly affects the level of expected return, whereas the unexpected predictor continues to influence the efficiency of the time-varying expected return through its correlation with the innovation to the expected return. Using both the dividend yield and repurchasing yield, Baranchuk and Xu [3] were able to demonstrate the very different role played by the two predictors. The repurchasing yield affects the expected return directly, whereas dividend yield works through the indirect channel affecting the precision in the estimate of the expected return. From a technical perspective, such an elaborated model structure also avoids the potential spurious regression and possesses the ability to incorporate economic prior.

Out-of-sample Predictive Power

If predictability is due to time-varying expected returns, a representative investor will not attempt to make abnormal returns since both his/her risk exposure and risk tolerance change over time. However, a nonrepresentative (isolated) investor might be able to take the advantage of return predictability in order to outperform the market. Goyal and Welch [37] have run a horse racing on the out-of-sample predictive power for a model based on unconditional forecast *versus* models with conditional forecast using different predictors including the following:

- the dividend–price ratio and the dividend yield [10];
- the earnings price ratio and dividend–earnings (payout) ratio [26];
- the short-term interest rate [7];
- the term spread and the default spread [7, 16, 23];
- the inflation rate [14, 18];
- the book-to-market ratio [25, 32];
- the consumption, wealth, and income ratio [27]; and
- the aggregate net issuing activity [2].

After comparing the (conditional) root-mean-squared errors (RMSEs) with respect to the predicted returns to the (unconditional) RMSEs using a simple sample mean, Goyal and Welch [37] concluded that in-sample predictability can be very different from out-of-sample performance. In most cases, the unconditional RMSEs are smaller than the conditional RMSEs. Therefore, they believe that most results from predictive regressions are just statistical illusions.

Similar to the idea of using prior information to improve the predictive power for future returns as in [12], prior economic constraints are valuable information and should be used simultaneously. Campbell and Yogo [11] recognized that if we are really predicting the expected returns in a predictive regression, we should throw out the negative predicted returns since the expected returns should always be positive. By constraining the predicted returns to be nonnegative, Campbell and Yogo [11] found that most predictors in the above list are indeed useful in predicting future returns even out of sample.

In a related study, Xu [38] recognized that, given the low R^2 s in predictive regressions, it is very difficult to provide accurate prediction about the

magnitude of future returns due to errors in the parameter estimates. One has to at least estimate two parameters in a predictive regression, while a sample mean corresponds to only one parameter estimate. The additional estimation error could easily overwhelm the benefit of using predictors. Therefore, in out-of-sample studies, a more useful question to ask is whether we are able to predict the *direction* of future market movement. On the basis of this idea, Xu [38] had studied the economic significance of the following trading strategy:

Trading strategy: Invest in a risky asset today only if the predicted future asset return is positive.

Under the t -distributed return assumption, there exists a moderate condition, under which the trading strategy will outperform a buy-and-hold strategy. Using inflation, relative interest rate, and dividend–price ratio as predictors, Xu [38] had shown that such a trading strategy could potentially double the performance of a buy-and-hold strategy over the sample period from 1952 to 1998.

Concluding Comments

Given the vast literature on predictability, in this article attention was focused on the questions of the existence of predictability and the interpretation of predictability. Return predictability has always been a challenge to the EMH. The traditional view on the evidence is either denying the evidence with the help of statistical methods or attributing the phenomenon to market frictions. For example, most predictors, except for the past returns, are persistent. Such a statistical property may result in a spurious regression. Predictors are also imperfect, which will bring in an error-in-variables problem in estimation. Many market microstructure effects, such as bid–ask bounce and nonsynchronous trading, may induce autocorrelation in a short-run and among small stocks. The modern view, however, takes a more positive approach by recognizing the time-varying risk premium due to changes in either investment opportunities or investors' risk tolerance. If this is indeed the case, many variables that predict business cycles should also help to predict returns, for example, the interest rate. Other variables that contain a price component can also predict returns because prices reflect expectations and should summarize all future changes in the expected return or CF distributions.

Many statistical issues can be dealt with a more elaborated model structure and the use of economic prior. For example, if predictability is due to changes in the risk premium, we can model the expected returns explicitly as following an $AR(1)$ process and impose the nonnegativity constraint. We can also alleviate the market microstructure effects by using low-frequency returns such as monthly or quarterly returns. From an empirical perspective, however, we should not expect to find huge return predictability. It seems to be odd that some economic agents do not try to explore economic profits, even though they are not subject to the kind of risks that a representative investor might expose to. Indeed, many studies tend to find economically weak but statistically significant evidence of predictability.

Overall, we believe that the evidence points to the direction of predictable returns even under careful statistical inference. If this is the case, will evidence on predictability have any implications on asset pricing? The answer is yes as evident from the literature on testing the conditional asset pricing models. Predictability also has implications on investors' asset allocation decisions [22]. If returns are positively correlated over time, investors might want to allocate less wealth to equity. This is because a risk averse investor understands that he/she will be subject to even larger downside risk if today's return is low. This is still an active area of research.

End Notes

^aSuppose returns follow the following $AR(1)$ process,

$$r_t - \mu = \theta(r_{t-1} - \mu) + \epsilon_t$$

the true Sharpe ratio defined as μ/σ_ϵ can be expressed as

$$\frac{\mu}{\sigma_\epsilon} = \frac{1}{\sqrt{1 - \theta^2}} \frac{\mu}{\sigma_y}$$

^bWhen the discount rate is not zero, we can define a discounted process such that it is a martingale.

^cThe positive autocorrelation in the equal-weighted index shown in Table 1 is also consistent with the nonsynchronous trading story.

^dSince holdings in the future contracts are much smaller than the market capitalization of the 500 largest companies, the evidence could be consistent with our argument that the representative investors will not try to trade on the predictability, while nonrepresentative investor in a segment of the market could.

References

- [1] Ang, A. & Bekaert, G. (2007). Stock return predictability: is it there? *Review of Financial Studies* **20**, 651–707.
- [2] Baker, M. & Wurgler, J. (2000). The equity share in new issues and aggregate stock returns, *Journal of Finance* **55**, 2219–2257.
- [3] Baranchuk, N. & Xu, Y. (2007). *What Predicts Stock Returns?—The Role of Expected versus Unexpected Predictors*. working paper, University of Texas at Dallas.
- [4] Boudoukh, J., Michaely, R., Richardson, M.P. & Roberts, M.R. (2007). On the importance of measuring payout yield: implications for empirical asset pricing, *Journal of Finance* **62**, 877–915.
- [5] Boudoukh, J., Richardson, M.P. & Whitelaw, R.F. (1994). Tale of three schools: insights on autocorrelations of short-horizon stock returns, *Review of Financial Studies* **7**, 539–573.
- [6] Boudoukh, J., Richardson, M. & Whitelaw, R.F. (2008). The myth of long-horizon predictability, *Review of Financial Studies* **21**, 1533–1575.
- [7] Campbell, J.Y. (1987). Stock returns and term structure, *Journal of Financial Economics* **18**, 373–399.
- [8] Campbell, J.Y. (1991). A variance decomposition for stock returns, *Economic Journal* **101**, 157–179.
- [9] Campbell, J.Y., Lo, A.W. & MacKinlay, C.A. (1997). *The Econometrics of Financial Markets*, Princeton University Press, Princeton.
- [10] Campbell, J.Y. & Shiller, R. (1988). Stock prices, earnings, and expected dividends, *Journal of Finance* **43**, 661–676.
- [11] Campbell, J.Y. & Yogo, M. (2006). Efficient tests of stock return predictability, *Journal of Financial Economics* **81**, 27–60.
- [12] Cochrane, J.H. (2008). The dog that did not bark: a defense of return predictability, *Review of Financial Studies* **21**, 1533–1575.
- [13] Conrad, J. & Kaul, G. (1988). Time-variation in expected returns, *The Journal of Business* **61**, 409–425.
- [14] Fama, E. (1981). Stock returns, real activity, inflation, and money, *American Economic Review* **71**, 545–565.
- [15] Fama, E. & French, K. (1988). Permanent and temporary components of stock prices, *Journal of Finance* **96**, 246–273.
- [16] Fama, E. & French, K. (1989). Business conditions and expected returns on stock and bonds, *Journal of Financial Economics* **25**, 23–49.
- [17] Fama, E. & French, K. (1996). Multifactor explanations of asset pricing anomalies, *Journal of Finance* **51**, 55–84.
- [18] Fama, E. & Schwert, G.W. (1977). Asset returns and inflation, *Journal of Financial Economics* **5**, 115–146.
- [19] Ferson, W.E. & Harvey, C.R. (1991). The variation of economic risk premiums, *Journal of Political Economy* **99**, 385–415.

-
- [20] Ferson, W.E., Sarkissian, S. & Simin, T.T. (2003). Spurious regressions in financial economics? *Journal of Finance* **58**, 1393–1414.
 - [21] Granger, C.W.J. & Newbold, P. (1974). Spurious regressions in economics, *Journal of Econometrics* **14**, 111–120.
 - [22] Kandle, S. & Stambaugh, R.F. (1996). On the predictability of stock returns: an asset allocation perspective, *The Journal of Finance* **51**, 385–424.
 - [23] Keim, D. & Stambaugh, R. (1986). Predicting returns in stock and bond markets, *Journal of Financial Economics* **17**, 357–390.
 - [24] Kirby, C. (1997). Measuring the predictability in stock and bond returns, *Review of Financial Studies* **10**, 579–630.
 - [25] Kothari, S. & Shanken, J. (1997). Book-to-market, dividend yield, and expected market returns: a time-series analysis, *Journal of Financial Economics* **44**, 169–203.
 - [26] Lamont, O. (1998). Earnings and expected returns, *Journal of Finance* **53**, 1563–1587.
 - [27] Lettau, M. & Ludvigson, S. (2001). Consumption, aggregate wealth, and expected stock returns, *Journal of Finance* **56**, 515–849.
 - [28] Lewellen, J. (2004). Predicting returns with financial ratios, *Journal of Financial Economics* **74**, 209–235.
 - [29] Lo, A. & MacKinlay, A.C. (1990). An econometric analysis of nonsynchronous-trading, *Journal of Econometrics* **45**, 181–212.
 - [30] Pastor, L. & Stambaugh, R.F. (2008). *Predictive Systems: Living with Imperfect Predictors*, Working Paper, 12814, NBER.
 - [31] Pastor, L. & Stambaugh, R.F. (2007). *Predictive Systems: Living with Imperfect Predictors*. NBER, Working Paper.
 - [32] Pontiff, J. & Schall, L.D. (1998). Book-to-market ratio as predictors of market returns, *Journal of Financial Economics* **49**, 141–160.
 - [33] Richardson, M. & Smith, T. (1991). Tests of financial models in the presence of overlapping observations, *Review of Financial Studies* **4**, 227–254.
 - [34] Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market, *Journal of Finance* **39**, 1127–1140.
 - [35] Schwert, G. (2003). Anomalies and market efficiency, in *Handbook of Economics and Finance*, G. Constantinides, M. Harris & R. Stulz, eds, North Holland, Amsterdam, Netherlands, Chapter 17.
 - [36] Stambaugh, R.R. (1999). Predictive regressions, *Journal of Financial Economics* **54**, 375–421.
 - [37] Welch, I. & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction, *Review of Financial Studies* **21**, 1533–1575.
 - [38] Xu, Y. (2004). Small levels of predictability and large economic gains, *Journal of Empirical Finance* **11**, 247–275.

Related Articles

Capital Asset Pricing Model; Efficient Market Hypothesis; Expectations Hypothesis; Risk Premia.

YEXIAO XU

Real Options

Real options theory is about decision making and value creation in an uncertain world. It owes its success to its ability to reconcile frequently observed investment behaviors that are seemingly inconsistent with rational choices at the firm level. For instance, Dixit [15] uses real options to explain why firms undertake investments only if they expect a yield in excess of a required hurdle rate, thus violating the Marshallian theory of long- and short-run equilibria.^{a,b} This is because, relative to a setting in which there is no uncertainty, unforeseeable future payouts discourage commitment to a project unless the expected profitability of the project is sufficiently high. The real options methodology allows to identify and value risky investments and, under certain conditions, to even take advantage of uncertainty. Indeed, as we shall see, this valuation approach insures investments against possible adverse outcomes while retaining upside potential.^c

Definition of a Real Option

A real option gives its holder the right, but not the obligation, to take an action (e.g., deferring, expanding, contracting, or abandoning) for a specified price, called the *exercise—or strike—price*, on or before some specified future date. We can identify at least six factors that affect the value of a real option: the value of the underlying risky asset (i.e., the project, investment, or acquisition); the exercise price; the volatility of the value of the underlying asset; the time to expiration of the option; the interest rate; and the dividend rate of the underlying asset (i.e., the cash outflows or inflows over the life of the option). If the value of the underlying project, its standard deviation, or the time to expiration increase, so too does the value of the option. The value of the (call) option also increases if the risk-free rate of interest goes up. Lost dividends decrease the value of the option.^d A higher exercise price reduces (augments) the value of a call (put) option.^e

The quantitative origins of real options derive from the seminal work of Black and Scholes [2] and Merton [32] on financial options pricing (*see Black–Scholes Formula*). These roots are evident in the assumptions that trading and decision making

take place in continuous time and that the underlying sources of uncertainty follow Brownian motions. Even though these assumptions may be unsuitable in some corporate contexts, they permit to derive precise theoretical solutions, thereby proving to be essential.^{f,g} The focus of this earlier literature has been on valuing individual real options: the option to expand a project, for instance, is an American call option (*see American Options*). So is a deferral option that gives a firm the right to delay the start of a project. The option to abandon a project, or to scale back by selling a fraction of it for a fixed price, is formally an American put (*see American Options*). Real-world projects, however, are often more complex in that they involve a collection of real options, whose values may interact. The recent development in financial options interdependencies has enabled a smoother transition from a theoretical stage to an application stage.^h Margrabe's [29] valuation of an option to exchange one risky asset for another (*see Margrabe Formula*) finds immediate application in the modeling of switching options, which allow a firm to switch between two modes of operation. Geske [19] values options on options—called *compound options*—which may be applied to growth opportunities that become available only if earlier investments are undertaken. Phased investments belong to this category. Thus, almost paradoxically, in this relatively new field of research, the mathematically most complex models, which apply sophisticated contingent claims analysis techniques, entail a great wealth of factual applications.ⁱ Moreover, numerous studies show that real options represent a sizable fraction of a firm's value; both Kester [25] and Pindyck [35], for instance, estimate that the value of a firm's growth options is more than half its market value of equity if demand volatility exceeds 20%. For this reason, the theory of real options has gained significant importance among management practitioners whose choices determine the success or failure of their enterprises. Amram and Kulatilaka [1] collect several case studies to show practitioner audiences how real options can improve capital investment planning and results. In particular, they list three real options characteristics that are of great use to managers: (i) options payoffs are contingent on the manager's decisions; (ii) options valuations are aligned with financial market valuations; and (iii) options thinking can be used to design and manage strategic investments proactively. The real options paradigm,

however, is only the last stage in the evolution of valuation models. The traditional approach to valuing investment projects, which owes its origins to John Hicks and Irving Fisher, is based on net present value. This technique involves discounting expected net cash flows from a project at a discount rate that reflects the risk of those cash flows, called the *risk-adjusted* discount rate. Brennan and Trigeorgis [8] characterize this first-phase models as *static*, or *mechanistic*. The second-phase models are *control-able* cash-flow models, in which projects can be managed actively in response to the resolution of exogenous uncertainties. Since they ignore strategic investment, both first- and second-phase models often lead to suboptimal decisions. *Dynamic, game-theoretic* options models assume that projects can be managed actively, instead.^j These models take into account not only the resolution of exogenous uncertainties but also the actions of outside parties. For this reason, an area of immense importance within game-theoretic options models concerns market competition and strategy.

Strategic firm interactions are isomorphic to a portfolio of real options.^k Furthermore, the payouts of a project (as well as its value) can be seen as the outcome of a game among the inside agent, outside agents, and nature. Dixit [14] and Williams [40] were the first to consider real options within an equilibrium context. Smit and Ankum [37], among others, study competitive reactions within a game-theoretic framework under different market structures. In the same line of research is Grenadier's [21] analysis of a perfectly competitive real-estate market with stochastic demand and time to build.^l

Solution of the Basic Model

Besides particular cases, all investment expenditures have two important characteristics. First, they are at least partly irreversible, and second, they can be delayed so that the firm has the opportunity to wait for new information to arrive before committing any resources.

The most basic continuous-time model of irreversible investment was originally developed by McDonald and Siegel [31]. In their problem, a firm must decide when to invest in a single risky project, denoted by V , with a fixed known cost I . The project is assumed to follow a geometric Brownian motion

with expected return and volatility indicated by μ and σ , respectively. The project's payout rate equals δ . Formally, the process can be written as

$$\frac{dV}{V} = (\mu - \delta) dt + \sigma dz \quad (1)$$

where dz is the increment of a Wiener process and $(dz)^2 = dt$.^{m,n} In addition, denote the value of the firm's investment opportunity (its option to invest) by $F(V)$. It can be shown that the optimal rule is to invest at the date τ^* when the project's value first exceeds a certain optimal threshold V^* . This rule maximizes

$$F(V) = \max_{\tau} E[(V_{\tau} - I)e^{-\mu\tau}], \quad V_0 = V \quad (2)$$

over all possible stopping times τ , where E is the expectation operator. Prior to undertaking the project the only return to holding the investment option is its capital appreciation, so that

$$\mu F(V) dt = E[dF(V)] \quad (3)$$

Expanding $dF(V)$ using Itô's lemma yields

$$dF(V) = F'(V) dV + \frac{1}{2} F''(V) (dV)^2 \quad (4)$$

where primes indicate derivatives. Lastly, substituting equation (1) in (4) and taking expectations on both sides gives

$$\frac{1}{2} \sigma^2 V^2 F''(V) + (\mu - \delta) V F'(V) - \mu F(V) = 0 \quad (5)$$

Equation (5) must be solved simultaneously for the project value $F(V)$ and the optimal investment threshold V^* , subject to three boundary conditions:

$$F(0) = 0 \quad (6)$$

$$F(V^*) = V^* - I \quad (7)$$

$$F'(V^*) = 1 \quad (8)$$

Equation (6) is equivalent to stating that the investment option is worthless when the project's outcome is null. Equations (7) and (8) indicate the payoff and marginal value associated with the optimum. To derive V^* , we must guess a functional form

that satisfies equation (5) and verify if it works. In particular, if we take $F(V) = AV^\beta$, then

$$V^* = \frac{\beta I}{\beta - 1} \quad (9)$$

and

$$\beta = \frac{1}{2} - \frac{\mu - \delta}{\sigma^2} + \sqrt{\left(\frac{\mu - \delta}{\sigma^2} - \frac{1}{2}\right)^2 + \frac{2\mu}{\sigma^2}} \quad (10)$$

The optimal rule is to invest when the value of the project exceeds the cost by a factor $\frac{\beta}{\beta - 1} > 1$. This result is in contrast with net present value, which prescribes to invest as long as the value of the project exceeds the cost ($V^* = I$). However, since the latter rule does not account for uncertainty and irreversibility, it is incorrect and it leads to suboptimal decisions.

Furthermore, as it is apparent from the solution, the higher the risk of the project, measured by σ , the larger are the value of the option and the opportunity cost of investing. Increasing values of the growth rate, μ , also cause $F(V)$ and V^* to be higher. On the other hand, larger expected payout rates, δ , lower both $F(V)$ and V^* as holding the option becomes more expensive.

Dixit and Pindyck [16] show how the optimal investment rule can be found by using both dynamic programming (as it is done above) and contingent claims analysis.^o

Contingent claims methods require one important assumption: stochastic changes in the value of the project must be *spanned* by existing assets in the economy (see **Complete Markets**). Specifically, capital markets must be sufficiently *complete* so that one could find an asset, or construct a dynamic portfolio of assets, the price of which is perfectly correlated with the value of the project (see **Risk-neutral Pricing**).^{p,q} This assumption allows properly taking into account all the flexibility (options) that the project might have and using all the information contained in market prices (e.g., futures prices) when such prices exist.^r If the sources of uncertainty in a project are not traded assets (examples of which are product demand uncertainty, geological uncertainty, technological uncertainty, cost uncertainty, etc.), an equilibrium model of asset prices can be used to value the contingent claim.^s

Numerical Methods in Real Options

In practice, most real option problems must be solved using numerical methods. Until recently, these methods were so complex that only few companies found it practical to use them when formulating operating strategies. However, advances in both computational power and understanding of the techniques over the last 20 years have made it feasible to apply real options thinking to strategic decision making. Numerical solutions give not only the value of the project but also the optimal strategy for exercising the options.^t The simplest real option problems involving one or two state variables can be more conveniently solved using binomial or trinomial trees in one or two dimensions (see **Finite Element Methods**).^u When a problem involves more state variables, perhaps path dependent, the more practical solution is to use Monte Carlo simulation methods (see **Monte Carlo Simulation**).^{v,w} In order to do so, we use the assumption that properly anticipated prices (or cash flows) fluctuate randomly. Regardless of the pattern of cash flows that a project is expected to have, the changes in its present value will follow a random walk. This theorem, attributable to Paul Samuelson, allows us to combine any number of uncertainties by using Monte Carlo techniques, and to produce an estimate of the present value of a project conditional on the set of random variables drawn from their underlying distributions. More generally, there are two types of numerical techniques for option valuation: (i) those that approximate the underlying stochastic processes directly and (ii) those approximating the resulting partial differential equation. The first category includes lattice approaches and Monte Carlo simulations. Examples of the second category include numerical integration (see **Quadrature Methods**); and the implicit/explicit finite difference schemes (see **Finite Difference Methods for Barrier Options**; **Finite Difference Methods for Early Exercise Options**) used by Brennan [6], Brennan and Schwartz [7], and Majd and Pindyck [28], among others.

Conclusions

The application of option concepts to value real assets has been an important growth area in the theory and practice of finance. The insights and techniques

derived from option pricing have proven capable of quantifying the managerial operating flexibility and strategic interactions thus far ignored by conventional net present value and other quantitative approaches. This flexibility represents a substantial part of the value of many projects and neglecting it can undervalue investments and induce a misallocation of resources. By explicitly incorporating management flexibility into the analysis, real options have provided the tools for properly valuing corporate resources and capital budgeting.

End Notes

^aMarshall's [30] analysis states that if price exceeds long-run average cost, then existing firms expand and new ones enter a business.

^bSymmetrically, firms often do not exit a business for lengthy periods, even after the price falls substantially below long-run average cost. This phenomenon is dubbed *hysteresis*.

^cAmram and Kulatilaka [1], Brennan and Trigeorgis [8], Copeland and Antikarov [10], Dixit and Pindyck [16], Grenadier [21], Schwartz and Trigeorgis [36], and Smit and Trigeorgis [38] represent core reference volumes on real investment decisions under uncertainty. The survey article by Boyer *et al.* [4] is a noteworthy collection of all most notable contributions to the literature on strategic investment games, from the pioneering works of Gilbert and Harris [20] and Fudenberg and Tirole [18] to more recent contributions.

^dFor a thorough examination of the variables driving real options' analysis, the reader is referred to [10], Chapter 1.

^eAn interesting example on the effect of an option's exercise price on its value is presented by Moel and Tufano [33]. They study the bidding for rights to explore and develop a copper mine in Peru. A peculiar aspect of the transaction is the nature of the bidding rules that bidders were required to follow by the Peruvian government. Each bid was required to specify the minimum amount that the bidder would spend on developing the property if they decided to go ahead after exploration. This is equivalent to allowing the bidders to specify the exercise price of their development option. This structure gave rise to incentives that affected the amount that firms would offer, thus inducing successful bidders to make uneconomic investments.

^fBoyarchenko and Levendorskii [3] relax these assumptions and show how to analyze firm decisions in discrete time.

^gCox, Ross, and Rubinstein's [12] binomial approach enables a more simplified valuation of options in discrete time.

^hDetemple [13] provides a complete treatment of American-style derivatives pricing. He analyzes in detail

both plain and exotic contingent claims and presents recent results on the numerical computation of optimal exercise boundaries, hedging prices, and hedging portfolios.

ⁱFlexible manufacturing, natural resource investments, land development, leasing, large-scale energy projects, research and development, and foreign investment are all examples of real options cases.

^jTrigeorgis and Mason [39] remark that option valuation can be seen as a special version of decision tree analysis. Decision scientists propose the use of decision tree analysis [34] to capture the value of operating flexibility associated with many projects.

^kLuerhman [27] explains how a business strategy compares to a series of options more than to a single option. *De facto*, executing a strategy almost always involves making a sequence of decisions: some actions are taken immediately, while others are deliberately deferred.

^lThe time-to-build and continuous-time features of Grenadier's [21] model translate into an infinite state space. Despite this, he is able to determine the optimal construction rules by engineering an artificial economy with a finite state space in which the equilibrium strategy is identical to that of the true economy.

^mAccording to equation (1), the current project value is known but its future values are uncertain.

ⁿChapters 3 and 4 in [16] provide a thorough overview of the mathematical tools necessary to study investment decision using a continuous-time approach.

^oAlthough equivalent, the two methodologies are conceptually rather different: while the former lies on the option's value satisfying the Bellman equation, the latter is founded on the construction of a risk-free portfolio formed by a long position in the firm's option and a short position in units of the firm's project. Chapter 5 in [16] presents a detailed explanation, along with a guided derivation, of the optimal rule obtained on adopting each technique.

^pDuffie [17] gives great emphasis to the implications of complete markets for asset pricing under uncertainty.

^qHarrison and Kreps [22], Harrison and Pliska [23], and others have shown that, in complete markets, the absence of arbitrage implies the existence of a probability distribution such that securities are priced on the basis of their discounted (at the risk-free rate) expected cash flows, where expectation is determined under the *risk-neutral probability measure*. If all risks can be hedged, this probability is unique. The critical advantage of working in the risk-neutral environment is that it is a convenient environment for option pricing.

^rThe reader is referred to [36] for a more rigorous discussion on the application of contingent claims analysis to determine a project's optimal operating policy.

^sSee [11] for the derivation of a fundamental partial differential equation that must be satisfied by the value of all contingent claims on the value of state variables that are not traded assets.

[†]Broadie and Detemple [9] conduct a careful evaluation of the many methods for computing American option prices.

^uBoyle [5] shows how lattice frameworks can be extended to handle two state variables.

^vIn the last few years, methods have been developed, which allow using simulations for solving American-style options. For example, Longstaff and Schwartz [26] developed a least-squares Monte Carlo approach to compare the value of immediate exercise with the conditional expected value from continuation.

^wHull and White [24] suggest a control variate technique to improve computational efficiency when a similar derivative asset with an analytic solution is available.

References

- [1] Amram, M. & Kulatilaka, N. (1999). *Real Options: Managing Strategic Investment in an Uncertain World*, Harvard Business School Press, Boston, MA.
- [2] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *The Journal of Political Economy* **18**(3), 637–654.
- [3] Boyarchenko, S. & Levendorskii, S. (2000). Entry and exit strategies under Non-Gaussian distributions, in *Project Flexibility, Agency, and Competition*, M. Brennan & L. Trigeorgis, eds, Oxford University Press, Inc., New York, NY, pp. 71–84.
- [4] Boyer, R., Gravelle, E. & Lasserre, P. (2004). *Real Options and Strategic Competition: A Survey*. Working Paper.
- [5] Boyle, P. (1988). A lattice framework for option pricing with two state variables, *The Journal of Financial and Quantitative Analysis* **23**(1), 1–12.
- [6] Brennan, M. (1979). The pricing of contingent claims in discrete time models, *The Journal of Finance* **34**(1), 53–68.
- [7] Brennan, M. & Schwartz, E. (2001). Finite differences methods and jump processes arising in the pricing of contingent claims: a synthesis, in *Real Options and Investment Under Uncertainty: Classical Readings and Recent Contributions*, E. Schwartz & L. Trigeorgis, eds, The MIT Press, Cambridge, MA, pp. 559–570.
- [8] Brennan, M. & Trigeorgis, L. (2000). *Project Flexibility, Agency, and Competition*, Oxford University Press, Inc., New York, NY.
- [9] Broadie, M. & Detemple, J. (1996). American option valuation: new bounds, approximations, and a comparison of existing methods, *The Review of Financial Studies* **9**(4), 1211–1250.
- [10] Copeland, T. & Antikarov, V. (2001). *Real Options: A Practitioner's Guide*, W.W. Norton & Company, New York.
- [11] Cox, J., Ingersoll, J. & Ross, S. (1985). An intertemporal general equilibrium model of asset prices, *Econometrica* **53**(2), 363–384.
- [12] Cox, J., Ross, S. & Rubinstein, M. (1979). Option pricing: a simplified approach, *The Journal of Financial Economics* **7**(3), 229–263.
- [13] Detemple, J. (2005). *American-Style Derivatives: Valuation and Computation*, Chapman & Hall/CRC.
- [14] Dixit, A. (1989). Entry and exit decisions under uncertainty, *The Journal of Political Economy* **97**(3), 620–638.
- [15] Dixit, A. (1992). Investment and hysteresis, *The Journal of Economic Perspectives* **6**(1), 107–132.
- [16] Dixit, A. & Pindyck, R. (1994). *Investment Under Uncertainty*, Princeton University Press, Princeton, NJ.
- [17] Duffie, D. (1996). *Dynamic Asset Pricing Theory*, Princeton University Press, Princeton, NJ.
- [18] Fudenberg, D. & Tirole, J. (1985). Preemption and rent equalization in the adoption of new technology, *The Review of Economic Studies* **52**(3), 383–401.
- [19] Geske, R. (1979). A note on analytical valuation formula for unprotected American call options on stocks with known dividends, *The Journal of Financial Economics* **7**, 375–380.
- [20] Gilbert, R. & Harris, R. (1984). Competition with lumpy investment, *RAND Journal of Economics* **15**(2), 197–212.
- [21] Grenadier, S. (2000). Strategic options and product market competition, in *Project Flexibility, Agency, and Competition*, M. Brennan, & L. Trigeorgis, eds, Oxford University Press, Inc., New York, NY, pp. 275–296.
- [22] Harrison, M. & Kreps, D. (1979). Martingales and arbitrage in multiperiod securities markets, *The Journal of Economic Theory* **20**(3), 381–408.
- [23] Harrison, J. & Pliska, S. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and Their Applications* **11**, 215–260.
- [24] Hull, J. & White, A. (1988). The use of control variate technique in option pricing, *The Journal of Financial and Quantitative Analysis* **23**(3), 237–251.
- [25] Kester, W. (2001). Today's options for tomorrow's growth, in *Real Options and Investment Under Uncertainty: Classical Readings and Recent Contributions*, E. Schwartz & L. Trigeorgis, eds, The MIT Press, Cambridge, MA, pp. 33–46.
- [26] Longstaff, F. & Schwartz, E. (2001). Valuing American options by simulations: a simple least-squares approach, *The Review of Financial Studies* **14**(1), 113–147.
- [27] Luehrman, T. (2001). Strategy as a portfolio of real options, in *Real Options and Investment Under Uncertainty: Classical Readings and Recent Contributions*, E. Schwartz & L. Trigeorgis, eds, The MIT Press, Cambridge, MA, pp. 385–404.
- [28] Majd, S. & Pindyck, R. (1987). Time to build, option value, and investment decisions, *The Journal of Financial Economics* **18**(1), 7–27.
- [29] Margrabe, W. (1978). The value of an option to exchange one asset for another, *The Journal of Finance* **33**(1), 177–186.

- [30] Marshall, A. (1890). *Principles of Economics*, Macmillan and Co, London.
- [31] McDonald, R. & Siegel, D. (1986). The value of waiting to invest, *The Quarterly Journal of Economics* **101**(4), 707–728.
- [32] Merton, R. (1973). Theory of rational option pricing, *Bell Journal of Economics* **4**(1), 141–183.
- [33] Moel, A., Tufano, P., Brennan, M. & Trigeorgis, L. (2000). Bidding for the antamina mine: valuation and incentives in a real options context, in *Project Flexibility, Agency, and Competition*, Oxford University Press, London, pp. 128–150.
- [34] Myers, S. (2001). Finance theory and financial strategy, in *Real Options and Investment Under Uncertainty: Classical Readings and Recent Contributions*, E. Schwartz & L. Trigeorgis, eds, The MIT Press, Cambridge, MA, pp. 19–32.
- [35] Pyndick, R., Schwartz, E. & Trigeorgis, L. (eds) (2001). Irreversible investment, capacity choice, and the value of the firm, in *Real Options and Investment Under Uncertainty: Classical Readings and Recent Contributions*, The MIT Press, Cambridge, MA, pp. 313–334.
- [36] Schwartz, E. & Trigeorgis, L. (2001). *Real Options and Investment Under Uncertainty: Classical Readings and Recent Contributions*, The MIT Press, Cambridge, MA.
- [37] Smit, H. & Ankum, L. (1993). A real options and game-theoretic approach to corporate investment strategy under competition, *Financial Management* **22**(3), 241–250.
- [38] Smit, H. & Trigeorgis, L. (2004). *Strategic Investment: Real Options and Games*, Princeton University Press, Princeton, NJ.
- [39] Trigeorgis, L. & Mason, S. (2001). Valuing managerial flexibility, in *Real Options and Investment Under Uncertainty: Classical Readings and Recent Contributions*, E. Schwartz & L. Trigeorgis, eds, The MIT Press, Cambridge, MA, pp. 47–60.
- [40] Williams, J. (1993). Equilibrium and options on real assets, *The Review of Financial Studies* **6**(4), 825–850.

Further Reading

Grenadier, S. (2000). *Game Choices: The Intersection of Real Options and Game Theory*, Risk Books, London.

Related Articles

Black–Scholes Formula; Option Pricing: General Principles; Options: Basic Definitions; Swing Options.

DORIANA RUFFINO

Employee Stock Options

Employee stock options (ESOs) are call options issued by a company and given to its employees as part of their remuneration. The rationale is that granting the employee options will align his or her interests with those of the firm's shareholders. This is particularly relevant for managers and Chief Executive Officers (CEOs) whose behavior has more impact on firm value than that of lower ranked employees.

ESOs are prevalent in both the United States and Europe. In the fiscal year 1999, 94% of the S&P 500 companies granted options to their top executives, and the value at the grant date represented 47% of total pay for the CEOs [14]. The 2005 Mercer Compensation Survey [34] reports that over 75% of CEOs receive option grants and options account for 32% of CEO pay. The Hay Group's 2006 European Executive Pay Survey [15] found that 55% of the companies in the study used stock options.

ESOs are American call options on the company stock granted to the employee. They typically have a number of characteristics that distinguish them from financial options; see [38] and [35] for overviews. There is usually an initial *vesting* period during which the options cannot be exercised. Cliff vesting is a structure where all options granted on a given date become exercisable after an initial period, usually 2–4 years. Stepped vesting refers to a structure where a proportion of an option grant becomes exercisable each year, for example, 10% after one year, then 20%, 30%, and 40% each subsequent year. The most common structure is straight vesting where the proportions are equal, say one-third of the grant is exercisable after each of the first three years (see [2, 30], and [25]). During this period, typically, the employee must forfeit the remaining unvested options if he or she resigned or is fired. Clearly, if there is no vesting period, the options are American style, whereas, in the limit, as the vesting period approaches maturity, the options become European (see **American Options; Call Options** for descriptions of European and American options).

After the vesting period, the options may be exercised at any time up to and including the maturity date. These options are typically long dated with a 10-year maturity being most common. The employee

is not able to sell or transfer the options at any time. This is in keeping with the alignment or incentive effect of options. The option terms are modified if the employee exits the firm either because he or she is fired, leaves, retires, or dies. These “sunset rules” vary widely across firms (see [8] for details), but typically the employee is given a period of time in which to exercise the options or forfeit them. The length of time is generally longest if the employee retires and shortest if the employee leaves or is fired. In addition to being unable to unwind an option position by selling it, employees are typically also restricted from short selling the stock of their company and thus are very restricted in terms of hedging their option exposure [5].

There have been a number of empirical studies of ESO exercise patterns. Huddart and Lang [23] study exercise behavior in a sample of eight firms that volunteered internal records on option grants and exercises from 1982 to 1994. They find a pervasive pattern of option exercises well before expiration—the mean fraction of option life elapsed at the time of exercise varied from 0.26 to 0.79 over companies. Bettis *et al.* [2] analyze a unique database of more than 140 000 option exercises by corporate executives at almost 4000 firms during the period 1996 through 2002. They find 10-year options were exercised a median of 4.25 years before expiry. A further feature documented in the data is that of block exercise. Huddart and Lang [23] find that the mean fraction of options from a single grant exercised by an employee at one time varied from 0.18 to 0.72 over employees at a number of companies. Similarly, Aboody [1] reports yearly mean percentages of options exercised over the life of 5 and 10 year options, showing exercises are spread over the life of the options. Some of these block exercises are due to the nature of the vesting structure—for instance, Huddart and Lang [23] find spikes on vest dates corresponding to large exercises on those dates—but there are also other block exercises on dates that cannot be explained by vesting.

There are many questions of interest—including “What is the employee's optimal exercise policy?”, “What are the options worth to him or her?”, “What is the corresponding cost to the company of granting the options?” The employee's exercise policy and option value should incorporate the features described above—his or her inability to hedge being key. The cost to the company should reflect the

value of the option liability to the issuing corporation. This usually entails the assumption that shareholders are well diversified, so the cost should be the risk-neutral option value conditional on the optimal exercise behavior of the employee. This distinction between the option value to the employee (often called *subjective value*) and the cost to the company is important and arises because the employee cannot perfectly hedge the risk of the option exposure, while shareholders are typically assumed to be well diversified.

The need to quantify the company cost is particularly relevant in light of changes in accounting rules, which require companies to expense options at the grant date. In 1995, the Financial Accounting Standards Board (FASB) set a standard to require firms to expense stock options using “fair value”. However, this included the possibility to calculate the option cost to the firm as the option’s intrinsic value at the grant date. Perhaps motivated by this, companies mainly granted options that were at-the-money thus calculating a zero value for the expense. The huge growth of employee options and a series of corporate scandals led to pressure for changes to these rules, and new regulations (FASB 123R in the United States, International Financial Reporting Standards (IFRS) 2 in Europe) were introduced in 2004. From 2005 onward, these regulations required companies to use a “fair value method” of accounting for the expense of employee options, and although recommendations are made concerning appropriate methods, there is still much scope for interpretation by companies. For instance, use of the (European) Black–Scholes price with an estimated “expected term” is an acceptable and popular approach. Despite these changes, the granting of options that are at-the-money is still typical.

To take into account the nonhedgability aspect of employee options, we need to move outside of the complete market or risk-neutral pricing framework to an incomplete setting (see **Complete Markets**). There have been many papers in the literature in this direction, beginning with [22, 31, 32], and [14], amongst others. These papers typically develop binomial models that take trading restrictions and employee risk aversion into account and compute a certainty equivalent or subjective value for the employee options. These models make the simplistic assumption that any nonoption wealth is invested in a riskless bank account, and most treat the options as

European. Also in a binomial model, Cai and Vijn [3] and Carpenter [5] assume nonoption or outside wealth is invested in a Merton-style portfolio, but only allow for a one-off choice of this portfolio.

Many of the papers mentioned above observe that the utility-based or subjective valuation to the employee is much lower than the equivalent Black–Scholes value (the value obtained in an equivalent complete market setting); however, this is not universally true in models where nonoption wealth is invested in a riskless bond [14]. Generally, however, the (subjective) value of the options to the employee is less than the cost of the options to the company because of the employee’s hedging restrictions.

These models have been extended to incorporate the impact of optimal investment of outside wealth in a market or risky asset, rather than just a bank account. This was tackled in the natural setting of utility-indifference pricing (see [19] for a survey containing many references) for European options by Henderson [17]. This allows the employee to reduce risk by partial hedging in the market asset, which would seem to reflect what can be done in practice. The basic setup for continuous-time models with hedging in the market asset is as follows. The market M follows a geometric Brownian motion

$$dM/M = \mu dt + \sigma dB \quad (1)$$

where μ, σ are constants, and B is standard Brownian motion. Let W be a standard Brownian motion and assume $dBdW = \rho dt$. We can write $dW = \rho dB + \sqrt{1 - \rho^2}dZ$ for Z a Brownian motion independent of B . The company stock S also follows a geometric Brownian motion:

$$\begin{aligned} dS/S &= \nu dt + \eta dW \\ &= \nu dt + \eta(\rho dB + \sqrt{1 - \rho^2}dZ) \end{aligned} \quad (2)$$

The term $\rho^2\eta^2$ represents the hedgable or market component of the total risk of the stock and $(1 - \rho^2)\eta^2$ is the unhedgable or idiosyncratic/firm-specific risk of the stock. When $\rho^2 = 1$ all the risk can be hedged and an employee with an option on the stock S is able to perfectly hedge the risk he or she faces. (To avoid arbitrage, we should have $\nu - r = (\mu - r)\eta/\sigma$. More generally, CAPM imposes the relation

$v - r = (\mu - r)\eta\rho/\sigma$; see **Capital Asset Pricing Model**).

The employee can invest in a riskless asset with interest rate r and hold a cash amount θ_t in the market at time t . The dynamics of the wealth account X are then

$$dX = \theta dM/M + r(X - \theta)dt \quad (3)$$

If the employee is granted λ European call options with strike K then he or she solves

$$V(t, X_t, S_t, \lambda) = \sup_{\theta_t: \theta_t \geq t} E_t[U(X_T + \lambda(S_T - K)^+)] \quad (4)$$

Under the assumption of exponential utility, closed-form solutions are obtained for the value function. The utility-based or utility-indifference value p of the λ options solves $V(t, x + p, S_t, 0) = V(t, x, S_t, \lambda)$. In such models, it is straightforward to show that, in the limit, as the (absolute value of) correlation between the company stock and market approaches one, the Black–Scholes or complete market value is recovered. This value is then an upper bound on the utility-based valuation. In a European option setting, the Black–Scholes value represents the cost to the company, so we see the value to the employee is lower than the cost to the company. The other comparison of interest is to consider what difference the ability to undertake partial hedging in the market makes. The ability to partially hedge is valuable to the employee and his or her utility-based or subjective option value is higher than without the hedging/investment opportunity. In other words, the subjective value increases in (absolute value of) correlation. Similar to the models without the market asset, the higher the employee's risk aversion, the lower the utility-based option value.

Of course, as we described earlier, employee stock options are American options, and allow for early exercise once the options have vested. Some of the aforementioned papers also treat American style options and the general intuition is that hedging restrictions of the employee result in an earlier exercise and a lower subjective value than the equivalent Black–Scholes (complete market) American option. In the continuous-time model with investment in the market asset, closed-form results are found under the assumptions of exponential utility and perpetual options in [18] and numerical solutions for finite maturity in [33]. Kadam *et al.* [29]

considered the case of the perpetual option but without the partial hedging in the market. The exercise threshold and option values both decrease with risk aversion and increase with (absolute value of) correlation. Just as in the European case, the ability to partially hedge risk is valuable to the employee. He or she places a higher value on the option and waits longer to exercise it. It is also possible that stock volatility reduces the option value in some scenarios because of the interaction of the convex payoff with the concave utility function; see [17, 18, 33], and also [37]. Since the cost to the company is just the risk-neutral option value conditional on optimal exercise by the employee, it is also decreasing with risk aversion and increasing with (absolute value of) correlation [13]. Detemple and Sundaresan [9] and Ingersoll [25] also allow for optimal investment in a market portfolio and consider numerical approaches to the marginal pricing of small quantities of options.

As mentioned earlier, the data indicates that employees exercise options in a number of tranches on different occasions. Consideration of models that only allow for one option or one exercise time is not consistent with this observation. Vesting is one feature that clearly encourages such block exercise behavior, and indeed, Huddart and Lang [23] observe that many exercises take place immediately when the options vest. However, vesting does not appear to explain all of the intertemporal exercises, since not all exercises occur immediately upon vesting. Another reason for intertemporal exercise is risk aversion and the inability to hedge risk due to restrictions. Jain and Subrahmanian [26] consider a binomial model for a risk-averse employee who is granted a number of options. Grasselli [12] extends the binomial framework to include optimal investment in a correlated market asset. These papers find numerically that optimal behavior is to exercise options when the stock price reaches a boundary and the discrete nature of the model results in exercise occurring at a discrete set of dates or stock price levels. Rogers and Scheinkman [36] make similar observations numerically in a discrete approximation to a continuous-time model without investment opportunities in a market asset.

Grasselli and Henderson [13] show that under the assumption of exponential utility and perpetual options, closed-form solutions can be derived for the multiple-option problem with investment opportunities in a market asset. In fact, they show that

given N options, there are N unique stock price thresholds at which the employee should exercise an option. These thresholds are obtained using a recursive relation. The price thresholds are increasing as the quantity of options falls. In other words, when the employee has fewer options remaining, he or she is exposed to less risk, and thus is willing to wait for a higher price threshold before exercising further options. Similar comparative statics apply as in the single American option case—thresholds, option values, and company cost are decreasing in risk aversion and increasing in (absolute value of) correlation. In addition, they show that the cost to the company is underestimated if a single optimal exercise threshold is used. Since, in reality, options are not exercised one at a time, the paper also introduces a transaction cost on exercise, which restores block exercise as the optimal solution, again found in closed form. Leung and Sircar [33] consider the finite-maturity version of the problem, which leads to numerical solution of the free-boundary problem. They also include features such as vesting and job termination risk.

As described earlier, option terms change upon departure of an employee from the company and this should be incorporated into pricing models. Employee departure is typically modeled by an exogenous exponentially distributed time with constant intensity, independent of the stock price, similar to a reduced-form approach in credit modeling. (*see Structural Default Risk Models*). The papers [5, 6, 27, 39], and [33], among others, incorporate departure into a variety of setups in this manner.

Although we do not discuss estimation in any detail here, it is clear that estimation of such models is difficult. The models require estimates of risk aversion, outside wealth, and employee departure rate, which are not easily obtained. Bettis *et al.* [2] and Carpenter [5] have attempted calibration exercises on utility style models to exercise data; however, many simplifying assumptions have to be made due to data limitations. For example, they assume an option grant is exercised on one date only rather than on multiple occasions. Perhaps surprising is the finding of Carpenter [5] that after a calibration to data, a reduced-form model of employee departure is as capable as a utility-maximizing model in explaining option exercises. This finding motivates another strand of the literature, which models option exercise exogenously by postulating an exercise boundary in terms of the moneyiness of the

option; see [24] and [7]. This style of model has the attraction of simplicity and is much easier for calibration since the employee's risk aversion is no longer used. For this reason, it may well be a fruitful approach for calculating an approximation to the cost of the options to the company for accounting purposes.

We now turn to briefly discuss a number of other features relevant in employee compensation. Typically employees receive new grants of options periodically; however, companies also engage in resetting (where the option strike of existing options is adjusted downward when the options are out-of-the-money) and reloading (where additional options are granted automatically when existing options are exercised [10]).

Besides the traditional employee options described in this article, companies have increasingly granted performance-based options, which link option vesting or exercise to the achievement of market or accounting-based performance targets. These options are very popular in Europe, but have, until recently, been less common in the United States; see [11] and references therein. Compensation linked to accounting data is potentially open to manipulation and managers with such options may be motivated to inflate earnings. There is a large literature on the connection between compensation involving accounting-based targets and earnings management, either of a direct nature [4]) or accrual-based management or manipulation Healy [16].

Performance-based options can also have exercise prices contingent on performance relative to a comparison group—these are known as indexed options; see Johnson and Tian [28] who value such options in a risk-neutral framework using techniques from exchange or Margrabe options (*see Margrabe Formula*). Managers are then rewarded as a function of relative performance relative to a peer group rather than on absolute performance [20].

Other important issues that have not been discussed here include the impact of dilution—when options are exercised, the company typically issues new shares. Another important issue is the influence the CEO has on the stock price *via* his or her effort or choice of projects/risk. The problem of how best to compensate managers, given the benefits of improved incentives and the costs of inefficient risk-sharing, is the subject of a large literature on the principal agent problem; see the classic reference [21].

References

- [1] Aboody, D. (1996). Market valuation of employee stock options, *Journal of Accounting and Economics* **22**, 357–391.
- [2] Bettis, J.C., Bizjak, J.M. & Lemmon, M.L. (2005). Exercise behavior, valuation and the incentive effects of employee stock options, *Journal of Financial Economics* **76**, 445–470.
- [3] Cai, J. & Vijh, A. (2005). Executive stock and option valuation in a two state-variable framework, *Journal of Derivatives* **12**, 19–27.
- [4] Camara, A. & Henderson, V. (2007). *Performance Based Compensation and Direct Earnings Management*, Working paper.
- [5] Carpenter, J.N. (1998). The exercise and valuation of executive stock options, *Journal of Financial Economics* **48**, 127–158.
- [6] Carr, P. & Linetsky, V. (2000). The valuation of executive stock options in an intensity-based framework, *European Finance Review* **4**, 211–230.
- [7] Cvitanic, J., Wiener, Z. & Zapatero, F. (2008). Analytic pricing of employee stock options, *Review of Financial Studies* **21**, 683–724.
- [8] Dahiya, S. & Yermack, D. (2008). You can't take it with you: Sunset provisions for equity compensation when managers retire, resign or die, *Journal of Corporate Finance* **14**, 499–511.
- [9] Detemple, J. & Sundaresan, S. (1999). Nontraded asset valuation with portfolio constraints: a binomial approach, *Review of Financial Studies* **12**, 835–872.
- [10] Dybvig, P. & Lowenstein, M. (2003). Employee reload options: pricing, hedging and optimal exercise, *Review of Financial Studies* **12**, 145–171.
- [11] Gerakos, J.J., Goodman, T.H., Ittner, C.D. & Larcker, D.F. (2005). *The Adoption and Characteristics of Performance Stock Options*, Working paper.
- [12] Grasselli, M. (2005). *Nonlinearity, Correlation and the Valuation of Employee Options*, Working paper.
- [13] Grasselli, M. & Henderson, V. (2009). Risk aversion, effort and block exercise of executive stock options, *Journal of Economic Dynamics and Control* **33**, 109–127.
- [14] Hall, B.J. & Murphy, K.J. (2002). Stock options for undiversified executives, *Journal of Accounting and Economics* **33**, 3–42.
- [15] Hay Group (2007). 2006 European executive pay survey, *The Executive Edition* (1), 11–12.
- [16] Healy, P.M. (1985). The effect of bonus schemes on accounting decisions, *Journal of Accounting and Economics* **7**, 85–107.
- [17] Henderson, V. (2005). The impact of the market portfolio on the valuation, incentive and optimality of executive stock options, *Quantitative Finance* **5**(1), 35–47.
- [18] Henderson, V. (2007). Valuing the option to invest in an incomplete market, *Mathematics and Financial Economics* **1**, 103–128.
- [19] Henderson, V. & Hobson, D. (2009). Utility indifference pricing—an overview, in *Indifference Pricing*, R. Carmona, ed, Princeton University Press, Chapter 2.
- [20] Holmstrom, B. (1982). Moral hazard in teams, *Bell Journal of Economics* **13**, 1324–1340.
- [21] Holmstrom, B. & Milgrom, P. (1987). Aggregation and linearity in the provision of intertemporal incentives, *Econometrica* **55**, 303–328.
- [22] Huddart, S. (1994). Employee stock options, *Journal of Accounting and Economics* **18**, 207–231.
- [23] Huddart, S. & Lang, M. (1996). Employee stock option exercises: an empirical analysis, *Journal of Accounting and Economics* **21**, 5–43.
- [24] Hull, J. & White, A. (2002). How to value employee stock options, *Financial Analysts Journal* **60**(1), 114–119.
- [25] Ingersoll, J.E. (2006). The subjective and objective evaluation of compensation stock options, *Journal of Business* **79**, 453–487.
- [26] Jain, A. & Subramanian, A. (2004). The intertemporal exercise and valuation of employee stock options, *The Accounting Review* **79**(3), 705–743.
- [27] Jennergren, L. & Naslund, B. (1993). A comment on “Valuation of executive stock options and the FASB proposal”, *The Accounting Review* **68**, 179–183.
- [28] Johnson, S. & Tian, Y. (2000). Indexed executive stock options, *Journal of Financial Economics* **57**, 35–64.
- [29] Kadam, A., Lakner, P. & Srinivasan, A. (2005). *Executive Stock Options: Value to the Executive and Cost to the Firm*, Working paper, City University.
- [30] Kole, S. (1997). The complexity of compensation contracts, *Journal of Financial Economics* **43**, 79–104.
- [31] Kulatilaka, N. & Marcus, A.J. (1994). Valuing employee stock options, *Financial Analysts Journal* November–December, 46–56.
- [32] Lambert, R.A., Larcker, D.F. & Verrecchia, R.E. (1991). Portfolio considerations in valuing executive compensation, *Journal of Accounting Research* **29**(1), 129–149.
- [33] Leung, T. & Sircar, R. (2009). Accounting for risk aversion, vesting, job termination risk and multiple exercises in valuation of employee stock options, *Mathematical Finance* **19**(1), 99–128.
- [34] Mercer Human Resource Consulting. (2006). *2005 CEO Compensation Survey and Trends*.
- [35] Murphy, K.J. (1999). Executive compensation, in *Handbook of Labor Economics*, O. Ashenfelter & D. Card, eds, North Holland, Vol. 3.
- [36] Rogers, L.C.G. & Scheinkman, J. (2007). Optimal exercise of executives stock options, *Finance and Stochastics* **11**, 357–372.
- [37] Ross, S.A. (2004). Compensation, incentives and the duality of risk aversion and riskiness, *Journal of Finance* **59**(1), 207–225.
- [38] Rubinstein, M. (1995). On the accounting valuation of employee stock options, *Journal of Derivatives* **3**, 8–24.

- [39] Sircar, R. & Xiong, W. (2007). A general framework for evaluating executive stock options, *Journal of Economic Dynamics and Control* **31**(7), 2317–2349.

Further Reading

Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**(3), 637–654.

Related Articles

American Options; Black–Scholes Formula; Call Options; Capital Asset Pricing Model; Complete Markets; Structural Default Risk Models.

VICKY HENDERSON & JIA SUN

Arbitrage Strategy

It is difficult to imagine a normative condition that is more widely accepted and unquestionable in the minds of anyone involved in the field of quantitative finance other than the *absence of arbitrage opportunities* in a financial market. Put plainly, an *arbitrage strategy* allows a financial agent to make certain profit out of nothing, that is, out of zero initial investment. This has to be disallowed on economic basis if the market is in equilibrium state, as opportunities for riskless profit would result in an instantaneous movement of prices of certain financial instruments.

Let us give an illustrative example of an arbitrage strategy in the foreign exchange market, commonly called the *triangular arbitrage*. Suppose that Mary, in Paris, is buying^a the US dollar for €0.685. Tom, in San Francisco, is buying Japanese yen for \$0.009419. Finally, Toru, in Tokyo, is buying one euro for ¥155.02. All these transactions are supposed to be able to occur at the *same time*. There is something worth noting in the situation just described—something that could allow you to make riskless profit. Let us see how. You borrow \$10 000 from your rich aunt Clara and tell her you will return the money in a matter of minutes. First, you approach Mary and change all your dollars to euros. This means that you will get €6850. With the euros in hand, you contact Toru and change them into yen—you will get $¥(6850 \times 155.02) = ¥1\,061\,887$. Finally, you call Tom, wire him all your yen and change them back to dollars, which gets you $\$(1\,061\,887 \times 0.009419) \equiv \$10\,001.91$. You give the \$10 000 back to your aunt Clara as promised, and you have managed to create \$1.91 out of thin air.

Although the above-mentioned example is oversimplistic, it gives a clear idea of what arbitrage is: a position on a combination of assets that requires zero initial capital and results in a profit with *no* risk involved. Let us now take a step further and see what will happen under the situation of the preceding example. As more and more investors become aware of the discrepancy between prices, they will all try to use the same smart strategy that you used for their benefit. Everyone will be trying to exchange US dollars for euros in the first step of the arbitrage, which will drive Mary to start buying the US dollar for less than €0.685 because of the high demand for the euros she is selling. Similarly, Tom will start buying

Japanese yen for less than \$0.009419 and Toru will be buying euro for less than ¥155.02. Very soon, the situation will be such that nobody is able to make a riskless profit anymore.

The economic rationale behind asking for nonexistence of arbitrage opportunities is based exactly on the discussion in the previous paragraph. If arbitrage opportunities were present in the market, a multitude of investors would try to take advantage of them simultaneously. Therefore, there would be an almost instantaneous move of the prices of certain financial instruments as a response to a supply–demand imbalance. This price movement will continue until any opportunity for riskless profit is no longer available.

It is important to note that the preceding, somewhat theoretical, discussion does *not* imply that arbitrage opportunities *never* exist in practice. On the contrary, it has been observed that opportunities for some, albeit usually minuscule, riskless profit appear frequently as a consequence of the huge amount of distant geographic trading locations, as well as a result of the numerous financial products that have sprung up and are sometimes interrelated in complicated ways. Realizing that such opportunities exist is a matter of rapid access to information that a certain group of investors, so-called *arbitrageurs*, has. It is rather the *existence* of arbitrageurs acting in financial markets that ensures that when arbitrage opportunities exist, they will be fleeting.

The principle of not allowing for arbitrage opportunities in financial markets has far-reaching consequences and has immensely boosted research in quantitative finance. The ground-breaking papers of Black (*see* **Black, Fischer**) and Scholes [1] and Merton (*see* **Merton, Robert C.**) [3], published in 1973, were the first instances explaining how absence of arbitrage opportunities leads to rational pricing and hedging formulas for European-style options in a geometric Brownian motion financial model.^b This idea was consequently taken up and generalized by many authors and has led to a profound understanding of the interplay between the economics of financial markets and the mathematics of stochastic processes, with deep-reaching results—*see* **Fundamental Theorem of Asset Pricing; Risk-neutral Pricing; Equivalent Martingale Measures; and Free Lunch** for some amazing developments on this path.

We close the discussion of arbitrages on an amusing note. Such is the firm belief on the principle of not allowing for arbitrage opportunities in financial modeling that even jokes have been created in order to substantiate it further. We quote directly from Chapter 1 of [2], which can be used as an excellent introduction to arbitrage theory: *A professor working in Mathematical Finance and a normal^c person go on a walk and the normal person sees a €100 bill lying on the street. When the normal person wants to pick it up, the professor says: “Don’t try to do that. It is absolutely impossible that there is a €100 bill lying on the street. Indeed, if it were lying on the street, somebody else would have picked it up before you.”*

End Notes

^a. All the prices referred to in this example are *bid* prices of the currencies involved.

^b. For historical perspectives regarding option pricing and hedging, see **Black, Fischer; Merton, Robert C.; Arbitrage: Historical Perspectives; and Option Pricing Theory: Historical Perspectives**. For a more thorough quantitative treatment, see **Risk-neutral Pricing**.

^c. Is this bold distancing from normality of mathematical finance professors, clearly implied from the authors of [2], a decisive step toward illuminating the perception they have of their own personalities? Or is it just a gimmick used to add another humorous ingredient to the joke? The answer is left for the reader to determine.

References

- [1] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *The Journal of Political Economy* **81**, 637–654.
- [2] Delbaen, F. & Schachermayer, W. (2006). *The Mathematics of Arbitrage*, Springer Finance, Springer-Verlag, Berlin.
- [3] Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.

Further Reading

- Dalang, R.C., Morton, A. & Willinger, W. (1990). Equivalent martingale measures and no-arbitrage in stochastic securities market models, *Stochastics and Stochastics Reports* **29**, 185–201.
- Delbaen, F. (1992). Representing martingale measures when asset prices are continuous and bounded, *Mathematical Finance* **2**, 107–130.
- Delbaen, F. & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing, *Mathematische Annalen* **300**, 463–520.
- Delbaen, F. & Schachermayer, W. (1998). The fundamental theorem of asset pricing for unbounded stochastic processes, *Mathematische Annalen* **312**, 215–250.
- Elworthy, K.D., Li, X.-M. & Yor, M. (1999). The importance of strictly local martingales; applications to radial Ornstein-Uhlenbeck processes, *Probability Theory and Related Fields* **115**, 325–355.
- Föllmer, H. & Schied, A. (2004). *Stochastic Finance*, de Gruyter Studies in Mathematics, extended Edition, Walter de Gruyter & Co., Berlin, Vol. 27.
- Hull, J.C. (2008). *Options, Futures, and Other Derivatives*, 7th Edition, Prentice Hall.
- Michael Harrison, J. & Kreps, D.M. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- Michael Harrison, J. & Pliska, S.R. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and Their Applications* **11**, 215–260.
- Shreve, S.E. (2004). *Stochastic Calculus for Finance. I: The Binomial Asset Pricing Model*, Springer Finance, Springer-Verlag, New York.

Related Articles

Black, Fischer; Equivalent Martingale Measures; Fundamental Theorem of Asset Pricing; Free Lunch; Good-deal Bounds; Merton, Robert C.; Ross, Stephen; Risk-neutral Pricing.

CONSTANTINOS KARDARAS

Fundamental Theorem of Asset Pricing

Consider a financial market modeled by a price process S on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The *fundamental theorem of asset pricing*, which is one of the pillars supporting the modern theory of Mathematical Finance, states that the following two statements are *essentially* equivalent:

1. S does not allow for arbitrage (NA).
2. There exists a probability measure Q on the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which is equivalent to $\mathbb{P}$ and under which the process is a martingale.

We have formulated this theorem in vague terms, which will be made precise in the sequel: we formulate versions of this theorem that use precise definitions and avoid the use of the word *essentially*.

The story of this theorem started—like most of modern Mathematical Finance—with the work of Black (see **Black, Fischer**), Scholes [3], and Merton (see **Merton, Robert C.**) [25]. These authors consider a model $S = (S_t)_{0 \leq t \leq T}$ of geometric Brownian motion proposed by Samuelson (see **Samuelson, Paul A.**) [30], which is widely known today as the Black–Scholes model. Presumably every reader of this article is familiar with the well-known technique to price options in this framework (see **Risk-neutral Pricing**): one changes the underlying measure $\mathbb{P}$ to an equivalent measure Q under which the discounted stock price process is a martingale. Subsequently, one prices options (and other derivatives) by simply taking expectations with respect to this “risk neutral” or “martingale” measure Q .

In fact, this technique was *not* the novel feature of [3, 25]. It was used by actuaries for some centuries and it was also used by Bachelier [2] in 1900 who considered Brownian motion (which, of course, is a martingale) as a model $S = (S_t)_{0 \leq t \leq T}$ of a stock price process. In fact, the prices obtained by Bachelier (see **Bachelier, Louis (1870–1946)**) by this method were—at least for the empirical data considered by Bachelier himself—very close to those derived from the celebrated Black–Merton–Scholes formula ([34]).

The decisive *novel feature* of the Black–Merton–Scholes approach was the argument that links this pricing technique with the notion of arbitrage: the payoff function of an option can be precisely *replicated* by trading dynamically in the underlying stock. This idea, which is credited in footnote 3 of [3] to Merton, opened a completely new perspective on how to deal with options, as it linked the pricing issue with the idea of hedging, that is, dynamically trading in the underlying asset.

The technique of replicating an option is completely absent in Bachelier’s early work; apparently, the idea of “spanning” a market by forming linear combinations of primitive assets first appears in the Economics literature in the classic paper by Arrow (see **Arrow, Kenneth**) [1]. The mathematically delightful situation, that the market is complete in the sense that all derivatives can be replicated, occurs in the Black–Scholes model as well as in Bachelier’s original model of Brownian motion (see **Second Fundamental Theorem of Asset Pricing**). Another example of a model in continuous time sharing this property is the compensated Poisson process, as observed by Cox and Ross (see **Ross, Stephen**) [4]. Roughly speaking, these are the only models in continuous time sharing this seductively beautiful “martingale representation property” (see [16, 39] for a precise statement on the uniqueness of these families of models).

Appealing as it might be, the consideration of “complete markets” as above is somewhat dangerous from an economic point of view: the precise replicability of options, which is a sound mathematical theorem in the framework of the above models, may lead to the illusion that this is also true in economic reality. However, these models are far from matching reality in a one-to-one manner. Rather they only highlight important aspects of reality and therefore should not be considered as ubiquitously appropriate.

For many purposes, it is of crucial importance to put oneself into a more general modeling framework.

When the merits as well as the limitations of the Black–Merton–Scholes approach unfolded in the late 1970s, the investigations on the fundamental theorem of asset pricing started. As Harrison and Pliska formulate it in their classic paper [15]: “it was a desire to better understand their formula which originally motivated our study, ...”.

The challenge was to obtain a deeper insight into the relation of the following two aspects: on

one hand, the methodology of pricing by taking expectations with respect to a properly chosen “risk neutral” or “martingale” measure Q ; on the other hand, the methodology of pricing by “no arbitrage” considerations. Why, after all, do these two seemingly unrelated approaches yield identical results in the Black–Merton–Scholes approach? Maybe even more importantly: how far can this phenomenon be extended to more involved models?

To the best of the author’s knowledge, the first person to take up these questions in a systematic way was Ross (*see* **Ross, Stephen**) [29]; see also [4, 27, 28]. He chose the following setting to formalize the situation: fix a topological, ordered vector space (X, τ) , modeling the possible cash flows (e.g., the payoff function of an option) at a fixed time horizon T . A good choice is, for example, $X = L^p(\Omega, \mathcal{F}, \mathbb{P})$, where $1 \leq p \leq \infty$ and $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$ is the underlying filtered probability space. The set of marketed assets M is a subspace of X .

In the context of a stock price process $S = (S_t)_{0 \leq t \leq T}$ as above, one might think of M as all the outcomes of an initial investment $x \in \mathbb{R}$ plus the result of subsequent trading according to a predictable trading strategy $H = (H_t)_{0 \leq t \leq T}$. This yields (in discounted terms) an element

$$m = x + \int_0^T H_t dS_t \quad (1)$$

in the set M of marketed claims. It is natural to price the above claim m by setting $\pi(m) = x$, as this is the net investment necessary to finance the above claim m .

For notational convenience, we shall assume in the sequel that S is a one-dimensional process. It is straightforward to generalize to the case of d risky assets by assuming that S is $\mathbb{R}^d$ -valued and replacing the above integral by

$$m = x + \int_0^T \sum_{i=1}^d H_t^i dS_t^i \quad (2)$$

Some words of warning about the stochastic integral (1) seem necessary. The precise admissibility conditions, which should be imposed on the stochastic integral (1), in order to make sense both mathematically as well as economically, are a subtle issue. Much of the early literature on the fundamental theorem of asset pricing struggled exactly with this

question. An excellent reference is [14]. Ross [29] circumvented this problem by deliberately leaving this issue aside and simply starting with the modeling assumption that the subset $M \subseteq X$ as well as a pricing operator $\pi : M \rightarrow \mathbb{R}$ are *given*.

Let us now formalize the notion of arbitrage. In the above setting, we say that the **no arbitrage** assumption is satisfied if, for $m \in M$, satisfying $m \geq 0$, $\mathbb{P}$ -a.s. and $\mathbb{P}[m > 0] > 0$, we have $\pi(m) > 0$. In prose, this means that it is not possible to find a claim $m \in M$, which bears no risk (as $m \geq 0$, $\mathbb{P}$ -a.s.), yields some gain with strictly positive probability (as $\mathbb{P}[m > 0] > 0$), and such that its price $\pi(m)$ is less than or equal to zero.

The question that now arises is whether it is possible to extend $\pi : M \rightarrow \mathbb{R}$ to a nonnegative, continuous linear functional $\pi^* : X \rightarrow \mathbb{R}$.

What does this have to do with the issue of martingale measures? This theme was developed in detail by Harrison and Kreps [14]. Suppose that $X = L^p(\Omega, \mathcal{F}, \mathbb{P})$ for some $1 \leq p < \infty$, that the price process $S = (S_t)_{0 \leq t \leq T}$ satisfies $S_t \in X$, for each $0 \leq t \leq T$, and that M contains (at least) the “simple integrals” on the process $S = (S_t)_{0 \leq t \leq T}$ of the form

$$m = x + \sum_{i=1}^n H_i (S_{t_i} - S_{t_{i-1}}) \quad (3)$$

Here $x \in \mathbb{R}$, $0 = t_0 < t_1 < \dots < t_n = T$ and $(H_i)_{i=1}^n$ is a (say) bounded process which is *predictable*, that is, H_i is $\mathcal{F}_{t_{i-1}}$ -measurable. The sums in equation (3) are the Riemann sums corresponding to the stochastic integrals (1). The Riemann sums (3) have a clear-cut economic interpretation [14]. In equation (3) we do not have to bother about subtle convergence issues as only finite sums are involved in the definition. It is therefore a traditional (minimal) requirement that the Riemann sums of the form (3) are in the space M of marketed claims; naturally, the price of a claim m of the form (3) should be defined as $\pi(m) = x$.

Now suppose that the functional π , which is defined for the claims of the form (3) can be extended to a continuous, nonnegative functional π^* defined on $X = L^p(\Omega, \mathcal{F}, \mathbb{P})$. If such an extension π^* exists, it is induced by some function $g \in L^q(\Omega, \mathcal{F}, \mathbb{P})$, where $\frac{1}{p} + \frac{1}{q} = 1$. The nonnegativity of π^* is tantamount to $g \geq 0$, $\mathbb{P}$ -a.s., and the fact that $\pi^*(1) = 1$ shows that g is the density of a probability measure Q with Radon–Nikodym derivative $\frac{dQ}{d\mathbb{P}} = g$.

If we can find such an extension π^* of π , we thus find a probability measure Q on $(\Omega, \mathcal{F}, \mathbb{P})$ for which

$$\pi^* \left(\sum_{i=1}^n H_i (S_{t_i} - S_{t_{i-1}}) \right) = \mathbb{E}_Q \left[\sum_{i=1}^n H_i (S_{t_i} - S_{t_{i-1}}) \right] \quad (4)$$

for every bounded predictable process $H = (H_i)_{i=1}^n$ as above, which is tantamount to $(S_t)_{0 \leq t \leq T}$ being a martingale (see [Th. 2] [14], or [Lemma 2.2.6] [11]).

To sum up, in the case $1 \leq p < \infty$, finding a continuous, nonnegative extension $\pi^* : L^p(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{R}$ of π amounts to finding a $\mathbb{P}$ -absolutely continuous measure Q with $\frac{dQ}{d\mathbb{P}} \in L^q$ and such that $(S_t)_{0 \leq t \leq T}$ is a martingale under Q .

At this stage, it becomes clear that in order to find such an extension π^* of π , the Hahn–Banach theorem should come into play in some form, for example, in one of the versions of the separating hyperplane theorem.

In order to be able to do so, Ross assumes ([p. 472] [29]) that “...we will endow X with a strong enough topology to insure that the positive orthant $\{x \in X | x > 0\}$ is an open set, ...”. In practice, the only infinite-dimensional ordered topological vector space X , such that the positive orthant has nonempty interior, is $X = L^\infty(\Omega, \mathcal{F}, \mathbb{P})$, endowed with the topology induced by $\|\cdot\|_\infty$.

Hence the two important cases, applying to Ross’ hypothesis, are when either the probability space Ω is finite, so that $X = L^p(\Omega, \mathcal{F}, \mathbb{P})$ simply is finite dimensional and its topology does not depend on $1 \leq p \leq \infty$, or if $(\Omega, \mathcal{F}, \mathbb{P})$ is infinite and $X = L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ equipped with the norm $\|\cdot\|_\infty$.

After these preparations we can identify the two convex sets to be separated: let $A = \{m \in M : \pi(m) \leq 0\}$ and B be the interior of the positive cone of X . Now make the easy, but crucial, observation: these sets are disjoint if and only if the no-arbitrage condition is satisfied. As one always can separate an open convex set from a disjoint convex set, we find a functional $\tilde{\pi}$, which is strictly positive on B , while $\tilde{\pi}$ takes nonpositive values on A . By normalizing $\tilde{\pi}$, that is, letting $\pi^* = \tilde{\pi}(1)^{-1} \tilde{\pi}$ we have thus found the desired extension.

In summary, the first precise version of the fundamental theorem of asset pricing is established in [29], the proof relying on the Hahn–Banach theorem. There are, however, serious limitations: in the case of

infinite $(\Omega, \mathcal{F}, \mathbb{P})$, the present result only applies to $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ endowed with the norm topology. In this case, the continuous linear functional π^* only is in $L^\infty(\Omega, \mathcal{F}, \mathbb{P})^*$ and not necessarily in $L^1(\Omega, \mathcal{F}, \mathbb{P})$; in other words, we cannot be sure that π^* is induced by a probability measure Q , as it may happen that $\pi^* \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})^*$ also has a singular part.

Another drawback, which already appears in the case of finite-dimensional Ω (in which case π^* certainly is induced by some Q with $\frac{dQ}{d\mathbb{P}} = g \in L^1(\Omega, \mathcal{F}, \mathbb{P})$) is the following: we cannot be sure that the function g is strictly positive $\mathbb{P}$ -a.s. or, in other words, that Q is equivalent to $\mathbb{P}$.

After this early work by Ross, a major advance in the theory was achieved between 1979 and 1981 by three seminal papers [14, 15, 24] by Harrison, Kreps, and Pliska. In particular, [14] is a landmark in the field. It uses a similar setting as [29], namely, an ordered topological vector space (X, τ) and a linear functional $\pi : M \rightarrow \mathbb{R}$, where M is a linear subspace of X . Again the question is whether there exists an extension of π to a linear, continuous, strictly positive $\pi^* : X \rightarrow \mathbb{R}$. This question is related in [14] to the issue of whether (M, π) is viable as a model of economic equilibrium. Under proper assumptions on the convexity and continuity of the preferences of agents, this is shown to be equivalent to the extension discussed above.

The paper [14] also analyzes the case when Ω is finite. Of course, only processes $S = (S_t)_{t=0}^T$ indexed by finite, discrete time $\{0, 1, \dots, T\}$ make sense in this case. For this easier setting, the following precise theorem was stated and proved in the subsequent paper [15] by Harrison and Pliska:

Theorem 1 ([Th. 2. 7.] [15]): *suppose the stochastic process $S = (S_t)_{t=0}^T$ is based on a finite, filtered, probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t=0}^T, \mathbb{P})$. The market model contains no-arbitrage possibilities if and only if there is an equivalent martingale measure for S .*

The proof again relies on a (finite-dimensional version) of the Hahn–Banach theorem plus an extra argument making sure to find a measure Q , which is equivalent to $\mathbb{P}$. Harrison and Pliska thus have achieved a precise version of the above meta-theorem in terms of equivalent martingale measures, which does not use the word “essentially”. Actually, the theme of the Harrison–Pliska theorem goes back

much further, to the work of Shimony [35] and Kemeny [22] on symbolic logic in the tradition of Carnap, de Finetti, and Ramsey. These authors showed that, in a setting with only finitely many states of the world, a family of possible bets does not allow (by taking linear combinations) for making a riskless profit (i.e., one certainly does not lose but wins with strictly positive probability), if and only if there is a probability measure Q on these finitely many states, which prices the possible bets by taking conditional Q -expectations.

The restriction to finite Ω is very severe in applications: the flavor of the theory, building on Black–Scholes–Merton, is precisely the concept of *continuous time*. Of course, this involves infinite probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$.

Many interesting questions were formulated in the papers [14, 15] hinting on the difficulties to prove a version of the fundamental theorem of asset pricing beyond the setting of finite probability spaces.

A major breakthrough in this direction was achieved by Kreps [24]: as above, let $M \subseteq X$ and a linear functional $\pi : M \rightarrow \mathbb{R}$ be given. The typical choice for X will now be $X = L^p(\Omega, \mathcal{F}, \mathbb{P})$, for $1 \leq p \leq \infty$, equipped with the topology τ of convergence in norm, or, if $X = L^\infty(\Omega, \mathcal{F}, \mathbb{P})$, equipped with the Mackey topology τ induced by $L^1(\Omega, \mathcal{F}, \mathbb{P})$. This setting will make sure that a continuous linear functional on (X, τ) will be induced by a measure Q , which is absolutely continuous with respect to $\mathbb{P}$.

The no-arbitrage assumption means that $M_0 := \{m \in M : \pi(m) = 0\}$ intersects the positive orthant X_+ of X only in $\{0\}$. In order to obtain an extension of π to a continuous, linear functional $\pi^* : X \rightarrow \mathbb{R}$ we have to find an element in $(X, \tau)^*$, which separates the convex set M_0 from the disjoint convex set $X_+ \setminus \{0\}$, that is, the positive orthant of X with 0 deleted.

Easy examples show that, in general, this is not possible. In fact, this is not much of a surprise (if X is infinite-dimensional) as we know that some topological condition is needed for the Hahn–Banach theorem to work.

It is always possible to separate a *closed* convex set from a disjoint *compact* convex set by a continuous linear functional. In fact, one may even get strict separation in this case. It is this version of the Hahn–Banach theorem that Kreps eventually applies.

But how? After all, neither M_0 nor $X_+ \setminus \{0\}$ are closed in (X, τ) , let alone compact.

Here is the ingenious construction of Kreps: define

$$A = \overline{M_0 - X_+} \quad (5)$$

where the bar denotes the closure with respect to the topology τ . We shall require that A still satisfies

$$A \cap X_+ = \{0\} \quad (6)$$

This property is baptized as “*no free lunch*” by Kreps:

Definition 1 [24]: *The financial market defined by (X, τ) , M , and π admits a free lunch if there are nets $(m_\alpha)_{\alpha \in I} \in M_0$ and $(h_\alpha)_{\alpha \in I} \in X_+$ such that*

$$\lim_{\alpha \in I} (m_\alpha - h_\alpha) = x \quad (7)$$

for some $x \in X_+ \setminus \{0\}$.

It is easy to verify that the negation of the above definition is tantamount to the validity of equation (6).

The economic interpretation of the “no free lunch” condition is a sharpening of the “no-arbitrage condition”. If the latter is violated, we can simply find an element $x \in X_+ \setminus \{0\}$, which also lies in M_0 . If the former fails, we cannot quite guarantee this, but we can find $x \in X_+ \setminus \{0\}$, which *can be approximated* in the τ -topology by elements of the form $m_\alpha - h_\alpha$. The passage from m_α to $m_\alpha - h_\alpha$ means that agents are allowed to “throw away money”, that is, to abandon a positive element $h_\alpha \in X_+$. This combination of the “free disposal” assumption with the possibility of passing to limits is crucial in Kreps’ approach (5) as well as in most of the subsequent literature. It was shown in [Ex. 3.3] [32]; ([33]) that the (seemingly ridiculous) “free disposal” assumption cannot be dropped.

Definition (5) is tailor-made for the application of Hahn–Banach. If the no free lunch condition (6) is satisfied, we may, for any $h \in X_+$, separate the τ -closed, convex set A from the one-point set $\{h\}$ by an element $\pi_h \in (X, \tau)^*$. As $0 \in A$, we may assume that $\pi_h|_A \leq 0$ while $\pi_h(h) > 0$. We thus have obtained a nonnegative (as $-X_+ \subseteq A$), continuous linear functional π_h , which is strictly positive on a given $h \in X_+$. Supposing that X_+ is τ -separable (which is the case in the above setting of L^p -spaces if $(\Omega, \mathcal{F}, \mathbb{P})$ is countably generated), fix a dense sequence $(h_n)_{n=1}^\infty$ and find strictly positive scalars $\mu_n > 0$ such that $\pi^* = \sum_{n=1}^\infty \mu_n \pi_{h_n}$ converges to a

probability measure in $(X, \tau)^* = L^q(\Omega, \mathcal{F}, \mathbb{P})$, where $\frac{1}{p} + \frac{1}{q} = 1$. This yields the desired extension π^* of π which is *strictly positive* on $X_+ \setminus \{0\}$.

We still have to specify the choice of (M_0, π) . The most basic choice is to take for given $S = (S_t)_{0 \leq t \leq T}$ the space generated by the “simple integrands” (3) as proposed in [14]. We thus may deduce from Kreps’ arguments in [24] the following version of the fundamental theorem of asset pricing.

Theorem 2 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be countably generated and $X = L^p(\Omega, \mathcal{F}, \mathbb{P})$ endowed with the norm topology τ , if $1 \leq p < \infty$, or the Mackey topology induced by $L^1(\Omega, \mathcal{F}, \mathbb{P})$, if $p = \infty$.*

Let $S = (S_t)_{0 \leq t \leq T}$ be a stochastic process taking values in X . Define $M_0 \subseteq X$ to consist of the simple stochastic integrals $\sum_{i=1}^n H_i(S_{t_i} - S_{t_{i-1}})$ as in equation (3).

Then the “no free lunch” condition (5) is satisfied if and only if there is a probability measure Q with $\frac{dQ}{d\mathbb{P}} \in L^q(\Omega, \mathcal{F}, \mathbb{P})$, where $\frac{1}{p} + \frac{1}{q} = 1$, such that $(S_t)_{0 \leq t \leq T}$ is a Q -martingale.

This remarkable theorem of Kreps sets new standards. For the first time, we have a mathematically precise statement of our meta-theorem applying to a general class of models *in continuous time*. There are still some limitations, however.

When applying the theorem to the case $1 \leq p < \infty$, we find the requirement $\frac{dQ}{d\mathbb{P}} \in L^q(\Omega, \mathcal{F}, \mathbb{P})$ for some $q > 1$, which is not very pleasant. After all, we want to know what exactly corresponds (in terms of some no-arbitrage condition) to the existence of an equivalent martingale measure Q . The q -moment condition is unnatural in most applications. In particular, it is not invariant under equivalent changes of measures as is done often in the applications.

The most interesting case of the above theorem is $p = \infty$. However, in this case, the requirement $S_t \in X = L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ is unduly strong for most applications. In addition, for $p = \infty$, we run into the subtleties of the Mackey topology τ (or the weak-star topology, which does not make much of a difference) on $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$. We shall discuss this issue below.

The “heroic period” of the development of the fundamental theorem of asset pricing marked by Ross [29], Harrison–Kreps [14], Harrison–Pliska [15], and Kreps [24], put the issue on safe mathematical grounds and brought some spectacular results. However, it still left many questions open; quite a number

of them were explicitly stated as open problems in these papers.

Subsequently a rather extensive literature developed, answering these problems and opening new perspectives. We cannot give a full account on all of this literature and refer, for example, to the monograph [11] for more extensive information. We can give an outline.

As regards the situation for $1 \leq p \leq \infty$ in Kreps’ theorem, this issue was further developed by Duffie and Huang [12] and, in particular, by Stricker [36]. This author related the no free lunch condition of Kreps to a theorem by Yan [37] obtained in the context of the Bichteler–Dellacherie theorem on the characterization of semimartingales. Using Yan’s theorem, Stricker gave a different proof of Kreps’ theorem, which does not need the assumption that $(\Omega, \mathcal{F}, \mathbb{P})$ is countably generated.

A beautiful extension of the Harrison–Pliska theorem was obtained in 1990 by Dalang, Morton, and Willinger [5]. They showed that, for an $\mathbb{R}^d$ -valued process $(S_t)_{t=0}^T$ in finite discrete time, the no-arbitrage condition is indeed equivalent to the existence of an equivalent martingale measure. The proof is surprisingly tricky, at least for the case $d \geq 2$. It is based on the measurable selection theorem (the suggestion to use this theorem is acknowledged to Delbaen). Different proofs of the Dalang–Morton–Willinger theorem have been given in [17, 20, 21, 26, 31].

An important question left unanswered by Kreps was whether one can, in general, replace the use of nets $(m_\alpha - h_\alpha)_{\alpha \in I}$, indexed by α ranging in a general ordered set I , simply by sequences $(m_n - h_n)_{n=1}^\infty$. In the context of continuous processes, $S = (S_t)_{0 \leq t \leq T}$, a positive answer was given by Delbaen in [6], if one is willing to make the harmless modification to replace the deterministic times $0 = t_0 \leq t_1 \leq \dots \leq t_n = T$ in equation (3) by *stopping times* $0 = \tau_0 \leq \tau_1 \leq \dots \leq \tau_n = T$. A second case, where the answer to this question is positive, are processes $S = (S_t)_{t=0}^\infty$ in infinite, discrete time as shown in [32].

The Banach–Steinhaus theorem implies that, for a sequence $(m_n - h_n)_{n=1}^\infty$ converging in $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ with respect to the weak-star (or Mackey) topology, the norms $(\|m_n - h_n\|_\infty)_{n=1}^\infty$ remain bounded (“uniform boundedness principle”). Therefore, it follows that in the above two cases of continuous processes $S = (S_t)_{0 \leq t \leq T}$ or processes $(S_t)_{t=0}^\infty$ in infinite, discrete time, the “no free lunch” condition of Kreps can be equivalently replaced by the “no free lunch

with bounded risk” condition introduced in [32]: in equation (7) above, we additionally impose that $(\|m_\alpha - h_\alpha\|_\infty)_{\alpha \in I}$ remains bounded. In this case, we have that there is a constant $M > 0$ such that $m_\alpha \geq -M$, $\mathbb{P}$ -a.s. for each $\alpha \in I$, which explains the wording “bounded risk”.

However, in the context of general semimartingale models $S = (S_t)_{0 \leq t \leq T}$, a counter-example was given by Delbaen and the author in ([Ex. 7.8] [7]) showing that the “no free lunch with bounded risk” condition does not imply the existence of an equivalent martingale measure. Hence, in a general setting and by only using simple integrals, there is no possibility of getting any more precise information on the free lunch condition than the one provided by Kreps’ theorem.

At this stage it became clear that, in order to obtain sharper results, one has to go beyond the framework of simple integrals (3) and rather use general stochastic integrals (1). After all, the simple integrals are only a technical gimmick, analogous to step functions in measure theory. In virtually all the applications, for example, the replication strategy of an option in the Black–Scholes model, one uses general integrals of the form (1).

General integrands pose a number of questions to be settled. First of all, the integral (1) has to be mathematically well defined. The theory of stochastic calculus starting with K. Itô, and developed in particular by the Strasbourg school of probability around Meyer, provides very precise information on this issue: there is a good integration theory for a given stochastic process $S = (S_t)_{0 \leq t \leq T}$ if and only if S is a semimartingale (theorem of Bichteler–Dellacherie).

Hence, mathematical arguments lead to the model assumption that S has to be a semimartingale. However, what about an economic justification of this assumption? Fortunately, the economic reasoning hints in the same direction. It was shown by Delbaen and the author that, for a locally bounded stochastic process $S = (S_t)_{0 \leq t \leq T}$, a very weak form of Kreps’ “no free lunch” condition involving simple integrands (3), implies already that S is a semimartingale (see [Theorem 7.2] [7], for a precise statement).

Hence, it is natural to assume that the model $S = (S_t)_{0 \leq t \leq T}$ of stock prices is a semimartingale so that the stochastic integral (3) makes sense mathematically, for all S -integrable, predictable processes $H = (H_t)_{0 \leq t \leq T}$. As pointed out, [14, 15] impose, in addition, an admissibility condition to rule out doubling strategies and similar schemes.

Definition 2 ([Def. 2.7] [7]): An S -integrable predictable process $H = (H_t)_{0 \leq t \leq T}$ is called *admissible* if there is a constant $M > 0$ such that

$$\int_0^t H_u dS_u \geq -M, \quad \text{a.s., for } 0 \leq t \leq T \quad (8)$$

The economic interpretation is that the economic agent, trading according to the strategy, has to respect a finite credit line M .

Let us now sketch the approach of [7]. Define

$$K = \left\{ \int_0^T H_t dS_t : H \text{ admissible} \right\} \quad (9)$$

which is a set of (equivalence classes of) random variables. Note that by equation (6) the elements $f \in K$ are uniformly bounded from below, that is, $f \geq -M$ for some $M \geq 0$. On the other hand, there is no reason why the positive part f_+ should obey any boundedness or integrability assumption.

As a next step, we “allow agents to throw away money” similarly as in Kreps’ work [24]. Define

$$\begin{aligned} C &= \{g \in L^\infty(\Omega, \mathcal{F}, \mathbb{P}) : g \leq f \text{ for some } f \in K\} \\ &= [K - L_+^0(\Omega, \mathcal{F}, \mathbb{P})] \cap L^\infty(\Omega, \mathcal{F}, \mathbb{P}) \end{aligned} \quad (10)$$

where $L_+^0(\Omega, \mathcal{F}, \mathbb{P})$ denotes the set of nonnegative measurable functions.

By construction, C consists of bounded random variables, so that we can use the functional analytic duality theory between L^∞ and L^1 . The difference of the subsequent definition to Kreps’ approach is that it pertains to the norm topology $\|\cdot\|_\infty$ rather than to the Mackey topology on $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 3 ([2.8] [11]): A locally bounded semimartingale $S = (S_t)_{0 \leq t \leq T}$ satisfies the *no free lunch with vanishing risk condition* if

$$\bar{C} \cap L_+^\infty(\Omega, \mathcal{F}, \mathbb{P}) = \{0\} \quad (11)$$

where $\bar{C}$ denotes the $\|\cdot\|_\infty$ -closure of C .

Here is the translation of equation (11) into prose: the process S fails the above condition if there is a function $g \in L_+^\infty(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{P}[g > 0] > 0$ and a sequence $(f^n)_{n=1}^\infty$ of the form

$$f^n = \int_0^T H_t^n dS_t \quad (12)$$

where H^n are admissible integrands, such that

$$f_n \geq g - \frac{1}{n} \quad a.s. \quad (13)$$

Hence the condition of no free lunch with vanishing risk is intermediate between the (stronger) no free lunch condition of Kreps and the (weaker) no-arbitrage condition. The latter would require that there is a nonnegative function g with $\mathbb{P}[g > 0] > 0$, which is of the form

$$g = \int_0^T H_t dS_t \quad (14)$$

for an admissible integrand H . Condition (13) does not quite guarantee this, but something — at least from an economic point of view — very close: we can *uniformly* approximate from below such a g by the outcomes f_n of admissible trading strategies.

The main result of Delbaen and the author [7] reads as follows.

Theorem 3 ([Corr. 1.2] [7]): *Let $S = (S_t)_{0 \leq t \leq T}$ be a locally bounded real-valued semimartingale.*

There is a probability measure Q on $(\Omega, \mathcal{F})$, which is equivalent to $\mathbb{P}$ and under which S is a local martingale if and only if S satisfies the condition of no free lunch with vanishing risk.

This is a mathematically precise theorem, which, in my opinion, is quite close to the vague “meta-theorem” at the beginning of this article. The difference to the intuitive “no arbitrage” idea is that the agent has to be willing to sacrifice (at most) the quantity $\frac{1}{n}$ in equation (13), where we may interpret $\frac{1}{n}$ as, say, 1 cent.

The proof of the above theorem is rather long and technical and a more detailed discussion goes beyond the scope of this article. To the best of the author’s knowledge, no essential simplification of this proof has been achieved so far ([19]).

Mathematically speaking, the statement of the theorem looks very suspicious at first glance: after all, the *no free lunch with vanishing risk* condition pertains to the *norm topology* of $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$. Hence it seems that, when applying the Hahn–Banach theorem, one can only obtain a linear functional in $L^\infty(\Omega, \mathcal{F}, \mathbb{P})^*$, which is not necessarily of the form $\frac{dQ}{d\mathbb{P}} \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, as we have seen in Ross’ work [29].

The reason why the above theorem, nevertheless, is true is a little miracle: it turns out ([Th. 4.2] [7])

that, under the assumption of no free lunch with vanishing risk, the set C defined in equation (10) is *automatically* weak-star closed in $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$. This pleasant fact is not only a crucial step in the proof of the above theorem; maybe even more importantly, it also found other applications. For example, to find general existence results in the theory of utility optimization (see **Expected Utility Maximization: Duality Methods**) it is of crucial importance to have a closedness property of the set over which one optimizes: for these applications, the above result is very useful [23].

Without going into the details of the proof, the importance of certain elements in the set K is pointed out. The admissibility rules out the use of doubling strategies. The opposite of such a strategy can be called a *suicide strategy*. It is the mathematical equivalent of making a bet at the roulette, leaving it as well as all gains on the table as long as one keeps winning, and wait until one loses for the first time. Such strategies, although admissible, do not reflect economic efficiency. More precisely, we define the following.

Definition 4 *An admissible outcome $\int_0^T H_t dS_t$ is called maximal if there is no other admissible strategy H' such that $\int_0^T H'_t dS_t \geq \int_0^T H_t dS_t$ with $\mathbb{P}[\int_0^T H'_t dS_t > \int_0^T H_t dS_t] > 0$*

In the proof of Theorem 6, these elements play a crucial role and the heart of the proof consists in showing that every element in K is dominated by a maximal element. However, besides their mathematical relevance, they also have a clear economic interpretation. There is no use in implementing a strategy that is not maximal as one can do better. Nonmaximal elements can also be seen as bubbles [18].

In Theorem 6, we only assert that S is a *local* martingale under Q . In fact, this technical concept cannot be avoided in this setting. Indeed, fix an S -integrable, predictable, admissible process $H = (H_t)_{0 \leq t \leq T}$ as well as a bounded, predictable, strictly positive process $(k_t)_{0 \leq t \leq T}$. The subsequent identity holds true trivially.

$$\int_0^t H_u dS_u = \int_0^t \frac{H_u}{k_u} d\tilde{S}_u, \quad 0 \leq t \leq T \quad (15)$$

where

$$\tilde{S}_u = \int_0^u k_v dS_v, \quad 0 \leq u \leq T \quad (16)$$

The message of equations (15) and (16) is that the class of processes obtained by taking admissible stochastic integrals on S or $\tilde{S}$ simply coincide. An easy interpretation of this rather trivial fact is that the possible investment opportunities do not depend on whether stock prices are denoted in euros or in cents (this corresponds to taking $k_t \equiv 100$ above).

However, it may very well happen that $\tilde{S}$ is a martingale while S only is a local martingale. In fact, the concept of local martingales may even be *characterized* in these terms ([Proposition 2.5] [10]): a semimartingale S is a local martingale if and only if there is a strictly positive, decreasing, predictable process k such that $\tilde{S}$ defined in equation (16) is a martingale.

Again we want to emphasize the role of the maximal elements. It turns out ([8, 11]) that if $\int_0^T H_t dS_t$ is maximal, *if and only if* there is an equivalent local martingale measure Q such that the process $\int_0^t H_u dS_u$ is a martingale and not just a local martingale under Q . One can show ([9, 11]) that for a given sequence of maximal elements $\int_0^T H_t^n dS_t$, one can find one and the same equivalent local martingale measure Q such that *all* the processes $\int_0^t H_u^n dS_u$ are Q -martingales. Another useful and related characterization ([8, 11]) is that if a process $V_t = x + \int_0^t H_u dS_u$ defines a maximal element $\int_0^T H_u dS_u$ and remains strictly positive, the whole financial market can be rewritten in terms of V as a new numéraire without losing the no-arbitrage properties. The change of numéraire and the use of the maximal elements allows to introduce a numéraire invariant concept of admissibility, see [9] for details. An important result in this article is that the sum of maximal elements is again a maximal element.

Theorem 6 above still contains one severe limitation of generality, namely, the local boundedness assumption on S . As long as we only deal with continuous processes S , this requirement is, of course, satisfied. However, if one also considers processes with jumps, in most applications it is natural to drop the local boundedness assumption.

The case of general semimartingales S (without any boundedness assumption) was analyzed in [10]. Things become a little trickier as the concept of local martingales has to be weakened even further: we refer

to **Equivalent Martingale Measures** for a discussion of the concept of sigma-martingales. This concept allows to formulate a result pertaining to a perfectly general setting.

Theorem 4 ([Corr. 1.2][7]): *Let $S = (S_t)_{0 \leq t \leq T}$ be an $\mathbb{R}^d$ -valued semimartingale.*

There is a probability measure Q on $(\Omega, \mathcal{F})$, which is equivalent to $\mathbb{P}$ and under which S is a sigma-martingale if and only if S satisfies the condition of no free lunch with vanishing risk with respect to admissible strategies.

One may still ask whether it is possible to formulate a version of the fundamental theorem, which does not rely on the concepts of local or sigma-, but rather on “true” martingales.

This was achieved by Yan [38] by applying a clever change of numéraire technique, (see **Change of Numéraire** also [Section 5] [13]): let us suppose that $(S_t)_{0 \leq t \leq T}$ is a *positive* semimartingale, which is natural if we model, for example, prices of shares (while the previous setting of not necessarily positive price processes also allows for the modeling of forwards, futures etc.).

Let us weaken the admissibility condition (8) above, by calling a predictable, S -integrable process *allowable* if

$$\int_0^t H_u dS_u \geq -M(1 + S_t) \quad a.s., \text{ for } 0 \leq t \leq T \quad (17)$$

The economic idea underlying this notion is well known and allows for the following interpretation: an agent holding M units of stock and bond may, in addition, trade in S according to the trading strategy H satisfying equation (17); the agent will then remain liquid during $[0, T]$.

By taking $S + 1$ as new numéraire and replacing admissible by allowable trading strategies, Yan obtains the following theorem.

Theorem 5 ([Theorem 3.2] [38]) *Suppose that S is a positive semimartingale.*

There is a probability measure Q on $(\Omega, \mathcal{F})$, which is equivalent to $\mathbb{P}$ and under which S is a martingale if and only if S satisfies the condition of no free lunch with vanishing risk with respect to allowable trading strategies.

References

- [1] Arrow, K. (1964). The role of securities in the optimal allocation of risk-bearing, *Review of Economic Studies* **31**, 91–96.
- [2] Bachelier, L. (1964). Théorie de la Spéculation, *Annales Scientifiques de l'É Normale Supérieure* **17**, 21–86. English translation in: Cootner, P. (ed), *The Random Character of Stock Market Prices*, MIT Press.
- [3] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- [4] Cox, J. & Ross, S. (1976). The valuation of options for alternative stochastic processes, *Journal of Financial Economics* **3**, 145–166.
- [5] Dalang, R.C., Morton, A. & Willinger, W. (1990). Equivalent Martingale measures and no-arbitrage in stochastic securities market model, *Stochastics and Stochastic Reports* **29**, 185–201.
- [6] Delbaen, F. (1992). Representing martingale measures when asset prices are continuous and bounded, *Mathematical Finance* **2**, 107–130.
- [7] Delbaen, F. & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing, *Mathematische Annalen* **300**, 463–520.
- [8] Delbaen, F. & Schachermayer, W. (1995). The no-arbitrage condition under a change of numéraire, *Stochastics and Stochastic Reports* **53**, 213–226.
- [9] Delbaen, F. & Schachermayer, W. (1997). The Banach space of workable contingent claims in arbitrage theory, *Annales de l'Institut Henri Poincaré (B) Probability and Statistics* **33**, 113–144.
- [10] Delbaen, F. & Schachermayer, W. (1998). The fundamental theorem of asset pricing for unbounded stochastic processes, *Mathematische Annalen* **312**, 215–250.
- [11] Delbaen, F. & Schachermayer, W. (2006). *The Mathematics of Arbitrage*, Springer Finance, Springer, p. 371.
- [12] Duffie, D. & Huang, C.F. (1986). Multiperiod security markets with differential information; martingales and resolution times, *Journal of Mathematical Economics* **15**, 283–303.
- [13] Guasoni, P., Rásonyi, M. & Schachermayer, W. (2009). The fundamental theorem of asset pricing for continuous processes under small transaction costs, *Annals of Finance*, forthcoming.
- [14] Harrison, J.M. & Kreps, D.M. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- [15] Harrison, J.M. & Pliska, S.R. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and their Applications* **11**, 215–260.
- [16] Harrison, J.M. & Pliska, S.R. (1983). A stochastic calculus model of continuous trading: complete markets, *Stochastic Processes and their Applications* **11**, 313–316.
- [17] Jacod, J. & Shiryaev, A.N. (1998). Local martingales and the fundamental asset pricing theorems in the discrete-time case, *Finance and Stochastics* **2**(3), 259–273.
- [18] Jarrow, R., Protter, P. & Shimbo, K. (2007). Asset price bubbles in complete markets, in *Advances in Mathematical Finance*, Appl. Numer. Harmon. Anal., Birkhäuser, Boston, Boston MA, pp. 97–121.
- [19] Kabanov, Y.M. (1997). On the FTAP of Kreps-Delbaen-Schachermayer (English), in *Statistics and Control of Stochastic Processes*, Y.M. Kabanov ed., World Scientific, Singapore, pp. 191–203. The Liptser Festschrift. Papers from the Steklov seminar held in Moscow, Russia, 1995–1996.
- [20] Kabanov, Y.M. & Kramkov, D. (1994). No-arbitrage and equivalent martingale measures: an elementary proof of the Harrison–Pliska theorem, *Theory of Probability and its Applications* **39**(3), 523–527.
- [21] Kabanov, Y.M. & Stricker, Ch. (2001). A teachers' note on no-arbitrage criteria, *Séminaire de Probabilités XXXV, Springer Lecture Notes in Mathematics* **1755**, 149–152.
- [22] Kemeny, J.G. (1955). Fair bets and inductive probabilities, *Journal of Symbolic Logic* **20**(3), 263–273.
- [23] Kramkov, D. & Schachermayer, W. (1999). The asymptotic elasticity of utility functions and optimal investment in incomplete markets, *Annals of Applied Probability* **9**(3), 904–950.
- [24] Kreps, D.M. (1981). Arbitrage and equilibrium in economics with infinitely many commodities, *Journal of Mathematical Economics* **8**, 15–35.
- [25] Merton, R.C. (1973). The theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.
- [26] Rogers, L.C.G. (1994). Equivalent martingale measures and no-arbitrage, *Stochastics and Stochastic Reports* **51**(1–2), 41–49.
- [27] Ross, S. (1976). The arbitrage theory of capital asset pricing, *Journal of Economic Theory* **13**, 341–360.
- [28] Ross, S. (1977). Return, risk and arbitrage, *Risk and Return in Finance* **1**, 189–218.
- [29] Ross, S. (1978). A simple approach to the valuation of risky streams., *Journal of Business* **51**, 453–475.
- [30] Samuelson, P.A. (1965). Proof that properly anticipated prices fluctuate randomly, *Industrial Management Review* **6**, 41–50.
- [31] Schachermayer, W. (1992). A Hilbert space proof of the fundamental theorem of asset pricing in finite discrete time, *Insurance: Mathematics and Economics* **11**(4), 249–257.
- [32] Schachermayer, W. (1994). Martingale Measures for Discrete time Processes with Infinite Horizon, *Mathematical Finance* **4**, 25–56.
- [33] Schachermayer, W. (2005). A note on arbitrage and closed convex cones, *Mathematical Finance* (1), forthcoming.
- [34] Schachermayer, W. & Teichmann, J. (2005). How close are the option pricing formulas of Bachelier and Black-Merton-Scholes? *Mathematical Finance* **18**(1), 55–76.

- [35] Shimony, A. (1955). Coherence and the axioms of confirmation, *The Journal of Symbolic Logic* **20**, 1–28.
- [36] Stricker, Ch. (1990). Arbitrage et Lois de Martingale, *Annales de l'Institut Henri Poincaré—Probabilités et Statistiques* **26**, 451–460.
- [37] Yan, J.A. (1980). Caractérisation d' une classe d'ensembles convexes de L^1 ou H^1 , in *Séminaire de Probabilités XIV*, J. Azema, M. Yor, eds, Springer Lecture Notes in Mathematics 784, Springer, pp. 220–222.
- [38] Yan, J.A. (1998). A new look at the fundamental theorem of asset pricing, *Journal of Korean Mathematics Society* **35**, 659–673.
- [39] Yor, M. (1978). Sous-espaces denses dans L^1 ou H^1 et représentation des martingales, in *Séminaire de Probabilités XII*, Springer Lecture Notes in Mathematics, Springer, Vol. 649, pp. 265–309.

Related Articles

Arbitrage Strategy; Arrow, Kenneth; Change of Numeraire; Equivalent Martingale Measures; Martingales; Martingale Representation Theorem; Risk-neutral Pricing; Stochastic Integrals.

WALTER SCHACHERMAYER

Risk-neutral Pricing

A classical problem arising frequently in business is the valuation of future cash flows that are risky. By the term *risky* we mean that the payment is not of a deterministic nature; rather there is some uncertainty in the amount of the future cash flows. Of course, in real life, virtually everything happening in the future contains some element of uncertainty.

As an example, let us think of an investment project, say, a company plans to build a new factory. A classical way to proceed is to calculate a *net asset value*. One tries to estimate the future cash flows generated by the project in the subsequent periods. In the present example, they will initially be negative; this initial investment should be compensated by the positive cash flows in the later periods. Having fixed these estimates of the future cash flows for all periods, one calculates a *net asset value* by discounting these cash flows to the present date. But, of course, there is uncertainty involved in the estimation of the future cash flows and people doing these calculations are, of course, aware of that. The usual way to compensate for this uncertainty is to apply an interest rate that is higher than the *riskless^a rate of return* corresponding to the rate of return of government bonds.

The spread between the riskless rate of return and the interest rate used for discounting the future cash flows in the calculation of the net asset value can be quite substantial in order to compensate for the riskiness. Only if the net asset value, obtained by discounting with a rather high rate of return, remains positive, the management of the company will engage in the investment project.

Mathematically speaking, the above procedure may be described as follows: first, one determines the *expected values* of the future cash flows and, subsequently, one discounts by using an elevated discount factor. However, there is no systematic way of mathematically approaching the question of how the degree of uncertainty in the determination of the expected values can be quantified, and in which way this should be taken into account to determine the spread between the interest rates.

We now turn to a different approach, which interchanges the roles of taking expectations and discounting in taking the riskness of the cash flows into

account. This approach is used in modern mathematical finance, in particular, in the Black–Scholes formula. However, the idea goes back much further and the method was used by actuaries for centuries.

Think of a life insurance contract. To focus on the essential point, we consider the simplest case: a one-year death insurance. If the insured person dies within the subsequent year, the insured sum S , say $S = €1$, is paid out at the end of this year; if the insured person survives the year, nothing is paid, and the contract ends at the end of the year.

To calculate the premium^b for this contract, actuaries look up in their mortality tables^c the probability that the insured person dies within one year. The traditional notation for this probability is q_x , where x denotes the age of the insured person.

To calculate the premium for such a one-year death insurance contract, with S normalized to $S = 1$, actuaries apply the formula

$$P = \frac{1}{1+i} q_x \quad (1)$$

The term q_x is just the expected value of the future cash flow and i denotes “the” interest rate: hence the premium P is the *discounted expected value* of the cash flow at the end of the year.

It is important to note that actuaries use a “conservative” value for the interest rate, for example, $i = 3\%$. In practical terms, this corresponds quite well to the “riskless rate of return”. In any case, it is quite different, in practical as well as in theoretical terms, from the discount factors used to calculate the net asset value of a risky future cash flow according to the method stated above.

But, after all, the premium of our death insurance contract also corresponds to the present value of an uncertain future cash flow! How do actuaries account for the risk involved in this cash flow, if not *via* an appropriate choice of the interest rate?

The answer is simple when looking at equation (1): apart from the interest rate i the probability q_x of dying within the next year also enters the calculation of P . The art of the actuarial profession is to choose the “good” value for q_x . Typically, actuaries very well know the actual mortality probabilities in their portfolio of contracts, which often consists of several hundred thousand contracts; in other words, they have a very good understanding of what the “true value” of q_x is. However, they do not apply this “true value” in their premium calculations: in equation (1) they

would apply a value for q_x which is substantially higher than the “true” value of q_x . Actuaries speak about mortality tables of the first kind and the second kind.

Mortality tables of the second kind reflect the “true probabilities”. They are only used for the internal analysis of the profitability of the insurance company. On the other hand, in the daily life of actuaries only the mortality tables of the first kind, which properly display the “modified” probabilities q_x , are used. They are not only used for the calculation of premia but also for all quantities of relevance involved in an insurance policy, such as surrender values, reserves, and so on. This constitutes a big strength of the actuarial technique: actuaries are always armed with perfectly coherent logic when doing all these calculations. This logic is that of a fair game or, mathematically speaking, of a martingale. Indeed, *if the q_x would correctly model the mortality of the insured person and if i were the interest rate that the insurance company could precisely achieve when investing the premia, then the premium calculation (1) would make the insurance contract a fair game.*

It is important to note that this argument pertains only to a kind of virtual world, as it is precisely the task of actuaries to choose the mortalities q_x in a prudent way such that they do *not* coincide with the “true” probabilities. In the case of insurance contracts where the insurance company has to pay in the case of death, actuaries choose the probabilities q_x higher than the “true ones”. This happens in the simple example considered above. On the other hand, if the insurance company has to pay when the insured person is still alive, for example, in the case of a pension, actuaries use probabilities q_x which are lower than the “true ones”, in order to be on the safe side.

These actuarial techniques have been elaborated on as this will be helpful to more clearly understand the essence of the option pricing approach of Black, Scholes, and Merton. Their well-known model for the risky stock S and the risk-free bond are

$$\begin{aligned} dS_t &= S_t \mu dt + S_t \sigma dW_t \\ dB_t &= B_t r dt \end{aligned} \quad (2)$$

The task is to value a (European) derivative on the stock S at expiration time T , for example, $C_T = (S_T - K)_+$. As explained earlier (see **Complete**

Markets), the solution proposed by Black, Scholes, and Merton is

$$C_0 = e^{-rT} \mathbb{E}_Q[C_T] \quad (3)$$

The above equation is a perfect analog to the premium of a death insurance contract (1). The first term, taking care of the discounting, uses the “conservative” choice of a riskless interest rate r . The second term gives the expected value of the future cash flow, taken under the *risk-neutral* probability measure Q . This probability measure Q is chosen in such a way that the dynamics (2) of the stock under Q become

$$dS_t = S_t r dt + S_t \sigma dW_t \quad (4)$$

The point is that the drift term $S_t r dt$ of S under Q is in line with the growth rate of the risk-free bond

$$dB_t = B_t r dt \quad (5)$$

The interpretation of (4) is that *if the market were correctly modeled by the probability Q* , then the market was *risk neutral*. The mathematical formulation, $(e^{-rt} S_t)_{0 \leq t \leq T}$, that is, the stock price process discounted by the risk-free interest rate r , is a martingale under Q .

Similarly as in the actuarial context above, the mathematical model of a financial market under the risk-neutral measure Q pertains to a virtual world, and not to the real world. In reality, that is, under $\mathbb{P}$, we would typically have $\mu > r$. Fixing this case, Girsanov’s formula (see **Equivalence of Probability Measures; Stochastic Exponential**) tells us precisely that the probability measure Q represents a “prudent choice of probability”. It gives less weight than the original measure $\mathbb{P}$ to the events which are favorable for the buyer of a stock, that is, when S_T is large. On the other hand, Q gives more weight than $\mathbb{P}$ to unfavorable events, that is, when S_T is small. This can be seen from Girsanov’s formula

$$\frac{dQ}{d\mathbb{P}} = \exp \left[-\frac{\mu - r}{\sigma} W_T - \frac{(\mu - r)^2}{2\sigma^2} T \right] \quad (6)$$

and the dynamics of the stock price process S under $\mathbb{P}$ resulting from (2)

$$S_T = S_0 \exp \left[\sigma W_T + \left(\mu - \frac{\sigma^2}{2} \right) T \right] \quad (7)$$

Fixing a random element $\omega \in \Omega$, the Radon–Nikodym derivative $\frac{dQ}{dP}(\omega)$ is small iff $W_T(\omega)$ is large, and the latter is large iff $S_T(\omega)$ is large.

In many applications, it is not even necessary to consider the original “true” probability measure $\mathbb{P}$. There are hundreds of papers containing the sentence: “we work under the risk-neutral measure Q ”. This is parallel to the situation of an actuary in his/her daily work: He/she does not bother about the “true” mortality probabilities, but only about the probabilities listed in the mortality table of the first kind.

The history of the valuation formula (3), in fact, goes back much further than Black, Scholes, and Merton. Already in 1900, L. Bachelier applied this formula in his thesis [1] in order to price options. It seems worthwhile to have a closer look. Bachelier did not use a discount factor, such as e^{-rT} , in equation (3). The reason is that in 1900 prices underlying the option were denoted in forward prices at the Paris stock exchange (called “true prices” by Bachelier who also carefully adjusted for coupon payments; see [6] for details). As it is well known, when considering forward prices the discount factor disappears. In modern terminology, this fact boils down to “Black’s formula”.

As regards the second term in equation (3), Bachelier started from the very beginning with a martingale model, namely, (scaled) Brownian motion [6]

$$S_t = S_0 + \sigma W_t, \quad 0 \leq t \leq T \quad (8)$$

In other words, he also “worked assuming the risk-neutral probability”.

In fact, in the first pages of his thesis Bachelier does speak about *two kinds of probabilities*. The following is a quote from [1]:

- (i) The probability which might be called “mathematical”, which can be determined a priori and which is studied in games of chance.
 - (ii) The probability dependent on future events and, consequently impossible to predict in a mathematical manner.
- This latter is the probability that the speculator tries to predict.

Admitting a large portion of goodwill and hindsight knowledge one might interpret (i) something like the risk-neutral probability Q , while (ii) describes something like the historical measure $\mathbb{P}$.

Risk-neutral Pricing for General Models

In the Black–Scholes model (2) there is only one *risk-neutral* measure Q under which the discounted stock price process becomes a martingale.^d

This feature characterizes *complete financial markets* (see **Complete Markets**). In this case, we not only obtain from equation (3) a price C_0 for the derivative security C_T , but we get much more: the derivative can be perfectly replicated by starting at time $t = 0$ with the initial investment given by equation (3) and subsequent dynamical trading in the underlying stock S . This is the essence of the approach of Black, Scholes, and Merton; it has no parallel in the classical actuarial approach or in the work of L. Bachelier.

What happens in incomplete financial markets, that is, when there is more than one risk-neutral measure Q ? It has been shown by Harrison and Pliska [4] that equation (3) yields *precisely* all the *consistent pricing rules* for derivatives on S , when Q runs through the set of risk-neutral measures equivalent to $\mathbb{P}$. We denote the latter set by $\mathcal{M}^e(S)$. The term *consistent* means that there should be no-arbitrage possibilities when all possible derivatives on S are traded at the price given by equation (3).

But, what is the good choice of $Q \in \mathcal{M}^e(S)$? In general, this question is as meaningless as the question: what is the good choice of an element in some convex subset of a vector space? In order to allow for a more intelligent version of this question, one needs additional information. It is here that the original probability measure $\mathbb{P}$ comes into play again: a popular approach is to choose the element $Q \in \mathcal{M}^e(S)$ which is “closest” to $\mathbb{P}$.

In order to make this idea precise, fix a strictly convex function $V(y)$, for example,

$$V(y) = y(\ln(y) - 1), \quad y > 0 \quad (9)$$

$$\text{or} \quad V(y) = \frac{y^2}{2}, \quad y \in \mathbb{R} \quad (10)$$

Determine $\hat{Q} \in \mathcal{M}^e(S)$ as the optimizer of the optimization problem

$$\mathbb{E} \left[V \left(\frac{dQ}{dP} \right) \right] \rightarrow \min! \quad Q \in \mathcal{M}^e(S) \quad (11)$$

To illustrate things at the hand of the above examples: For $V(y) = y(\ln(y) - 1)$, this corresponds to

choosing the element $\hat{Q} \in \mathcal{M}^e(S)$ minimizing the relative entropy $H(Q|\mathbb{P}) = \mathbb{E}_Q\left[\ln\left(\frac{dQ}{d\mathbb{P}}\right)\right]$; for $V(y) = \frac{y^2}{2}$, this corresponds to choosing $Q \in \mathcal{M}^e(S)$ minimizing the L^2 -norm $\|\frac{dQ}{d\mathbb{P}}\|_{L^2(\mathbb{P})} = \mathbb{E}_{\mathbb{P}}\left[\left(\frac{dQ}{d\mathbb{P}}\right)^2\right]^{\frac{1}{2}}$.

Under appropriate conditions, the minimization problem (11) has a solution, which then is unique by the strict convexity assumption.

There is an interesting connection to the issue of **Utility Indifference Valuation**. Let $U(x)$ be the (negative) Legendre–Fenchel transform of V , that is,

$$U(x) = \inf_y \{-xy + V(y)\} \quad (12)$$

For the two examples above, we obtain

$$U(x) = -e^{-x} \quad (13)$$

$$\text{or } U(x) = -\frac{x^2}{2} \quad (14)$$

which may be interpreted as utility functions. It turns out that—under appropriate assumptions—the optimizer $\hat{Q}$ in equation (11) yields precisely the marginal utility indifference pricing rule when plugged into equation (3) (see **Utility Indifference Valuation**).

In particular, we may conclude that pricing by marginal utility [2, 3, 5] is a consistent pricing rule in the sense of Harrison and Kreps.

End Notes

^aIn real life nothing is actually *riskless*: in practice, the *riskless rate of return* corresponds to government bonds (provided that the government is reliable).

^bWe do not consider costs, taxes, and so on, which are eventually added to this premium; we only consider the “net premium”.

^cA mortality table (horrible word!) is nothing but a list of probabilities q_x , where x runs through the relevant ages, say $x = 18, \dots, 110$. The first mortality table was constructed by Edmond Halley in 1693.

^dTo be precise: this result only holds true if for the underlying filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$ we have $\mathcal{F} = \mathcal{F}_T$ and the filtration $(\mathcal{F}_t)_{0 \leq t \leq T}$ is generated by $(S_t)_{0 \leq t \leq T}$.

References

- [1] Bachelier, L. (1964). Théorie de la Spéculation, *Annales Scientifiques de l'École Normale Supérieure* **17**, 21–86; English translation in: P. Cootner (ed.) (1900). *The Random Character of stock market prices*, MIT Press.
- [2] Davis, M. (1997). Option pricing in incomplete markets, in *Mathematics of Derivative Securities*, M.A.H. Dempster & S.R. Pliska, eds, Cambridge University Press, pp. 216–226.
- [3] Foldes, D. (2000). Valuation and martingale properties of shadow prices: an exposition, *Journal of Economic Dynamics and Control* **24**, 1641–1701.
- [4] Harrison, J.M. & Pliska, S.R. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and their Applications* **11**, 215–260.
- [5] Rubinstein, M. (1976). The valuation of uncertain income streams and the pricing of options, *Bell Journal of Economics* **7**, 407–426.
- [6] Schachermayer, W. (2003). Introduction to the mathematics of financial markets, in *Lecture Notes in Mathematics 1816 - Lectures on Probability Theory and Statistics, Saint-Flour Summer School 2000* (Pierre Bernard, editor), S. Albeverio, W. Schachermayer & M. Talagrand, eds, Springer Verlag, Heidelberg, pp. 111–177.

Further Reading

- Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- Delbaen, F. & Schachermayer, W. (2006). *The Mathematics of Arbitrage*, Springer Finance, p. 371.
- Merton, R.C. (1973). The theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.
- Ross, S. (1978). A simple approach to the valuation of risky streams, *Journal of Business* **51**, 453–475.
- Samuelson, P.A. (1965). Proof that properly anticipated prices fluctuate randomly, *Industrial Management Review* **6**, 41–50.

Related Articles

Change of Numeraire; Complete Markets; Equivalent Martingale Measures; Fundamental Theorem of Asset Pricing; Model Calibration; Monte Carlo Simulation; Pricing Kernels; Stochastic Discount Factors.

WALTER SCHACHERMAYER

Hedging

In a complete market (*see Complete Markets*) derivative securities are redundant in the sense that they can be replicated by the gains from trading *via* a self-financing admissible strategy in the underlying asset. This replicating strategy is then called the *hedging strategy* for the claim.

More formally, we fix some filtered probability space $(\Omega, \mathcal{A}, (\mathcal{F}_t), P)$. The (discounted) price process of a risky asset is modeled by an $(\mathcal{F}_t)$ -adapted semimartingale S . A *claim* B is an $\mathcal{F}_T$ -measurable random variable, where T is the maturity of the claim. B is *attainable* if there exists a constant c and an admissible strategy ϑ such that

$$B = c + \int_0^T \vartheta_t dS_t \quad (1)$$

The quintuple $(\Omega, \mathcal{A}, (\mathcal{F}_t), P, S)$ models a financial market. A market is *complete* if all bounded claims are attainable. Finally, a market that is not complete is called *incomplete*.

In case there exists an equivalent martingale measure (*see Equivalent Martingale Measures*) Q for S in a complete market, it must be unique according to some version of the second fundamental theorem of asset pricing (*see Second Fundamental Theorem of Asset Pricing*). Moreover, S has the predictable representation property (PRP) (*see Martingale Representation Theorem*) with respect to (w.r.t.) $(Q, (\mathcal{F}_t))$, meaning that every $(Q, (\mathcal{F}_t))$ -martingale can be written as a sum of its initial value and a stochastic integral w.r.t. S . These facts can be used to show existence of an optimal hedging strategy as follows: we consider for each bounded claim B the associated Q -martingale V given by

$$V_t = E_Q[B | \mathcal{F}_t], \quad t \leq T \quad (2)$$

By the PRP, there exists an admissible strategy ϑ such that

$$V_t = V_0 + \int_0^t \vartheta_u dS_u, \quad t \leq T \quad (3)$$

In particular, for $t = T$, we get

$$B = E_Q[B] + \int_0^T \vartheta_t dS_t \quad (4)$$

To calculate ϑ , note that we can express ϑ as the (symbolic) differential of angle bracket processes (w.r.t. Q),

$$\vartheta_t = \frac{d \langle V, S \rangle_t}{d \langle S \rangle_t} \quad (5)$$

which then in turn can often be evaluated by assuming more specific structures for the price process S and the claim B .

Quadratic Risk Minimization

In incomplete markets, one can in general not hedge a claim perfectly, and hence, there will always be some remaining risk which can be minimized according to various criteria. The Föllmer–Sondermann (FS) [5] approach consists in an orthogonal projection in $L^2(Q)$ of a square-integrable claim B onto the subspace spanned by the constants and stochastic integrals w.r.t. the price process S (which we assume to be locally square-integrable). Here, Q is some martingale measure for S that has been obtained either *via* calibration or according to some optimality criterion.

More precisely, given a claim $B \in L^2(Q, \mathcal{F}_T)$, we want to minimize

$$E_Q \left[\left(B - c - \int_0^T \vartheta_t dS_t \right)^2 \right] \quad (6)$$

over all constants c and all $\vartheta \in L^2(S)$, that is, predictable processes ϑ , such that $E_Q \left[\int_0^T \vartheta_t^2 d[S]_t \right] < \infty$. Hence, the goal is to project B onto the linear space

$$K = \left\{ c + \int_0^T \vartheta_t dS_t : c \in \mathbb{R}, \vartheta \in L^2(S) \right\} \subset L^2(Q) \quad (7)$$

For ϑ as above we also denote

$$K_0 = \left\{ \int_0^T \vartheta_t dS_t : \vartheta \in L^2(S) \right\} \subset L^2(Q) \quad (8)$$

By its very construction, the stochastic integral yields an isometry (here, we understand $[S]$ as the measure on $[0, T]$ which is associated with the increasing process $[S]$)

$$K_0 \cong L^2(\Omega \times [0, T], Q \otimes [S]) \quad (9)$$

$$\int_0^T \vartheta_t dS_t \longleftrightarrow \vartheta \quad (10)$$

since we have

$$E_Q \left[\left(\int_0^T \vartheta_t dS_t \right)^2 \right] = E_Q \left[\int_0^T \vartheta_t^2 d[S]_T \right] \quad (11)$$

Hence, K_0 is isometrically isomorphic to an L^2 -space and therefore closed. Therefore, we can apply the theorem about the orthogonal projection in the Hilbert spaces to get a decomposition

$$B = c^B + \int_0^T \vartheta_t^B dS_t + L_T \quad (12)$$

where L_T is orthogonal to each element of K ; in particular, $E_Q[L_T] = 0$ since $1 \in K$. It follows that we have $c^B = E_Q[B]$, and ϑ^B is called the *FS optimal hedging strategy*. As processes, $L_t := E[L_T | \mathcal{F}_t]$ and S are strongly orthogonal in the sense that LS is a Q -martingale or equivalently, $\langle L, S \rangle = 0$, where the predictable covariation $\langle \cdot, \cdot \rangle$ here, refers to the measure Q . This implies

$$\int \vartheta^B d\langle V, S \rangle = \langle S, S \rangle \quad (13)$$

where $V_t := E_Q[B | \mathcal{F}_t]$ denotes the martingale generated by B . Moreover, a simple calculation yields

$$E_Q[L_T^2] = E_Q \left[\langle V, V \rangle_T - \int_0^T (\vartheta_t^B)^2 \langle S, S \rangle_t \right] \quad (14)$$

Equation (13) is sometimes written as

$$\vartheta^B = \frac{d\langle V, S \rangle}{d\langle S, S \rangle} \quad (15)$$

We call

$$V = c^B + \int \vartheta^B dS + L \quad (16)$$

the *Galtchouk–Kunita–Watanabe* (GKW) decomposition of B or rather V relative to S .

In some models, one can compute the optimal (risk-minimizing) hedging strategy by solving a partial integro-differential equation [1] or by a generalized Clark–Ocone formula from Malliavin calculus [2].

Utility-indifference Hedging

Let u be some utility function defined on the whole real line. If there exists a number π satisfying

$$\begin{aligned} & \sup_{\vartheta} E \left[u \left(x + \int_0^T \vartheta_t dS_t \right) \right] \\ &= \sup_{\vartheta} E \left[u \left(x + \int_0^T \vartheta_t dS_t + \pi - B \right) \right] \end{aligned} \quad (17)$$

then it is called *utility-indifference price* of the claim B . It is the threshold where the investor is indifferent whether just to maximize expected utility from a pure investment into the stock with the price process S or to sell in addition a claim B and collect a premium π for this.

The optimal strategies ϑ on both sides of equation (17) typically differ. The difference

$$\theta := \phi^B - \phi^0 \quad (18)$$

of the optimizers on the right- and the left-hand side respectively can be interpreted as a *utility-based hedging strategy*. It corresponds to the adjustment of the investor's portfolio made in order to account for the option.

Let us consider exponential utility

$$u(x) = 1 - \exp(-\alpha x) \quad (19)$$

where $\alpha > 0$. If θ^α denotes the exponential utility-based hedging strategy corresponding to selling α units of the claim B , then it turns out that under quite general conditions the associated normalized gains $\frac{1}{\alpha} \int_0^T \theta_t^\alpha dS_t$ converge in $L^2(Q^0)$, Q^0 being the minimal entropy martingale measure, to $\int_0^T \vartheta_t^B dS_t$. Here, ϑ^B is the integrand coming from the GKW decomposition (12) w.r.t. Q^0 ; see [6] and the references contained therein.

Further Approaches to Hedging

Ideally, one would like to find a hedging strategy that always allows one to superreplicate the claim B . Finding such a strategy is related to the optional decomposition theorem for supermartingales which are bounded from below. However, it turns out that pursuing such a *superhedging strategy* is too expensive in the sense that the corresponding price

typically equals the highest price consistent with no-arbitrage pricing, that is, it amounts to $\sup_Q E_Q[B]$, where the supremum is taken over all the equivalent martingale measures Q .

Therefore, it has been proposed by Föllmer and Leukert [3] to maximize the probability of a successful hedge given a certain amount of initial capital, a concept that they call *quantile hedging*. However, with this approach there is no protection for the worst case scenarios other than portfolio diversification, and technically, it might be difficult to implement this since it corresponds to hedging a knock-out option. The same authors [4], moreover, considered *efficient hedges* which minimize the expected shortfall weighted by some loss function. In this way, the investor may interpolate between the extremes of no hedge and a superhedge, depending on the accepted level of shortfall risk.

References

- [1] Cont R., Tankov P. & Voltchkova E. (2007). Hedging with options in presence of jumps, in *Stochastic Analysis and Applications: The Abel Symposium 2005 in honor of Kiyosi Ito*, F.E. Benth, G. Di Nunno, T. Lindstrom, B. Øksendal & T. Zhang, eds, Springer, pp. 197–218.
- [2] Di Nunno, G. (2002). Stochastic integral representation, stochastic derivatives and minimal variance hedging, *Stochastics and Stochastics Reports* **73**, 181–198.
- [3] Föllmer, H. & Leukert, P. (1999). Quantile hedging, *Finance and Stochastics* **3**, 251–273.
- [4] Föllmer, H. & Leukert, P. (2000). Efficient hedging: cost versus shortfall risk, *Finance and Stochastics* **4**, 117–146.
- [5] Föllmer, H. & Sondermann, D. (1986). Hedging of non-redundant contingent claims. Contributions to mathematical economics, in *Honor of G. Debreu*, W. Hildenbrand & A. Mas-Colell, eds, Elsevier Science Publications, North-Holland, pp. 205–223.
- [6] Kallsen, J. & Rheinländer, T. (2008). *Asymptotic Utility-based Pricing and Hedging for Exponential Utility*. Preprint.

Related Articles

Complete Markets; Delta Hedging; Equivalent Martingale Measures; Mean–Variance Hedging; Option Pricing; General Principles; Second Fundamental Theorem of Asset Pricing; Stochastic Integrals; Superhedging; Uncertain Volatility Model; Utility Indifference Valuation.

THORSTEN RHEINLÄNDER

Complete Markets

According to the arbitrage pricing of derivative securities, the arbitrage price of a financial derivative is defined as the wealth of a self-financing trading strategy based on traded primary assets, which replicates the terminal payoff at maturity (or, more generally, all cash flows) from the financial derivative. Hence, an important issue arises whether any financial derivative admits a replicating strategy in a given model; if this property holds, then the market model is said to be *complete*. Completeness of a market model ensures that any derivative security can be priced by arbitrage and hedged by a dynamic trading in primary traded assets. For example, in the framework of the Cox, Ross, and Rubinstein [9] model, not only the call and put options but also any path-independent or path-dependent contingent claim can be replicated by a dynamic trading in stock and bond. Similarly, the classic Black and Scholes [3] model enjoys the property of completeness, although a suitable technical assumption needs to be imposed on the class of considered contingent claims.

Even for an incomplete model, the class of hedgeable derivatives, formally represented by *attainable* contingent claims, can be sufficiently large for practical purposes. Therefore, completeness should not be seen as a necessary requirement, as opposed to the no-arbitrage property, which is an indispensable feature of any financial model used for arbitrage pricing of derivative securities.

Finite Market Models

The issue of completeness of a *finite* market model was analyzed, among others, by Taqqu and Willinger [24]. The finiteness of a market means that the underlying probability space is finite, $\Omega = \{\omega_1, \omega_2, \dots, \omega_d\}$, and trading activities may only occur at the finite set of dates, denoted as $\{0, 1, \dots, T\}$. As a standard example of a finite market model, one may quote, for instance, the Cox, Ingersoll, and Ross [9] binomial tree model (*see Binomial Tree*) or any its multinomial extensions.

Let $S^1, S^2, \dots, S^k$ be the stochastic processes describing the *spot* (or *cash*) prices of some non-dividend paying financial assets. As customary, we postulate that the price process of at least one asset

is given as a strictly positive process, so that it can be selected as a *numéraire asset*. Let us then assume that $S_t^k > 0$ for every $t \leq T$. To emphasize the special role of the process S^k , we will sometimes write B instead of S^k . We assume that all assets are perfectly divisible and the market is frictionless, that is, there are no restrictions on the short-selling of assets, transaction costs, taxes, and so on.

We consider a probability space $(\Omega, \mathcal{F}_T, \mathbb{P})$, which is equipped with a filtration $\mathbb{F} = (\mathcal{F}_t)_{t \leq T}$. A probability measure $\mathbb{P}$, to be interpreted as the real-life probability, is an arbitrary probability measure on $(\Omega, \mathcal{F}_T)$ such that $\mathbb{P}(\omega_i) > 0$ for every $i = 1, 2, \dots, d$. For convenience, we assume throughout that the σ -field $\mathcal{F}_0$ is trivial, that is, $\mathcal{F}_0 = \{\emptyset, \Omega\}$. All processes considered in what follows are assumed to be $\mathbb{F}$ -adapted.

Trading Strategies

The component ϕ_t^i of a trading strategy $\phi = (\phi^1, \phi^2, \dots, \phi^k)$ represents the number of units of the i th security held by an investor at time t . In other words, $\phi_t^i S_t^i$ is the amount of funds invested in the i th security at time t . Hence, the *wealth process* $V(\phi)$ of a trading strategy ϕ is given by the equality, for $t = 0, 1, \dots, T$,

$$V_t(\phi) = \sum_{i=1}^k \phi_t^i S_t^i \quad (1)$$

The initial wealth $V_0(\phi) = \phi_0 S_0$ is also referred to as the *initial cost* of ϕ .

A trading strategy ϕ is said to be *self-financing* whenever it satisfies the following condition, for every $t = 0, 1, \dots, T-1$,

$$\sum_{i=1}^k \phi_t^i S_{t+1}^i = \sum_{i=1}^k \phi_{t+1}^i S_{t+1}^i \quad (2)$$

In the financial interpretation, this condition means that the portfolio ϕ is revised at any date t in such a way that there are no infusions of external funds and no funds are withdrawn from the portfolio. We denote by Φ the vector space of all self-financing trading strategies. The *gains process* $G(\phi)$ of any trading strategy ϕ equals, for $t = 0, 1, \dots, T$,

$$G_t(\phi) = \sum_{u=0}^{t-1} \sum_{i=1}^k \phi_u^i (S_{u+1}^i - S_u^i) \quad (3)$$

with $G_0(\phi) = 0$. It can be checked that a trading strategy ϕ is self-financing if and only if the equality $V_t(\phi) = V_0(\phi) + G_t(\phi)$ holds for every $t = 0, 1, \dots, T$.

Replication and Arbitrage

A European contingent claim X with maturity T is an arbitrary $\mathcal{F}_T$ -measurable random variable. Since the space Ω is assumed to be a finite set with d elements, any claim X has the representation $X = (X(\omega_1), X(\omega_2), \dots, X(\omega_d)) \in \mathbb{R}^d$. Hence, the class $\mathcal{X}$ of all contingent claims that settle at T may be identified with the vector space $\mathbb{R}^d$.

A *replicating strategy* for the contingent claim X , which settles at time T , is a self-financing trading strategy ϕ such that $V_T(\phi) = X$. For any claim X , we denote by Φ_X the class of all replicating strategies for X .

The wealth process $V(\phi)$ of an arbitrary strategy ϕ from Φ_X is called a *replicating process* of X in $\mathcal{M}$. Finally, we say that a claim X is *attainable* in $\mathcal{M}$ if it admits at least one replicating strategy. We denote the class of all attainable claims by $\mathcal{A}$.

Definition 1 A market model $\mathcal{M}$ is said to be complete if every claim $X \in \mathcal{X}$ is attainable in $\mathcal{M}$ or, equivalently, if for every $\mathcal{F}_T$ -measurable random variable X there exists at least one trading strategy $\phi \in \Phi$ such that $V_T(\phi) = X$. In other words, a market model $\mathcal{M}$ is complete whenever $\mathcal{X} = \mathcal{A}$.

Let X be an arbitrary attainable claim that settles at time T . We say that X is *uniquely replicated* in $\mathcal{M}$ if it admits a unique replicating process in $\mathcal{M}$, that is, if the equality $V_t(\phi) = V_t(\psi)$, $t \in [0, T]$, holds for arbitrary trading strategies ϕ, ψ from Φ_X . Then the process $V(\phi)$ is termed the *wealth process* of X in $\mathcal{M}$.

Arbitrage Price

A trading strategy $\phi \in \Phi$ is called an *arbitrage opportunity* if $V_0(\phi) = 0$ and the terminal wealth of ϕ satisfies

$$\mathbb{P}(V_T(\phi) \geq 0) = 1 \quad \text{and} \quad \mathbb{P}(V_T(\phi) > 0) > 0 \quad (4)$$

where $\mathbb{P}$ is the real-world probability measure. We say that a market $\mathcal{M} = (S, \Phi)$ is *arbitrage free* if

there are no-arbitrage opportunities in the class Φ of all self-financing trading strategies.

It can be shown that if the market model $\mathcal{M}$ is arbitrage free, then any attainable contingent claim X is uniquely replicated in $\mathcal{M}$. The converse implication is not true, however, that is, the uniqueness of the wealth process of any attainable contingent claim does not imply the arbitrage-free property of a market, in general. Therefore, the existence and uniqueness of the wealth process associated with any attainable claim is insufficient to justify the term *arbitrage price*. Indeed, it is easy to give an example of a finite market in which all claims can be uniquely replicated, but there exists a strictly positive claim which can be replicated by a self-financing strategy with a negative initial cost.

Definition 2 Let the market model $\mathcal{M}$ be arbitrage free. Then the wealth process of an attainable claim X is called the *arbitrage price* of X in $\mathcal{M}$ and it is denoted by $\pi_t(X)$ for every $t = 0, 1, \dots, T$.

Risk-neutral Valuation Formula

Recall that we write $S^k = B$. Let us denote by S^* the process of *relative prices*, which equals, for every $t = 0, 1, \dots, T$,

$$\begin{aligned} S_t^* &= (S_t^1 B_t^{-1}, S_t^2 B_t^{-1}, \dots, S_t^k B_t^{-1}) \\ &= (S_t^{*1}, S_t^{*2}, \dots, S_t^{*(k-1)}, 1) \end{aligned} \quad (5)$$

where we denote $S^{*i} = S^i B^{-1}$. Recall that the probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $(\Omega, \mathcal{F})$ are said to be *equivalent* if, for any event $A \in \mathcal{F}$, the equality $\mathbb{P}(A) = 0$ holds if and only if $\mathbb{Q}(A) = 0$. Similarly, $\mathbb{Q}$ is said to be *absolutely continuous* with respect to $\mathbb{P}$ if, for any event $A \in \mathcal{F}$, the equality $\mathbb{P}(A) = 0$ implies that $\mathbb{Q}(A) = 0$. Clearly, if the probability measures $\mathbb{P}$ and $\mathbb{Q}$ are equivalent, then they are also equivalent to each other. The following concept is crucial in the so-called risk-neutral valuation approach.

Definition 3 A probability measure $\mathbb{P}^*$ on $(\Omega, \mathcal{F}_T)$ equivalent to $\mathbb{P}$ (absolutely continuous with respect to $\mathbb{P}$, respectively) is called a *equivalent martingale measure* for S^* (a *generalized martingale measure* for S^* , respectively) if the relative price S^* is a $\mathbb{P}^*$ -martingale with respect to the filtration $\mathbb{F}$.

An $\mathbb{F}$ -adapted, k -dimensional process $S^* = (S^{*1}, S^{*2}, \dots, S^{*k})$ is a $\mathbb{P}^*$ -martingale with respect to a filtration $\mathbb{F}$ if the equality

$$\mathbb{E}_{\mathbb{P}^*}(S_{t+1}^{*i} | \mathcal{F}_t) = S_t^{*i} \quad (6)$$

holds for every i and $t = 0, 1, \dots, T-1$.

We denote by $\mathcal{P}(S^*)$ and $\mathcal{Q}(S^*)$ the class of all equivalent martingale measures for S^* and the class of all generalized martingale measures for S^* , respectively, so that the inclusion $\mathcal{P}(S^*) \subseteq \mathcal{Q}(S^*)$ holds. It is not difficult to provide an example in which the class $\mathcal{P}(S^*)$ is empty, whereas the class $\mathcal{Q}(S^*)$ is not.

Definition 4 A probability measure $\mathbb{P}^*$ on $(\Omega, \mathcal{F}_T)$ equivalent to $\mathbb{P}$ (absolutely continuous with respect to $\mathbb{P}$, respectively) is called an equivalent martingale measure for $\mathcal{M} = (S, \Phi)$ (a generalized martingale measure for $\mathcal{M} = (S, \Phi)$, respectively) if for every trading strategy $\phi \in \Phi$ the relative wealth process $V^*(\phi) = V(\phi)B^{-1}$ is a $\mathbb{P}^*$ -martingale with respect to the filtration $\mathbb{F}$.

We write $\mathcal{P}(\mathcal{M})$ ($\mathcal{Q}(\mathcal{M})$, respectively) to denote the class of all equivalent martingale measures (of all generalized martingale measures, respectively) for $\mathcal{M}$. For conciseness, an equivalent martingale measure (a generalized martingale measure, respectively) is abbreviated as EMM (GMM, respectively). Note that an equivalent martingale measure is sometimes referred to as a *risk-neutral probability*.

It can be shown that a trading strategy ϕ is self-financing if and only if the relative wealth process $V^*(\phi) = V(\phi)B^{-1}$ satisfies, for every $t = 0, 1, \dots, T$,

$$V_t^*(\phi) = V_0^*(\phi) + \sum_{u=0}^{t-1} \sum_{i=1}^k \phi_u^i (S_{u+1}^{*i} - S_u^{*i}) \quad (7)$$

Therefore, for any $\phi \in \Phi$ and any GMM $\mathbb{P}^*$ the relative wealth $V^*(\phi)$ is a $\mathbb{P}^*$ -martingale with respect to the filtration $\mathbb{F}$. This leads to the following result.

Lemma 1 A probability measure $\mathbb{P}^*$ on $(\Omega, \mathcal{F}_T)$ is a GMM for the market model $\mathcal{M}$ if and only if it is a GMM for the relative price process S^* , that is, $\mathcal{P}(S^*) = \mathcal{P}(\mathcal{M})$ and $\mathcal{Q}(S^*) = \mathcal{Q}(\mathcal{M})$.

The next result shows that the existence of an EMM for $\mathcal{M}$ is a sufficient condition for the

no-arbitrage property of $\mathcal{M}$. Recall that trivially $\mathcal{P}(\mathcal{M}) \subseteq \mathcal{Q}(\mathcal{M})$ so that the class $\mathcal{Q}(\mathcal{M})$ is manifestly nonempty if $\mathcal{P}(\mathcal{M})$ is so.

Proposition 1 Assume that the class $\mathcal{P}(\mathcal{M})$ is nonempty. Then the market $\mathcal{M}$ is arbitrage free. Moreover, the arbitrage price process of any attainable contingent claim X , which settles at time T , is given by the risk-neutral valuation formula, for every $t = 0, 1, \dots, T$,

$$\pi_t(X) = B_t \mathbb{E}_{\mathbb{P}^*}(X B_T^{-1} | \mathcal{F}_t) \quad (8)$$

where $\mathbb{P}^*$ is any EMM (or GMM) for the market model $\mathcal{M}$.

It can be checked that the binomial tree model (see **Binomial Tree**) with deterministic interest rates is complete, whereas its extension in which the stock price is modeled by a trinomial tree is incomplete. Completeness relies, in particular, on the choice of traded primary assets. Hence, it is natural to ensure completeness of an incomplete model by adding new traded instruments (typically, plain-vanilla options).

Completeness of a Finite Market

We already know that if the set of equivalent martingale measures is nonempty, then the market model $\mathcal{M}$ is arbitrage free. It appears that this condition is also necessary for the no-arbitrage property of the market model $\mathcal{M}$.

Proposition 2 Suppose that the market model $\mathcal{M}$ is arbitrage free. Then the class $\mathcal{P}(\mathcal{M})$ of equivalent martingale measures for $\mathcal{M}$ is nonempty.

This leads to the following version of the *first fundamental theorem of asset pricing* (the *First FTAP*).

Theorem 1 A finite market model $\mathcal{M}$ is arbitrage free if and only if the class $\mathcal{P}(\mathcal{M})$ is nonempty, that is, there exists at least one equivalent martingale measure for $\mathcal{M}$.

In the case of a finite market model, this result was established by Harrison and Pliska [13]. For a probabilistic approach to the First FTAP we refer to Taqqu and Willinger [20], who examine the case of a finite market model, and to papers by Dalang *et al.* [10] and Schachermayer [23], who study the case of a discrete-time model with infinite state space.

The following fundamental result provides a relationship between the completeness property of a finite market model and the uniqueness (or nonuniqueness) of an EMM. Any result of this kind is commonly referred to as the *second fundamental theorem of asset pricing*.

Theorem 2 *Assume that a market model $\mathcal{M}$ is arbitrage free so that the class $\mathcal{P}(\mathcal{M})$ is nonempty. Then $\mathcal{M}$ is complete if and only if the uniqueness of an equivalent martingale measure for $\mathcal{M}$ holds.*

If an arbitrage-free market model is incomplete, not all claims are attainable and the class $\mathcal{P}(\mathcal{M})$ of equivalent martingale measures comprises more than one element. In that case, one can use the following result to determine whether a given contingent claim is attainable.

Corollary 1 *A contingent claim $X \in \mathcal{X}$ is attainable in an arbitrage-free market model $\mathcal{M}$ if and only if the map $\mathbb{P}^* \mapsto \mathbb{E}_{\mathbb{P}^*}(XB_T^{-1})$ from $\mathcal{P}(\mathcal{M})$ to $\mathbb{R}$ is constant.*

It follows from this result that if a claim is attainable, so that its arbitrage price is well defined, the price can be computed using the risk-neutral valuation formula under any of (possibly several) martingale measures. In addition, if the risk-neutral valuation formula yields the same result for any choice of an EMM for the market model at hand, then a given claim is necessarily attainable.

Multidimensional Black and Scholes Model

A multidimensional Black and Scholes model is a natural extension to a multiasset setup of the classic Black and Scholes [3] options pricing model. Let k denote the number of primary risky assets. For any $i = 1, \dots, k$, the price process S^i of the i th risky asset, referred to as the i th stock, is modeled as an Itô process (the dot $\cdot$ stands for the inner product in $\mathbb{R}^d$)

$$dS_t^i = S_t^i(\mu_t^i dt + \sigma_t^i \cdot dW_t) \quad (9)$$

with $S_0^i > 0$ or, more explicitly,

$$dS_t^i = S_t^i \left(\mu_t^i dt + \sum_{j=1}^d \sigma_t^{ij} dW_t^j \right) \quad (10)$$

where $W = (W^1, \dots, W^d)$ is a standard d -dimensional Brownian motion, defined on a filtered probability space $(\Omega, \mathbb{F}, \mathbb{P})$. We make the natural assumption that the underlying filtration $\mathbb{F}$ coincides with the filtration $\mathbb{F}^W$ generated by the Brownian motion W . The coefficients μ^i and σ^i follow bounded progressively measurable processes on the space $(\Omega, \mathbb{F}, \mathbb{P})$, with values in $\mathbb{R}$ and $\mathbb{R}^d$, respectively. An important special case is obtained by postulating that for every i the volatility coefficient σ^i is represented by a fixed vector in $\mathbb{R}^d$ and the appreciation rate μ^i is a real number.

For brevity, we write $\sigma = \sigma_t$ to denote the volatility matrix—that is, the time-dependent random matrix $[\sigma_t^{ij}]$, whose i th row specifies the volatility of the i th traded stock. The last primary security is the risk-free savings account B with the price process $S^{k+1} = B$ satisfying

$$dB_t = r_t B_t dt, \quad B_0 = 1 \quad (11)$$

for a bounded, nonnegative, progressively measurable interest rate process r . This means that, for every $t \in [0, T]$,

$$B_t = \exp \left(\int_0^t r_u du \right) \quad (12)$$

To ensure the absence of arbitrage opportunities, we postulate the existence of a d -dimensional, progressively measurable process γ such that the equality

$$r_t - \mu_t^i = \sum_{j=1}^d \sigma_t^{ij} \gamma_t^j = \sigma_t^i \cdot \gamma_t \quad (13)$$

is satisfied simultaneously for every $i = 1, \dots, k$ (for Lebesgue a.e. $t \in [0, T]$, with probability one). Note that the *market price for risk* γ is not uniquely determined, in general. Indeed, the uniqueness of a solution γ to this equation holds only if $d \leq k$ and the volatility matrix σ has the full rank for every $t \in [0, T]$.

For example, if $d = k$ and the volatility matrix σ is nonsingular (for Lebesgue a.e. $t \in [0, T]$, with probability one), then, for every $t \in [0, T]$,

$$\gamma_t = \sigma_t^{-1}(r_t \mathbf{1} - \mu_t) \quad (14)$$

where $\mathbf{1}$ denotes the d -dimensional vector with every component equal to one, and μ_t is the vector with components μ_t^i . For any process γ satisfying the

above equation, we introduce a probability measure $\mathbb{P}^*$ on $(\Omega, \mathcal{F}_T)$ by setting

$$\frac{d\mathbb{P}^*}{d\mathbb{P}} = \exp \left(\int_0^T \gamma_u \cdot dW_u - \frac{1}{2} \int_0^T \|\gamma_u\|^2 du \right), \quad \mathbb{P}\text{-a.s.} \quad (15)$$

provided that the right-hand side in the last formula is well defined. The Doléans (stochastic) exponential

$$\eta_t = \exp \left(\int_0^t \gamma_u \cdot dW_u - \frac{1}{2} \int_0^t \|\gamma_u\|^2 du \right) \quad (16)$$

is known to be a strictly positive supermartingale (but not necessarily a martingale) under $\mathbb{P}$, since it may happen that $\mathbb{E}_{\mathbb{P}^*}(\eta_T) < 1$. A probability measure $\mathbb{P}^*$ equivalent to $\mathbb{P}$ is well defined if and only if the process η follows a $\mathbb{P}$ -martingale, that is, when $\mathbb{E}_{\mathbb{P}^*}(\eta_T) = 1$. For the last property to hold, it is enough (but not necessary) that γ is a bounded process.

Assume that the class of martingale measures is nonempty. By virtue of the Girsanov theorem, the process W^* , which equals, for every $t \in [0, T]$,

$$W_t^* = W_t - \int_0^t \gamma_u du \quad (17)$$

is a d -dimensional standard Brownian motion on $(\Omega, \mathbb{F}, \mathbb{P}^*)$. It follows from the Itô formula that the discounted stock price $S_t^{*i} = S_t^i B_t^{-1}$ satisfies under $\mathbb{P}^*$

$$dS_t^{*i} = S_t^{*i} \sigma_t^i \cdot dW_t^* \quad (18)$$

for any $i = 1, \dots, k$. This means that the discounted prices of all stocks follow local martingales under $\mathbb{P}^*$, so that any probability measure described above is a martingale measure for our model and it corresponds to the choice of the savings account as the *numéraire asset*. The class of *tame strategies* relative to B is defined by postulating that the discounted wealth of a strategy follows a stochastic process bounded from below. The market model obtained in this way is referred to as the *multidimensional Black and Scholes model*.

In the classic version of the multidimensional Black and Scholes model, one postulates that $d = k$, the constant volatility matrix σ is nonsingular, and the appreciation rates μ_i and the continuously compounded interest rate r are constant. It is easily

seen that under these assumptions, the martingale measure $\mathbb{P}^*$ exists and is unique.

Completeness of the Multidimensional Black and Scholes Model

The completeness of the multidimensional Black and Scholes model is defined in much the same way as for a finite market model, except that certain technical restrictions need to be imposed on the class of contingent claims we wish to hedge and price. This is linked to the fact that not all self-financing trading strategies are deemed to be *admissible*. Some of them should be excluded in order to ensure the no-arbitrage property of the model (in addition to the existence of a martingale measure). Typically, one considers the class of tame strategies to play the role of admissible trading strategies.

The multidimensional Black and Scholes model is said to be *complete* if any $\mathbb{P}^*$ -integrable, bounded from below contingent claim X is attainable, that is, if for any such claim X there exists an admissible trading strategy ϕ such that $X = V_T(\phi)$. Otherwise, the market model is said to be *incomplete*.

Since, by assumption, the interest rate process r is nonnegative and bounded, the integrability and boundedness of X is therefore equivalent to the integrability and boundedness of the discounted claim X/B_T . It is not postulated that the uniqueness of an EMM holds, and thus the $\mathbb{P}^*$ -integrability of X refers to any EMM for the model. The next result establishes necessary and sufficient conditions for the completeness of the Black and Scholes market.

Proposition 3 *The following are equivalent:*

1. *the multidimensional Black and Scholes model is complete;*
2. *inequality $d \leq k$ holds and the volatility matrix σ has full rank for Lebesgue a.e. $t \in [0, T]$, with probability 1;*
3. *there exists a unique equivalent martingale measure $\mathbb{P}^*$ for discounted stock price S^{*i} for every $i = 1, \dots, k$.*

The classic one-dimensional Black and Scholes market model introduced in [3] is clearly a special case of the multidimensional Black and Scholes model. Hence, the above results apply also to the classic Black and Scholes market model, in which the

martingale measure $\mathbb{P}^*$ is well known to be unique. We conclude that the one-dimensional Black and Scholes market model is complete, that is, any $\mathbb{P}^*$ -integrable contingent claim is $\mathbb{P}^*$ -attainable and thus it can be priced by arbitrage.

In the general semimartingale framework, the equivalence of the uniqueness of an EMM and the completeness of a market model were conjectured by Harrison and Pliska [13, 14] (see also [18]). The case of the Brownian filtration is examined in [16]. Chatelain and Stricker [7, 8] provide definitive results for the case of continuous local martingales (see also [1, 20] for related results). They focus on the important distinction between the vector and componentwise stochastic integrals.

Local and Stochastic Volatility Models

Note that we have examined the completeness of the market model in which trading was restricted to a predetermined family of primary securities. In practice, several derivative securities are also traded either on organized exchanges or over-the-counter and thus they can be used to formally complete a given market model. Let us comment briefly on two classes of models in which, for simplicity, we assume that the bond price is deterministic.

Following Dupire [12], we define the stock price as a solution to the following stochastic differential equation:

$$dS_t = S_t(\mu(S_t, t) dt + \sigma(S_t, t) dW_t^*) \quad (19)$$

where $S_0 > 0$ and the function $\sigma : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}$ represents the so-called local volatility. In practice, the function σ is obtained by fitting the model to market quotes of traded options. Model of this form is complete and thus any derivative security with the stock price as an underlying asset can be hedged and priced by arbitrage (provided, of course, that the model is arbitrage free). Another example of a complete model in which the volatility follows a stochastic process is discussed by Hobson and Rogers [15].

In a typical stochastic volatility model, the stock price S is governed by the equation

$$dS_t = \mu(S_t, t) dt + \sigma_t S_t dW_t \quad (20)$$

where the *stochastic volatility* process σ satisfies

$$d\sigma_t = a(\sigma_t, t) dt + b(\sigma_t, t) d\widehat{W}_t \quad (21)$$

where W and $\widehat{W}$ are (possibly correlated) one-dimensional Brownian motions defined on some filtered probability space $(\Omega, \mathbb{F}, \mathbb{P})$. Owing to the presence of the Brownian motion $\widehat{W}$, stochastic volatility models are incomplete if stock and bond are the only trade primary assets. By postulating that some plain-vanilla options are traded, it is possible to complete a stochastic volatility model, however. Completeness of a model of financial market with traded call and put options and related topics, such as *static hedging* of exotic options, was examined by several authors: Bajeux-Besnainou and Rochet [2], Breeden and Litzenberger [4], Brown *et al.* [5], Carr *et al.* [6], Derman *et al.* [11], Madan and Milne [17], Nachman [19], Romano and Touzi [21], and Ross [22], to mention a few.

References

- [1] Artzner, P. & Heath, D. (1995). Approximate completeness with multiple martingale measures, *Mathematical Finance* **5**, 1–11.
- [2] Bajeux-Besnainou, I. & Rochet, J.-C. (1996). Dynamic spanning: are options an appropriate instrument, *Mathematical Finance* **6**, 1–16.
- [3] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economics* **81**, 637–654.
- [4] Breeden, D. & Litzenberger, R. (1978). Prices of state-contingent claims implicit in option prices, *Journal of Business* **51**, 621–651.
- [5] Brown, H., Hobson, D. & Rogers, L. (2001). Robust hedging of options, *Applied Mathematical Finance* **5**, 17–43.
- [6] Carr, P., Ellis, K. & Gupta, V. (1998). Static hedging of exotic options, *Journal of Finance* **53**, 1165–1190.
- [7] Chatelain, M. & Stricker, C. (1994). On componentwise and vector stochastic integration, *Mathematical Finance* **4**, 57–65.
- [8] Chatelain, M. & Stricker, C. (1995). Componentwise and vector stochastic integration with respect to certain multi-dimensional continuous local martingales, in *Seminar on Stochastic Analysis, Random Fields and Applications*, E. Bolthausen, M. Dozzi, F. Russo, eds, Birkhäuser, Boston, Basel, Berlin, pp. 319–325.
- [9] Cox, J.C., Ross, S.A. & Rubinstein, M. (1979). Option pricing: a simplified approach, *Journal of Financial Economics* **7**, 229–263.
- [10] Dalang, R.C., Morton, A. & Willinger, W. (1990). Equivalent martingale measures and no-arbitrage in

- stochastic securities market model, *Stochastics and Stochastic Reports* **29**, 185–201.
- [11] Derman, E., Ergener, D. & Kani, I. (1995). Static options replication, *Journal of Derivatives* **2**(4), 78–95.
 - [12] Dupire, B. (1994). Pricing with a smile, *Risk* **7**(1), 18–20.
 - [13] Harrison, J.M. & Pliska, S.R. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and their Applications* **11**, 215–260.
 - [14] Harrison, J.M. & Pliska, S.R. (1983). A stochastic calculus model of continuous trading: complete markets, *Stochastic Processes and their Applications* **15**, 313–316.
 - [15] Hobson, D.G. & Rogers, L.C.G. (1998). Complete model with stochastic volatility, *Mathematical Finance* **8**, 27–48.
 - [16] Jarrow, R.A. & Madan, D. (1991). A characterization of complete markets on a Brownian filtration, *Mathematical Finance* **1**, 31–43.
 - [17] Madan, D.B. & Milne, F. (1993). Contingent claims valued and hedged by pricing and investing in a basis, *Mathematical Finance* **4**, 223–245.
 - [18] Müller, S. (1989). On complete securities markets and the martingale property of securities prices, *Economics Letters* **31**, 37–41.
 - [19] Nachman, D. (1989). Spanning and completeness with options, *Review of Financial Studies* **1**, 311–328.
 - [20] Pratelli, M. (1996). Quelques résultats du calcul stochastique et leur application aux marchés financiers, *Astérisque* **236**, 277–290.
 - [21] Romano, M. & Touzi, N. (1997). Contingent claims and market completeness in a stochastic volatility model, *Mathematical Finance* **7**, 399–412.
 - [22] Ross, S.A. (1976). Options and efficiency, *Quarterly Journal of Economics* **90**, 75–89.
 - [23] Schachermayer, W. (1992). A Hilbert space proof of the fundamental theorem of asset pricing in finite discrete time, *Insurance: Mathematics and Economics* **11**, 249–257.
 - [24] Taqqu, M.S. & Willinger, W. (1987). The analysis of finite security markets using martingales, *Advances in Applied Probability* **19**, 1–25.

Related Articles

Binomial Tree; Local Volatility Model; Martingale Representation Theorem; Second Fundamental Theorem of Asset Pricing.

MAREK RUTKOWSKI

Equivalent Martingale Measures

The usual setting of mathematical finance is provided by a d -dimensional stochastic process $S = (S_t)_{0 \leq t \leq T}$ based on and adapted to a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$. This process S models the price evolution of d risky stocks, which is random.

To alleviate notation, we assume from the very beginning that these prices are denoted in discounted terms: fix a traded asset, the “bond”, as numéraire and express stock prices S in units of this bond. This simple and classical technique allows us to dispense with discount factors in the formulae below (compare Section 2.1 in [6] for more details).

A central topic in mathematical finance is to decide whether there is a probability measure Q , equivalent to $\mathbb{P}$, such that S is a martingale under Q . This is the theme of the fundamental theorem of asset pricing (*see Fundamental Theorem of Asset Pricing*). Once we know that there exist equivalent martingale measures, they can be used to determine risk-neutral prices of derivative securities by taking expectations under these measures (*see Risk-neutral Pricing*), and to replicate, respectively, sub- or superreplicate, the derivative.

In fact, we were less precise in the previous paragraph (as is usual in this context) by requiring that S is a *martingale*. It turns out that some technical care is needed here, involving the notions of *local martingales* and, more generally, of *sigma-martingales*. This article deals precisely with these technical variants of the concept of a martingale.

We start by giving precise definitions.

Definition 1 An $\mathbb{R}^d$ -valued stochastic process $(S_t)_{0 \leq t \leq T}$ based on and adapted to $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$ is called a

(i) *martingale* if

$$\mathbb{E}[S_t | \mathcal{F}_u] = S_u, \quad 0 \leq u \leq t \leq T \quad (1)$$

(ii) *local martingale* if there exists a sequence $(\tau_n)_{n=1}^\infty$ of $[0, T] \cup \{+\infty\}$ -valued stopping times, increasing a.s. to ∞ , such that the stopped processes $S_{t \wedge \tau_n}$ are all martingales, where

$$S_{t \wedge \tau_n} = S_t, \quad 0 \leq t \leq T \quad (2)$$

(iii) *sigma-martingale* if there is an $\mathbb{R}^d$ -valued martingale $M = (M_t)_{0 \leq t \leq T}$ and a predictable M -integrable $\mathbb{R}_+$ -valued process φ such that $S = \varphi \cdot M$.

The process $\varphi \cdot M$ is defined as the stochastic integral in the sense of semimartingales. The—by now well understood—underlying theory was developed notably by the school of P.A. Meyer in Strasbourg [10–12]:

$$(\varphi \cdot M)_t = \int_0^t \varphi_u dM_u, \quad 0 \leq t \leq T \quad (3)$$

It is not obvious, but true, that a local martingale is a sigma-martingale, so that (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) holds true above, while the reverse implications fail to hold true as we discuss later.

Why is it necessary to introduce these generalizations of the concept of a martingale? Let us start with a familiar example of a martingale, namely, geometric Brownian motion

$$M_t = \exp \left[W_t - \frac{t}{2} \right], \quad t \geq 0 \quad (4)$$

where the process $(W_t)_{t \geq 0}$ is a standard Brownian motion.

Clearly, $(M_t)_{t \geq 0}$ is a martingale (with reference to its natural filtration) when t ranges in $[0, \infty[$. But what happens if we include $t = \infty$ into the time set? It is straightforward to verify that

$$M_\infty := \lim_{t \rightarrow \infty} M_t \quad (5)$$

exists a.s. and equals

$$M_\infty = 0 \quad (6)$$

Hence we may well define the continuous process $(M_t)_{0 \leq t \leq \infty}$; this process is not a martingale any more as

$$1 = M_0 > \mathbb{E}[M_\infty] = 0 \quad (7)$$

In this example, the breakdown of the martingale property happens at $t = \infty$. However, it is purely formal to shift this problem to any other point $T \in]0, \infty[$, for example, $T = 1$. Indeed, letting

$$\begin{aligned} \tilde{M}_t &= M_{\tan(\frac{t\pi}{2})}, \quad 0 \leq t < 1 \\ \tilde{M}_1 &= M_\infty = 0 \end{aligned} \quad (8)$$

we find a process $(\tilde{M}_t)_{0 \leq t \leq 1}$, having a.s. continuous paths, which fails to be a martingale. However, it is intuitively clear that “locally”, that is, before t assumes the value 1, the process $(\tilde{M}_t)_{0 \leq t \leq 1}$ is “something like a martingale”. The good way to formalize this intuition is to find a “localizing” sequence of stopping times as in (ii) above. The canonical choice^a is

$$\tau_n = \inf\{t \in [0, 1] : |\tilde{M}_t| \geq n\} \quad (9)$$

which is a $[0, 1] \cup \{\infty\}$ -valued stopping time, if we define the infimum over the empty set to be equal to ∞ . It is straightforward to verify that $(\tau_n)_{n=1}^\infty$ satisfies the requirements of (ii) above.

In the above example it holds true that $(\tilde{M}_t)_{0 \leq t < 1}$ is a martingale, that is, when t ranges in $[0, 1[$. In other words, the problem only arises at $t = 1$. However, there is a more subtle phenomenon in this context, where the problem not only appears at one single value of t , but also for *all* t .

The canonical example for this phenomenon is the inverse of the three-dimensional Bessel process^b $(R_t)_{0 \leq t \leq 1}$. It may be defined by $R_0 = 1$ and

$$dR_t = R_t^2 dW_t, \quad 0 \leq t \leq 1 \quad (10)$$

It turns out that equation (10) well defines a stochastic process with continuous paths, which is a local martingale (define $(\tau_n)_{n=1}^\infty$ as in equation (9)). However, the function

$$t \rightarrow \mathbb{E}[R_t], \quad 0 \leq t \leq 1 \quad (11)$$

is *strictly decreasing* on the entire interval $[0, 1]$. Intuitively speaking, this may be interpreted as $(R_t)_{0 \leq t \leq 1}$ “losing mass in continuous time”. We leave it to the reader to develop his or her own intuition for this remarkable phenomenon. In any case, this example should convince the reader that the concept of *local martingales*, involving “localizing” sequences of stopping times, is a useful and natural notion.

To underline this claim even further, think of a diffusion process $(X_t)_{0 \leq t \leq 1}$ satisfying the equation

$$dX_t = \sigma(X_t, t) dW_t + \mu(X_t, t) dt, \quad 0 \leq t \leq 1 \quad (12)$$

A typical argument used, for example, in the derivation of the Black–Scholes partial differential equation (PDE) (compare **Complete Markets**

or [Vol. II] [13]), goes as follows: for the drift term $\mu(X_t, t)$ we have $\mu(X_t, t) \equiv 0$ if and only if $(X_t)_{0 \leq t \leq 1}$ is a martingale. This is a very useful argument. However, this argument is not quite complete, as a glance at equation (10) reveals where we only obtain a local martingale. The correct statement is the process X is a *local martingale* if and only if $\mu(X_t, t) = 0$, a.s. with respect to $d\mathbb{P} \otimes d\lambda$.

Local Martingales in Finance

As a concrete example of a local martingale modeling a price process we consider $\tilde{S} = (\tilde{S}_t)_{0 \leq t \leq T} = (\tilde{M}_t)_{0 \leq t \leq 1}$, the time-changed geometric Brownian motion defined in equation (8). We consider $\tilde{S}$ to be defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$, where the filtration $(\mathcal{F}_t)_{0 \leq t \leq T}$ is generated by $\tilde{S}$ and $\mathcal{F} = \mathcal{F}_T$. We conclude that $Q = \mathbb{P}$ is the unique probability measure Q on $\mathcal{F}$, which is equivalent to $\mathbb{P}$ and such that $\tilde{S}$ is a local martingale under Q . This quickly follows from the fact that $Q = \mathbb{P}$ is the unique probability measure on $\mathcal{F}$, equivalent to $\mathbb{P}$, such that the Brownian motion $W = (W_t)_{0 \leq t < \infty}$ in equation (4) is a martingale under Q .

The question arises whether $\tilde{S}$ defines a sound, that is, arbitrage-free model of a financial market. At first glance, things seem suspicious; after all, $(\tilde{S}_t)_{0 \leq t \leq T}$ is a ridiculous stock; it starts at $\tilde{S}_0 = 1$ and ends a.s. at $\tilde{S}_T = 0$. Hence buying a stock $\tilde{S}$ and holding it up to time $T = 1$ is a very “silly investment”. M. Harrison and St. Pliska [8] called such an investment a *suicide strategy*. But it is not forbidden to be silly.

A much more appealing investment strategy would be to go short in the stock at time 0, and to hold this short position up to time T , as this strategy yields a.s. a gain of one Euro. However, unfortunately, this is forbidden. To understand why this is the case, let us recall from **Fundamental Theorem of Asset Pricing** the definition of *admissibility*: a trading strategy $H = (H_t)_{0 \leq t \leq T}$ for a (general semimartingale) stock price process S is defined as a predictable strategy, which is S -integrable. By definition, the stochastic integral

$$(H \cdot S)_t = \int_0^t (H_u, dS_u), \quad 0 \leq t \leq T \quad (13)$$

then is well defined. We call H *admissible*, if there is a constant $C > 0$ such that a.s.

$$(H \cdot S)_t \geq -C, \quad 0 \leq t \leq T \quad (14)$$

The finite credit line C rules out doubling strategies and similar schemes that capitalize on taking higher and higher risks. A typical representative of such a “kind of doubling strategy” is the strategy of going short in the stock $(\tilde{S}_t)_{0 \leq t \leq T}$, which corresponds to taking $H_t \equiv -1$, for $0 \leq t \leq T$.

We now shall convince ourselves that local martingales yield sound, arbitrage-free models of financial markets. It turns out that it does not matter whether we start with a true martingale $S = (S_t)_{0 \leq t \leq T}$ or with a local martingale S , if we are only interested in the *admissible stochastic integrals* $H \cdot S$. Indeed, it was shown by J.P. Ansel and C. Stricker [1, Corr. 3.5] that, given a local martingale S and an admissible integrand H , the stochastic integral $H \cdot S$ is a local martingale and therefore (using once more the fact that $H \cdot S$ is bounded from below) a supermartingale. In particular, the process $H \cdot S$ cannot increase in expectation; but it may very well decrease in expectation as, for example, the process $\tilde{S}$ above.

The following characterization of local martingales (see [4, Prop. 2.5] for more on this issue) is useful in this context.

Proposition 1 *For an $\mathbb{R}^d$ -valued semimartingale S the following are equivalent.*

- (i) S is a local martingale.
- (ii) $S = \varphi \cdot M$, where M is an $\mathbb{R}^d$ -valued martingale, and φ is an $\mathbb{R}_+$ -valued, M -integrable, predictable, increasing process.

From this proposition and the trivial formula^c

$$H \cdot M = \frac{H}{\varphi} \cdot (\varphi \cdot M) = \frac{H}{\varphi} \cdot S \quad (15)$$

which holds true for every $\mathbb{R}^d$ -valued, predictable, M -integrable process $H = (H_t)_{0 \leq t \leq T}$ we deduce that the family of processes, which are stochastic integrals on the local martingale S coincides with the family of processes, which are stochastic processes on the martingale M . Also note that H is admissible for M if and only if $\frac{H}{\varphi}$ is admissible for S .

The bottom line of formula (15) is that *there is no difference between the stock price process S and M in Proposition 1* if we are only interested in the *admissible stochastic integrals* on these processes: these two families of stochastic integrals coincide.

Sigma-martingales

For continuous stock price processes S or, more generally, for locally bounded processes S , the concept of *local martingales* is sufficiently general to characterize those models that satisfy the condition of *no free lunch with vanishing risk* (see **Fundamental Theorem of Asset Pricing**).

However, if we pass to processes S that are not locally bounded any more, we still need one more step of generalization. The key concept for doing so was introduced by C. Chou [2] and M. Émery [7] under the name of “semimartingale de la classe (Σ_m) ”. In [4], F. Delbaen and this author took the liberty of calling these processes *sigma-martingales* S . The reason for this is that their relation to martingales is analogous to the relation between sigma-finite measures and finite measures, as seen from Definition 1 above. Also note the (only) difference between Definition 1 (iii) of a sigma-martingale, and the characterization of a local martingale as given in Proposition 1: in the latter the predictable, $\mathbb{R}_+$ -valued process φ is supposed to be *increasing* while there is no such restriction in the former one.

Here is the illuminating example, due to M. Émery [7] (compare [4, Ex. 2.2]), of the archetypical sigma-martingale, which fails to be a local martingale.

Example 1 We start with an exponentially distributed random variable τ and an independent Bernoulli random variable ε , that is, $\mathbb{P}[\varepsilon = 1] = \mathbb{P}[\varepsilon = -1] = \frac{1}{2}$. These random variables are based on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Define the process $M = (M_t)_{t \geq 0}$ by

$$M_t = \begin{cases} 0, & \text{for } 0 \leq t \leq \tau \\ \varepsilon, & \text{for } \tau \leq t \end{cases} \quad (16)$$

The verbal description goes like this: the process M remains at zero until time τ ; then a coin is flipped, independently of τ , and the process M continues at the level $+1$ or -1 , according to the result of this coin flip.

Denoting by $(\mathcal{F}_t)_{t \geq 0}$ the filtration generated by $(M_t)_{t \geq 0}$, it is rather obvious that $(M_t)_{t \geq 0}$ is a *martingale* in this filtration $(\mathcal{F}_t)_{t \geq 0}$. To keep in line with the above notation, we only consider the finite-horizon process $(M_t)_{0 \leq t \leq T}$, but the example could as well be presented for the infinite horizon.

Let $\varphi = (\varphi_t)_{0 \leq t \leq 1}$ be the *deterministic process*

$$\varphi_t = t^{-1}, \quad 0 \leq t \leq 1 \quad (17)$$

and define the stochastic integral $S = \varphi \cdot M$, for which we get

$$S_t = \begin{cases} 0, & \text{for } 0 \leq t \leq \tau \\ \tau^{-1}\varepsilon, & \text{for } \tau \leq t \end{cases} \quad (18)$$

The process $S = (S_t)_{0 \leq t \leq 1}$ is a well-defined stochastic integral (in the pointwise Stieltjes sense). The verbal description of S goes as follows: again S remains at 0 until time τ and then, depending on the sign of ε , it jumps to $+\tau^{-1}$ or $-\tau^{-1}$.

Is the process S a martingale? Morally speaking, one might think yes, as it has the same odds of jumping up or down.^d But this intuition goes wrong: indeed, the notion of martingale is based on some (conditional) expectations to be zero. When we do the calculations in the present example we end up with expressions of the form $\infty - \infty$, which creates a problem. Indeed, we have

$$\mathbb{E}[|S_t|] = \infty, \quad \text{for } 0 \leq t \leq 1 \quad (19)$$

as is easily seen from $\int_0^t u^{-1} du = \infty$, for $t > 0$. In fact, it is not hard to show [7] that, for every $(\mathcal{F}_t)_{0 \leq t \leq 1}$ —stopping time $\sigma : \Omega \rightarrow [0, 1]$, such that $\mathbb{P}[\sigma > 0] > 0$, we have

$$\mathbb{E}[|S_\sigma|] = \infty \quad (20)$$

It follows that S even *fails to be a local martingale*. But, of course, S is a sigma-martingale by its very construction.

The message of the above example is that the notion of sigma-martingale is tailor-made to save the intuition that—from a “moral point of view”—the above process S is “something like a martingale”.

Let us turn from moral considerations to finance again: the question arises as to whether a process $S = (S_t)_{0 \leq t \leq T}$, which is a sigma-martingale under some measure Q equivalent to $\mathbb{P}$, well defines a sound, that is, arbitrage-free, model of a financial market. The answer is analogous to the case of a local martingale, namely, a resounding *yes*. If $S = \varphi \cdot M$, for some $\mathbb{R}_+$ -valued predictable process φ , then again the “trivial formula” (15) above holds true. Hence again the families of *admissible stochastic integrals* on the processes S and M coincide. If these are the only relevant objects—as is the case for the classical approach to no-arbitrage theory as proposed by M. Harrison and S. Pliska [8]—the

processes S and M work equally well. In particular, the Ansel–Stricker Theorem carries over to sigma-martingales (see [4, Th. 5.5] for a somewhat stronger version of this result).

It is not hard to show that a locally bounded process, which is a sigma-martingale, is already a local martingale [4, Prop. 2.5 and 2.6]. Émery’s example shows that this is not the case any more if we drop the local boundedness assumption. From a financial point of view, however, the question of interest arises in a slightly different version. Is there an example of a process $S = (S_t)_{0 \leq t \leq T}$, which is a sigma-martingale, say under $\mathbb{P}$, but such that it *fails to be a local martingale under any probability measure Q equivalent to $\mathbb{P}$* ?

Émery’s original example does not provide a counterexample to this question; in this example, it is not hard to pass from $\mathbb{P}$ to Q such that S even becomes a Q -martingale. However, in [4, Ex. 2.3] a variant of Émery’s example has been constructed, which is a process S taking values in $\mathbb{R}^2$ answering the above question negatively. It seems worth mentioning that—to the best of the author’s knowledge—it is unknown whether there also is a counterexample of a process S , taking values only in $\mathbb{R}$, to this question.

Separating Measures

We have seen in the preceding sections that, for a process $S = (S_t)_{0 \leq t \leq T}$ which is a sigma-martingale under some probability measure Q and for each admissible integrand H , we have the inequality

$$\mathbb{E}_Q[(H \cdot S)_T] \leq 0 \quad (21)$$

Indeed, the theorem of Ansel–Stricker [1, Corr. 3.5] and its extension to sigma-martingales [4, Th. 5.5] imply that $H \cdot S$ is a local martingale and, using again the boundedness from below, the process $H \cdot S$ is a supermartingale.

The notion of a *separating measure* introduced by Y. Kabanov in [9], takes this inequality (21) as defining property. To formalize this idea, we assume that S is an $\mathbb{R}^d$ -valued semimartingale on some filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$. We say that a measure Q , equivalent to $\mathbb{P}$, is a *separating measure* for S if, for all admissible, predictable S -integrable integrands H , inequality (21) holds true.

If S is bounded, then it is straightforward to verify that the validity of inequality (21), for all admissible H , is tantamount to S being a martingale. It follows that, if S is locally bounded, then the validity of inequality (21), for all admissible H , is tantamount to S being a local martingale. Hence, we do not find anything new by using the notion of *separating measure* in the context of locally bounded semimartingales S . However, for semimartingales S that are not locally bounded, we do find something new; as observed above, if S is a sigma-martingale under Q then inequality (21) holds true, for all admissible H . But the converse does not hold true. The difference is illustrated by the subsequent easy one-period example. To stay in line with the present notation, we write it as an example in continuous time.

Example 2 Let X be an $\mathbb{R}$ -valued random variable, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which is unbounded from above and from below. For example, we may choose X to be normally distributed.

The process $S = (S_t)_{0 \leq t \leq 1}$ is defined as

$$S_t = \begin{cases} 0 & 0 \leq t < 1 \\ X & t = 1 \end{cases} \quad (22)$$

Defining $(\mathcal{F}_t)_{0 \leq t \leq 1}$ as the filtration generated by $S = (S_t)_{0 \leq t \leq 1}$, we find that the only $(\mathcal{F}_t)_{0 \leq t \leq 1}$ -predictable processes are the *constant processes* $H = (H_t)_{0 \leq t \leq 1}$. Among those, the only S -admissible predictable process is $H = 0$. Indeed, if $H = \text{const} \neq 0$, the process $H \cdot S$ is not bounded from below in the sense of inequality (14).

The condition (21) therefore is trivially satisfied, for *each* probability measure Q equivalent to $\mathbb{P}$. On the other hand, S is a martingale (or, equivalently, a local or a sigma-martingale) under Q if

$$\mathbb{E}_Q[X] = \mathbb{E}_Q[S_1] = 0 \quad (23)$$

Hence we see that, in this easy example, the class of separating measures Q is strictly bigger than the class of probability measures Q , equivalent to $\mathbb{P}$, under which S is a sigma-martingale.

Where does the nomenclature “separating measure” come from? This concept arises naturally as an intermediary step in the proof of **Fundamental Theorem of Asset Pricing** (compare [9] for a careful analysis of the arguments in [3] and [4] and, in particular, for the introduction of the name “separating measure”).

In the context of this theorem, after surmounting some difficulties, an application of the Hahn–Banach theorem plus an exhaustion argument due to J. Yan ([15], compare also [14]) provides a σ^* -continuous, linear functional $F : L^\infty(\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathbb{R}$ which *strictly separates* the set of random variables of the form^e

$$(H \cdot S)_T = \int_0^T (H_t, dS_t) \quad (24)$$

where H runs through the admissible integrands, from $L_+^\infty(\Omega, \mathcal{F}, \mathbb{P}) \setminus \{0\}$, that is, the positive orthant with the origin 0 deleted. Normalizing the functional F by $F(\zeta) = 1$, this translates into the fact that F is of the form

$$F(g) = \mathbb{E}_Q[g], \quad g \in L^\infty(\Omega, \mathcal{F}, \mathbb{P}) \quad (25)$$

where Q is a *separating measure*.

If the process S is bounded (respectively, locally bounded), it immediately follows that S is a martingale (respectively, a local martingale) under this separating measure Q , which concludes the proof of the fundamental theorem of asset pricing (see [3]).

If, however, S fails to be locally bounded, then we cannot conclude that S is “some kind of martingale” under the separating measure Q , as is illustrated by Example 2 above. Some further work is needed—which was carried out in [4]—to pass from the separating measure Q to a probability measure $\tilde{Q}$, which is equivalent to $\mathbb{P}$ and under which S is a sigma-martingale measure. It turns out that, in the setting of the fundamental theorem of asset pricing [4], the latter set is dense with reference to $\|\cdot\|_1$ in the set of separating measures for S . In particular, this set is nonempty, provided we have found a separating measure. This argument concludes the proof of the fundamental theorem of asset pricing also in the case of a general, $\mathbb{R}^d$ -valued semimartingale S .

End Notes

^aFor *continuous* local martingales $(M_t)_{t \geq 0}$ starting at $M_0 = 0$ the choice of stopping times *via* equation (9) always works, that is, gives a sequence $(\tau_n)_{n=1}^\infty$ satisfying the requirements of (ii) above. In the case of càdlàg local martingales this is not true any more and one may give examples of local martingales where equation (9) does not define a sequence of localizing stopping times.

^bThe name is derived from the following fact: let $(B_t)_{0 \leq t \leq 1} = (B_t^1, B_t^2, B_t^3)_{0 \leq t \leq 1}$ be an $\mathbb{R}^3$ -valued standard Brownian motion starting at $B_0 = (B_0^1, B_0^2, B_0^3) = (1, 0, 0)$. Let $R_t = \|B_t\|^{-1}$ where $\|\cdot\|$ denotes Euclidean norm on $\mathbb{R}^3$. Then $(R_t)_{0 \leq t \leq 1}$ satisfies equation (10), where $(W_t)_{0 \leq t \leq 1}$ is a one-dimensional Brownian motion adapted to the filtration generated by the three-dimensional Brownian $(B_t)_{0 \leq t \leq 1}$. We refer to [11] for a beautiful presentation of the theory of Bessel processes (compare also [5]).

^cIt is easy to verify that in Proposition 1 (as well as in Definition 1 (iii)), we may assume without loss of generality that φ takes its values in $]0, \infty[$ (or, equivalently, in $[1, \infty[$). See [4, Prop. 2.5] for details.

^dA precise statement is that the processes S and $-S$ have the same law, which obviously is the case.

^eTo be precise, we have to consider the random variables $(H \cdot S)_T \wedge C$, where C runs through $\mathbb{R}_+$, to make sure that these random variables are in $L^\infty(\Omega, \mathcal{F}, \mathbb{P})$.

References

- [1] Ansel, J.P. & Stricker, C. (1994). Couverture des actifs contingents et prix maximum, *Annales de l'Institut Henri Poincaré – Probabilités et Statistiques* **30**, 303–315.
- [2] Chou, C.S. (1977/78). Caractérisation d'une classe de semimartingales, in *Séminaire de Probabilités XIII*, Springer Lecture Notes in Mathematics, Springer, Vol. 721, pp. 250–252.
- [3] Delbaen, F. & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing, *Mathematische Annalen* **300**, 463–520.
- [4] Delbaen, F. & Schachermayer, W. (1998). The fundamental theorem of asset pricing for unbounded stochastic processes, *Mathematische Annalen* **312**, 215–250.
- [5] Delbaen, F. & Schachermayer, W. (1995). Arbitrage possibilities in Bessel processes and their relations to local martingales, *Probability Theory and Related Fields* **102**, 357–366.
- [6] Delbaen, F. & Schachermayer, W. (2006). *The Mathematics of Arbitrage*, Springer Finance, p. 371. ISBN: 3-540-21992-7.
- [7] Émery, M. (1980). Compensation de processus à variation finie non localement intégrables, in *Séminaire de Probabilités XIV*, J. Azema & M. Yor, eds, Springer Lecture Notes in Mathematics, Springer, Vol. 784, pp. 152–160.
- [8] Harrison, J.M. & Pliska, S.R. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and their Applications* **11**, 215–260.
- [9] Kabanov, Y.M. (1997). On the FTAP of Kreps-Delbaen-Schachermayer (English), in *Statistics and Control of Stochastic Processes*, Y.M. Kabanov, B.L. Rozovskii & A.N. Shiryaev, eds, The Liptser Festschrift. Papers from the Steklov Seminar held in Moscow, Russia, 1995–1996, World Scientific, Singapore, pp. 191–203.
- [10] Protter, P. (1990). Stochastic integration and differential equations. A new approach, in *Applications of Mathematics*, (2nd edition, 2003, corrected third printing: 2005) Springer-Verlag, Berlin, Heidelberg, New York, Vol. 21.
- [11] Revuz, D. & Yor, M. (1991). Continuous martingales and Brownian motion, in *Grundlehren der Mathematischen Wissenschaften*, 3rd edition, 1999, corrected third printing: 2005, Springer, Vol. 293.
- [12] Rogers, L.C.G. & Williams, D. (2000). *Diffusions, Markov Processes and Martingales*, Cambridge University Press, Vol. I and II.
- [13] Shreve, S. (2004). Stochastic calculus for finance, *Springer Finance* **I, II**, 208, 550.
- [14] Stricker, Ch. (1990). Arbitrage et Lois de martingale, *Annales de l'Institut Henri Poincaré – Probabilités et Statistiques* **26**, 451–460.
- [15] Yan, J.A. (1980). Caractérisation d'une classe d'ensembles convexes de L^1 ou H^1 , in *Séminaire de Probabilités XIV*, J. Azema & M. Yor, eds, Springer Lecture Notes in Mathematics, Springer, Vol. 784, pp. 220–222.

Related Articles

Arbitrage Strategy; Complete Markets; Free Lunch; Fundamental Theorem of Asset Pricing; Martingales; Minimal Entropy Martingale Measure; Minimal Martingale Measure; Risk-neutral Pricing; Second Fundamental Theorem of Asset Pricing.

WALTER SCHACHERMAYER

Second Fundamental Theorem of Asset Pricing

The *second fundamental theorem of asset pricing* concerns the mathematical characterization of the economic concept of *market completeness* for liquid and frictionless markets with an arbitrary number of assets. The theorem establishes the mathematical necessary and sufficient conditions in order to guarantee that every contingent claim on the market can be duplicated with a portfolio of primitive assets. For finite asset economies, completeness (i.e., *perfect replication* of every claim on the market by admissible self-financing strategies) is equivalent to uniqueness of the equivalent martingale measure. This result can be extended to market models with an infinite number of assets by defining completeness in terms of *approximate replication* of claims by attainable ones. Hence several definitions of completeness are possible, and in the sequel we will present and discuss them extensively.

Finite Number of Assets

The *second fundamental theorem* appeared in [9] under the assumption that the interest rate is zero and that the agent employs only simple trading strategies to address the following issue, raised in the economic literature [1, 20, 22]: given a financial market, which contingent claims are “spanned” by a given set of market securities?

In the seminal paper [7], it was already observed that in the idealized Black–Scholes market the cash flow of an option can be duplicated by managing a portfolio containing only stock and bond. A natural question is then as follows: for which contingent claim does this result hold in more general markets? When does it hold for *all* contingent claims on the market?

For markets with a finite number of asset prices, the answer to this problem was provided for the first time in [10, 11]. Here we follow the notation of [11] in order to state the *second fundamental theorem*.

Let $T < \infty$ be a fixed time horizon; consider a probability space $(\Omega, \mathcal{F}, P)$ endowed with a filtration $(\mathcal{F}_t)_{t \in [0, T]}$ satisfying the usual conditions and such that $\mathcal{F}_0$ contains only Ω and the null sets of P

and with $\mathcal{F}_T = \mathcal{F}$. Let $S = (S_t^0, \dots, S_t^d)_{t \in [0, T]}$ be a $(d + 1)$ -dimensional strictly positive semimartingale, whose components $S^0, \dots, S^d$ are right continuous with left limits. Moreover, we assume that $S_0^0 = 1$. Here, the stochastic process S_t^k represents the value at time t of the k th security on the market. The discounted price process $Z = (Z_t^1, \dots, Z_t^d)_{t \in [0, T]}$ is then defined by setting $Z^k = S^k/S^0$, for $k = 1, \dots, d$. Let $\mathbb{P}$ be the set of probability measures Q on $(\Omega, \mathcal{F})$ that are equivalent to P and such that Z is a (vector) martingale under Q . We assume that $\mathbb{P}$ is not empty, that is, that the market is arbitrage free (see **Fundamental Theorem of Asset Pricing**). We fix an element P^* in $\mathbb{P}$ and denote by E^* the expectation under P^* . Let $L(Z)$ denote the set of all vector-valued, predictable processes $H = (H_t^1, \dots, H_t^d)_{t \in [0, T]}$ that are *integrable* with respect to the semimartingale Z . For further details on $L(Z)$, we refer to Remark 1 below.

Definition 1 A stochastic process $\phi \in L(Z)$ is said to be an *admissible self-financing strategy* if

- (i) the discounted value process $V^*(\phi) := \sum_{k=1}^d \phi^k Z^k$ is almost surely nonnegative;
- (ii) $V^*(\phi)$ satisfies the self-financing condition

$$V_t^*(\phi) = V_0^*(H) + \int_0^t \sum_{k=1}^d \phi_s^k dZ_s^k, \quad t \in [0, T]; \quad (1)$$

- (iii) $V^*(\phi)$ is a martingale under P^* .

Condition (iii) is introduced here to rule out “certain foolish strategies that throw out money” [11], that is, for no-arbitrage reasons. Note also that in the preceding definition only the last condition may depend on the choice of the reference measure P^* .

A *contingent claim* X with maturity T is then represented by a nonnegative $(\mathcal{F}_T$ -measurable) random variable. Such a claim is said to be *attainable* if there exists an admissible trading strategy ϕ such that $V_T^*(\phi) = X/S_T^0$. The model is said to be *complete* if every claim^a is attainable.

Theorem 1 (The second fundamental theorem of asset pricing, [11]). *Let $\mathbb{P} \neq \emptyset$. Then the following statements are equivalent:*

- (i) The model is complete under P^* .

- (ii) Every P^* -martingale M can be represented in the form

$$M_t = M_0 + \int_0^t \sum_{k=1}^d H_s^k dZ_s^k, \quad t \in [0, T] \quad (2)$$

for some $H \in L(Z)$ (predictable representation property).

- (iii) $\mathbb{P}$ is a singleton, that is, there exists a unique equivalent martingale measure for Z .

The proof of this theorem relies on some results of [12, 14], Chapter XI, relating the representation property (1) to a condition involving a certain set of probability measures.

Remark 1 In Theorem 1 the definition of the space $L(Z)$ is crucial, as shown by a counterexample in [19]. From reference [16] we obtain that $L(Z)$ must be the largest class of integrands over which multidimensional integrals with respect to Z can be defined, as done implicitly in [11]. Hence by Theorem 4.6 of [12] we have that $L(Z)$ is the space of the vector-valued, predictable processes $H = (H_t^1, \dots, H_t^d)_{t \in [0, T]}$ such that

$$\int_0^t \sum_{i,j=1}^d H_s^i H_s^j d[Z^i, Z^j]_s, \quad t \in [0, T] \quad (3)$$

is locally integrable.

Completeness can be easily characterized in some particular cases, as shown by the following examples.

Example 1 Consider a market with a finite number of assets in discrete time $\{0, \dots, T\}$ and let $\mathcal{P}_t$ be the partition of Ω underlying $\mathcal{F}_t$. For each cell A of $\mathcal{P}_t$, $t \in \{0, \dots, T-1\}$, we define as *splitting index* of A the number $K_t(A)$ of cells of $\mathcal{P}_{t+1}$ that are contained in A . Then completeness can be characterized as follows.

Proposition 1 (Proposition 2.12 of [10]). *Let $\mathbb{P} \neq \emptyset$ and suppose that the securities are not redundant.^b Then the model is complete if and only if $K_t(A) = d+1$ for all $A \in \mathcal{P}_t$ and $t = 0, \dots, T-1$.*

Hence completeness is a matter of dimension. Corollary 4.2 of [23] shows that if the market is complete, then the splitting index $K_t(A)$ is determined by the price process S only, that is, for every

$t = 0, \dots, T$ and each $A \in \mathcal{P}_t$, we have $K_t(A) = \dim(\text{span}\{S_{t+1}(\omega) : \omega \in A\})$. Hence it is sufficient to check if the rank of the matrix with columns formed by the vectors $S_{t+1}(\omega)$, $\omega \in A$, equals the splitting index $K_t(A)$ of A . By using this geometric property of the sample paths of the price process, an algorithm is then provided in [23] to check if finite securities markets in discrete time are complete.

Example 2 In the case when security prices follow Itô processes on a multidimensional Brownian filtration, completeness of the market can be characterized in terms of the volatility matrix of the underlying asset prices, as shown in [3, 16, 18]. Consider a market with d risky assets given by Itô processes of the form

$$S_t^i = S_0^i \exp \left[\int_0^t \alpha_s^i ds - 1/2 \sum_{j=1}^n \int_0^t (\sigma_s^{ij})^2 ds + \sum_{j=1}^n \int_0^t \sigma_s^{ij} dW_s^j \right], \quad t \in [0, T] \quad (4)$$

$i = 1, \dots, d$, on the probability space $(\Omega, \mathcal{F}, P)$ endowed with the (augmented) natural filtration $(\mathcal{F}_t)_{t \in [0, T]}$ generated by the n -dimensional Brownian motion $W = (W_t^1, \dots, W_t^n)_{t \in [0, T]}$ with $\mathcal{F}_T = \mathcal{F}$. Here S^0 can be assumed constantly equal to 1 for the sake of simplicity. For $t \in [0, T]$ we denote by $\Sigma_t(\omega)$ the (random) volatility matrix, whose entries are given by

$$[\Sigma_t(\omega)]_{ij} = \sigma_t^{ij}(\omega), \quad i = 1, \dots, d, \quad j = 1, \dots, n \quad (5)$$

If for all $i = 1, \dots, d$, S_0^i is a positive constant, $(\alpha_t^i)_{t \in [0, T]}$ an adapted stochastic process with

$$\int_0^T |\alpha_s^i| ds < \infty, \quad \text{a.s.} \quad (6)$$

and $(\sigma_t^{ij})_{t \in [0, T]}$ are adapted stochastic processes with

$$\int_0^T (\sigma_s^{ij})^2 ds < \infty, \quad \text{a.s.} \quad (7)$$

for $j = 1, \dots, n$, then the following characterization of market completeness holds.

Theorem 2 (Theorem 4 of [3], Theorem 2.2 and 3.2 of [16]). *Let $\mathbb{P} \neq \emptyset$. Then the market is complete if and only if $P(\text{rank}(\Sigma_t) = d \text{ for almost all } t \in [0, T]) = 1$.*

For further references, see also Theorem 4.1 of [18]. Since there are n sources of randomness represented by the Brownian motions, it is natural to expect that n sufficiently independent asset prices are needed for completeness. Clearly, if $d < n$ the market cannot be complete.

Example 3 If price processes are discontinuous but with a finite number of jump sizes, then we obtain again a characterization of completeness in terms of the volatility matrix, as shown by the following theorem attributed to Bättig [3]. We set again $S^0 = 1$ and consider price processes driven by a multivariate point process^c μ with compensator $\nu(dt, dx) = K_t(dx)dt$ such that

$$S_t^i = S_0^i \varepsilon(R^i)_t, \quad t \in [0, T], \quad i = 1, \dots, d \quad (8)$$

with

$$R_t^i = \int_0^t \alpha_s^i ds + \int_{[0,t] \times E} \sigma^i(u, x)(\mu(du, dx) - \nu(du, dx)), \quad t \in [0, T], \quad i = 1, \dots, d, \quad (9)$$

where the $\sigma^i(t, x)$ s are bounded $d\mu \otimes dP$ a.e., $\mathcal{E}$ is the Doléans exponential (for the definition, we refer to Theorem I.4.61 of [13]), and $E \subset \mathbb{R}$. Note that here σ^i , μ , and ν may depend on ω , but for the sake of simplicity we do not indicate this dependence. In this context, asset prices may have jumps that can be thought of as the result of possible shocks that may trigger the market. If the cardinality $|E|$ of E is finite, we denote again by Σ_t the volatility matrix, whose row vectors are given by $(\sigma^i(t, x))_{x \in E}$, $i = 1, \dots, d$.

Theorem 3 (Theorem 5 of [3]). *Let $\mathbb{P} \neq \emptyset$, $|E| < \infty$ and $K_t(\{x\}) > 0$ for every $x \in E$. Then the market is complete if and only if $P(\text{rank}(\Sigma_t) = |E| \text{ for almost all } t \in [0, T]) = 1$.*

Furthermore, in the case of a finite number of jumps that may trigger the economy, the characterization of market completeness is similar to the Itô price process case, that is, one needs $|E|$ sufficiently independent processes for completeness in presence of the $|E|$

sources of randomness, given by the $|E|$ different possible shocks.

We have seen that the key to completeness is the predictable representation property. Hence, a natural question concerns the kind of martingales for which the predictable representation property is satisfied. For the continuous case, we have that the predictable representation property holds for diffusion processes that are martingales and have either Lipschitz coefficients [24] or a nondegenerate diffusion matrix and continuous coefficients [14]. The only one-dimensional martingales with stationary and independent increments that satisfy the predictable representation property are the Wiener and the Poisson martingales [25]. Hence the representation property holds for finite Lévy measures, but it fails for infinite Lévy measures. In the next section, we discuss the *second fundamental theorem* in the case of infinite dimensional financial markets.

Infinite Number of Assets

Many applications of hedging involve dynamic trading in principle in infinitely many securities, for example, in pricing of interest rate derivatives by using pure discount bonds or in the use of the term and strike structure of European put and call options to hedge exotic derivatives, when asset prices are driven by Lévy measures. Hence it is natural to develop infinite dimensional market models to address this kind of issues. The problem now is to establish if the *second fundamental theorem* still holds, and if the market is endowed with an infinite number of assets.

By defining a complete market *via* the density of a vector space, the *second fundamental theorem* is in [8] proved to hold true for (infinitely many) continuous and bounded asset price processes, if all the martingales with respect to the reference filtration $\mathcal{F}_t$ are continuous ([8], Theorem 6.7). In the case of a general filtration, Theorem 6.5 of [8] states that completeness is equivalent for P^* to be an extreme point of $\mathbb{P}$, that is, a weaker version of the *second fundamental theorem* holds.

The hypothesis of continuity cannot be dropped and in the presence of jumps (discontinuities) and infinitely many assets, a counterexample to the *second fundamental theorem* is provided in [2], where an economy with infinitely many assets is constructed,

in which the market is complete; yet, there exists an infinity of equivalent martingale measures. Since the formulation of this counterexample, many papers have studied the problem of extending the result of the *second fundamental theorem* to markets with infinitely many assets. Since many definitions of completeness are possible, the solution to the counterexample of [2] relies on the choice of the definition of completeness that is adopted. A first answer to this problem was provided in 1997 by Björk *et al.* [5, 6], where Theorem 6.11 shows that in the presence of infinitely many assets and a continuum of jump sizes, the uniqueness of the equivalent martingale measure is equivalent to the market being *approximately complete*, that is, every bounded contingent claim can be approached in $L^2(Q)$ for some $Q \in \mathbb{P}$ by a sequence of hedgeable claims.

In 1999, a number of papers appeared [3, 4, 15, 17] at the same time, where new definitions of market completeness were proposed in order to maintain the *second fundamental theorem*, even in complex economies. The equivalence between market completeness and uniqueness of the pricing measure is maintained by introducing a notion of market completeness that is independent both of the notion of no arbitrage and of a chosen equivalent martingale measure. In finite-dimensional markets, the definition of market completeness is given in terms of replicating value processes in economies without arbitrage possibilities and with respect to a given equivalent martingale measure. However, the issue of completeness is about the ability to replicate certain cash flows, and not about how these cash flows are valued or whether these values are arbitrage free. From this perspective, the appropriate measure to address the issue of completeness is the statistical probability measure P , and not an equivalent martingale measure that may also not exist. In reference [17], this new approach was also motivated by the empirical asset pricing literature. Moreover, an example in [3] shows an economy where the existence of an equivalent martingale measure precludes the possibility of market completeness. Hence in references [3, 4, 15, 17], the concept of exact (almost everywhere) replication of a contingent claim *via* an admissible portfolio is substituted by the notion of *approximation* of a contingent claim. The main outlines of this approach are the following.

Let $\mathbb{M}$ denote the space of the P -absolutely continuous signed measures on $\mathcal{F}_T$. Then $Q \in \mathbb{M}$ can

be interpreted as a market agent's personal way of assigning values to claims, that is, the set $\mathbb{M}$ represents the possible contingent claims valuation measures held by traders. An agent using the valuation measure $Q \in \mathbb{M}$ assigns to a contingent claim H the value $\int H dQ$. The fact that $\mathbb{M}$ is given by the P -absolutely continuous signed measures on $\mathcal{F}_T$ has two particular meanings: first that all traders agree on null events, and second, that there can be strictly positive random variables with negative personal value. For a given trader, represented by $Q \in \mathbb{M}$, two contingent claims H_1 and H_2 are approximately equal if

$$|\int (H_1 - H_2) dQ| < \epsilon \text{ for small } \epsilon > 0 \quad (10)$$

Denote the space of all bounded contingent claims by $\mathcal{C}$. The finite intersections of the sets of the form $B(H_1, \epsilon) = \{H_2 \in \mathcal{C} \mid |\int (H_1 - H_2) dQ| < \epsilon\}$, $H_1 \in \mathcal{C}$, and $\epsilon > 0$, give a basis for a topology τ^Q on $\mathcal{C}$. We endow $\mathcal{C}$ with the coarsest topology τ finer than all of the τ^Q , $Q \in \mathbb{M}$. This topology is now agent independent, that is, two claims are approximately equal if all the agents believe that their values are close. The topology τ is usually referred as the *weak* topology* on $\mathcal{C}$ [21].

An agent is then allowed to trade in a finite number of assets *via* self-financing, bounded, stopping time simple strategies that yield a bounded payoff at T . As in the previous section, a (bounded) claim is said to be attainable if it can be replicated by one of such strategies. In this setting, the market is said to be *quasicomplete* if any contingent claim $H \in \mathcal{C}$ can be approximated by attainable claims in the weak* topology induced by $\mathbb{M}$ on $\mathcal{C}$. Since the weak* topology as well as the trading strategies are agent measure independent, the same is true for this notion of completeness. Consider now the space $\mathbb{P}_\pm$ of the P -absolutely continuous signed martingale measures. Then the following generalized version of the *second fundamental theorem* holds.

Theorem 4 (The second fundamental theorem of asset pricing, Theorem 2 of [3], Theorem 1 of [4], Theorem 5 of [17]). *Let $\mathbb{P}_\pm \neq \emptyset$. Then there exists a unique P -absolutely continuous signed martingale measure if and only if the market is quasicomplete.*

The proof of this theorem relies on the theory of linear operators between locally convex topological vector spaces.

Since the market is endowed with an infinite number of assets, in principle, trading in infinitely many assets may be possible. To take this possibility into account, in [5, 6, 15, 17] portfolios consisting of infinitely many assets are allowed by considering measure-valued strategies. The result of Theorem 4 still holds in the case of market models where measure-valued strategies are allowed as shown in Theorem 6.11 of [5] and Theorem 2.1 of [15].

This approach resolves the paradox of the counterexample of [2], since the economy considered in [2] is incomplete under this new definition of market completeness. Moreover, if $\mathbb{P} \neq \emptyset$ and the number of assets is finite or the asset prices are given by continuous processes, then Theorem 5 of [4] shows that the market model is quasicomplete if and only if it is complete.

End Notes

^aWe say that a contingent claim is integrable if $E^*[X/S_T^0] < \infty$. By Definition 1, it follows that an attainable contingent claim is necessarily integrable. Hence we can restate the definition of market completeness as follows. The model is said to be *complete* if every integrable claim is attainable.

^bThe price process is said to contain a redundancy if $P(\alpha \cdot S_{t+1} = 0|A) = 1$ for some nontrivial vector α , some $t < T$, and some $A \in \mathcal{P}_t$.

^cLet E be a Blackwell space. An E -multivariate point process is an integer-valued random measure on $[0, T] \times E$ with $\mu([0, t] \times E) < \infty$ for every $\omega, t \in [0, T]$ (see Definition III.1.23 of [13]).

References

- [1] Arrow, K. (1964). The role of securities in the optimal allocation of risk-bearing, *Review of Economics Studies* **31**, 91–96.
- [2] Artzner, P. & Heath, D. (1995). Approximate completeness with multiple martingale measures, *Mathematical Finance* **5**, 1–11.
- [3] Bättig, R. (1999). Completeness of securities market models—an operator point of view, *The Annals of Applied Probability* **9**, 529–566.
- [4] Bättig, R. & Jarrow, R.A. (1999). The second fundamental theorem of asset pricing: a new approach, *The Review of Financial Studies* **12**, 1219–1235.
- [5] Björk, T., Di Masi, G., Kabanov, Y. & Runggaldier, W. (1997). Towards a general theory of bond markets, *Finance and Stochastics* **1**, 141–174.
- [6] Björk, T.G., Kabanov, Y. & Runggaldier, W. (1997). Bond market structure in the presence of marked point processes, *Mathematical Finance* **7**, 211–223.
- [7] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- [8] Delbaen, F. (1992). Representing martingale measures when asset prices are continuous and bounded, *Mathematical Finance* **2**, 107–130.
- [9] Harrison, J.M. & Kreps, D.M. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- [10] Harrison, J.M. & Pliska, S.R. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and Their Applications* **11**, 215–260.
- [11] Harrison, J.M. & Pliska, S.R. (1983). A stochastic calculus model of continuous trading: complete markets, *Stochastic Processes and Their Applications* **15**, 313–316.
- [12] Jacod, J. (1979). *Calcul Stochastique et Problèmes des Martingales*, Lectures Notes in Mathematics, No. 714, Springer-Verlag, Berlin, Heidelberg, New York.
- [13] Jacod, J. & Shiryaev, A.N. (1987). *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin, Heidelberg, New York.
- [14] Jacod, J. & Yor, M. (1977). Etude des solutions extrémales et représentation intégrales des solutions pour certains problèmes des martingales, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **38**, 83–125.
- [15] Jarrow, R.A., Jin, X. & Madan, D.B. (1999). The second fundamental theorem of asset pricing, *Mathematical Finance* **9**, 255–273.
- [16] Jarrow, R.A. & Madan, D.B. (1991). A characterization of complete security markets on a Brownian filtration, *Mathematical Finance* **1**, 31–43.
- [17] Jarrow, R.A. & Madan, D.B. (1999). Hedging contingent claims on semimartingales, *Finance and Stochastics* **3**, 111–134.
- [18] Londono, J.A. (2004). State tameness: a new approach for credit constraints, *Electronic Communications in Probability* **9**, 1–13.
- [19] Müller, S.M. (1989). On complete securities markets and the martingale property of securities, *Economics Letters* **31**, 37–41.
- [20] Ross, S. (1976). The arbitrage theory of capital asset pricing, *Journal of Economic Theory* **13**, 341–360.
- [21] Rudin, W. (1991). *Functional Analysis*, 2nd Edition, MacGraw-Hill, New York.
- [22] Stiglitz, J. (1972). On the optimality of the stock market allocation of investment, *Quarterly Journal of Economics* **86**, 25–60.

- [23] Taqqu, M.S. & Willinger, W. (1987). The analysis of finite security markets using martingales, *Advances in Applied Probability* **19**, 1–25.
- [24] Yamada, T. & Watanabe, S. (1971). On the uniqueness of solutions of stochastic differential equations. *Journal of Mathematics of Kyoto University* **11**, 155–167.
- [25] Yor, M. (1977). *Remarques Sur la Représentation des Martingales Comme intégrales Stochastiques*, Séminaire de probabilités de Strasbourg XI, Lecture Notes in Mathematics, No. 581, Springer, New York, pp. 502–517.

Related Articles

Equivalence of Probability Measures; Equivalent Martingale Measures; Fundamental Theorem of Asset Pricing; Hedging; Martingales; Martingale Representation Theorem;

FRANCESCA BIAGINI

Expected Utility Maximization: Duality Methods

Expected utility maximization has a long tradition in modern mathematical finance. It dates back to the 1950s [18] when it provided a theoretical foundation to the (Markowitz's) mean–variance asset allocation method (see **Risk–Return Analysis**). The objective of a rational and risk-averse agent acting is captured by a concave function, the utility U of the agent (see **Utility Function**). It is typically assumed that U is increasing since the agent prefers more wealth to less. Given his/her U , the agent chooses the portfolio P^* that maximizes the agent's *expected utility* over a horizon $[0, T]$.

Some famous case studies are considered in [12, 13], where the agent is planning for retirement in a Black–Scholes (and thus complete) financial market (see **Merton Problem**). The complete market framework (see **Complete Markets**) is a convenient mathematical idealization as any conceivable risk can be hedged by cleverly investing in the market. As a consequence, independently of the specific utility of the agent, the price of any claim is also uniquely assigned since by the no-arbitrage principle it must coincide with the initial value of the hedging portfolio.

In the more realistic situation of incomplete market, when there are, for example, intrinsic, nontraded sources of risk, both the valuation and the hedging problems become highly nontrivial issues. Expected utility maximization has also turned out to perform very well in the pricing problem in the general, incomplete market setup. The related pricing techniques are known as *pricing by marginal utility* and *indifference pricing* and are discussed briefly in this article (for more details see **Utility Indifference Valuation**).

The use of increasingly more complex probabilistic models of financial assets has continued to pose new mathematical challenges. If the setup is that of general non-Markovian diffusion or semimartingale models, direct methods from stochastic optimal control (as originally done by Merton and many others after him) become increasingly difficult to handle. As first suggested by Bismut [4], *convex duality* (see

Convex Duality) is a powerful alternative approach. In the mid-1980s, with the works of Pliska [14], He and Pearson [8], Karatzas *et al.* [10], and Cox and Huang [5] this new methodology started to fully develop. Relying on convex duality (see **Convex Duality**) and martingale (see **Martingales**) methods, it enables the treatment of the most general cases. The price to pay for the achieved generality is that the results obtained have a mathematical existence–uniqueness–characterization form. As is always the case, explicit calculations require the specification of a (very) tractable model.

The presentation given here is based on the convex duality approach, in a general semimartingale model. For a treatment of the same problem with martingale methods in a diffusion context, see **Expected Utility Maximization** or [9].

Examples

Consider an agent who is a price taker, that is, his/her actions do not affect market prices, and whose goal is to trade dynamically in a financial market up to a horizon T , in order to achieve maximum expected utility. A host of features can be taken into account, such as the initial endowment, the possibility of intertemporal consumption, and the presence of a random endowment at time T . A list of various situations is given in the following. The mathematical details are discussed in the next section.

1. Maximizing Utility of Terminal Wealth

The preferences of the investor are represented by a von Neumann–Morgenstern utility function

$$U : \mathbb{R} \rightarrow [-\infty, +\infty) \quad (1)$$

which must be not identical to $-\infty$, increasing, and concave.

Typical examples are $U(x) = \ln x$, $U(x) = \frac{1}{\alpha} x^\alpha$ with $\alpha < 1$, $\alpha \neq 0$, where it is intended that $U(x) = -\infty$ outside the domain, and $U(x) = -\frac{1}{\gamma} e^{-\gamma x}$ with $\gamma > 0$.

No consumption occurs before time T . The agent has the initial endowment x and can invest in the financial market. The resulting optimization problem is

$$\sup_{k \in K(x)} E[U(k)] \quad (2)$$

where $K(x)$ is the set of random wealths that can be obtained at time T (terminal wealths) with initial wealth x .

The formulation of the problem with random endowment, namely, when the agent receives at T an additional cashflow B (say, an option), is the following:

$$\sup_{k \in K(x)} E[U(k + B)] \quad (3)$$

as his/her terminal possible wealths now are of the form $k + B$.

2. Maximizing Utility of Consumption

Suppose that the agent is not particularly interested in consumption at the terminal time T , but rather he/she is willing to consume over the entire planning horizon. A consumption plan C for the agent is determined by its random rate of consumption $c(t)$ at time t for all $t \in [0, T]$. It is evident from the financial perspective that the rate $c(t)$ must be nonnegative, so the consumption in the interval $[t, t + dt]$ increases by the quantity $c(t)dt$. The goal of the agent is thus the selection of the best consumption plan over $[0, T]$, starting with an initial endowment $x \geq 0$. The utility function will now measure the degree of satisfaction with the intertemporal consumption or better with the rate of consumption. As this measure may change over the time, the utility also depends on the time parameter:

$$\bar{U} : [0, T] \times \mathbb{R} \rightarrow [-\infty, +\infty) \quad (4)$$

When t is fixed, then $\bar{U}(t, \cdot)$ is a utility function with the same properties as in case (1). As the rate of consumption cannot be negative, $\bar{U}(t, x) = -\infty$ when $x < 0$. The agent may clearly benefit from the opportunity of investing in the financial market, so in general his/her position can be expressed by a consumption plan C and a dynamically changing portfolio P . If $X^{C,P}(t)$ is the total wealth of the position (C, P) at time t , then as there is no inflow of cash the variation of the wealth in $[t, t + dt]$ must satisfy

$$dX^{C,P}(t) = -c(t)dt + dV^P(t) \quad (5)$$

where $dV^P(t)$ is the variation of the value of the portfolio P at time t due to market fluctuations. Let $\mathcal{A}(x)$ indicate the set of all such consumption plans—portfolios (C, P) when starting from

the wealth level x . The maximization is then that of the expected integrated utility from the rate of consumption:

$$\sup_{(C,P) \in \mathcal{A}(x)} E \left[\int_0^T \bar{U}(t, c(t)) dt \right] \quad (6)$$

3. Maximizing Utility of Terminal Wealth and Consumption

Alternatively, the agent may wish to maximize expected utility from terminal wealth and intertemporal consumption given his/her initial wealth $x \geq 0$. Therefore, there are two utilities, U and $\bar{U}$, from terminal wealth and from the rate of consumption, respectively. Let $\mathcal{A}(x)$ be the set of the possible consumption plans—portfolios (C, P) , obtained with initial wealth x , and let $X^{C,P}(T)$ be the terminal wealth from the choice (C, P) . Then the optimal consumption–investment is the couple (C^*, P^*) that solves

$$\sup_{(C,P) \in \mathcal{A}(x)} \left\{ E \left[\int_0^T \bar{U}(t, c(t)) dt \right] + E [U(X^{C,P}(T))] \right\} \quad (7)$$

The case selected in the following section for the illustration of the duality technique and the main results is the first, that is, utility maximization of terminal wealth. When intertemporal consumption is taken into account, similar results can be proved. In addition, case 3 turns out to be a superposition of cases 1 and 2, as shown in Chapters 3, 6 of [9].

Maximizing the Utility of (Discounted) Terminal Wealth

An analysis of any optimization problem relies on a precise definition of the domain of optimization and the objective function. Therefore, the study of maximization (2) requires a specification of

1. the financial market model and the admissible terminal wealths;
2. the technical assumptions on U ; and
3. some joint condition on the market model and the utility function.

1. The financial market model considered is frictionless and consists of N risky assets and one

risk-free asset (money market account). Although it is not necessary, for the sake of convenience, it is assumed that the risk-free asset, S^0 , is constantly equal to 1, that is, the prices are *discounted*. The N risky assets are globally indicated by $S = (S^1, \dots, S^N)$. The trading can occur continuously in $[0, T]$. $S = (S_t)_{t \leq T}$ is, in fact, an $\mathbb{R}^N$ -valued, continuous-time process, defined on a filtered probability space $(\Omega, (\mathcal{F}_t)_{t \leq T}, \mathbb{P})$. Since the wealth from an investment in this market is a (stochastic) integral, S is assumed to be a semimartingale so that the object “integral with respect to S ” is mathematically well defined (see **Stochastic Integrals**). For expository reasons, S is a *locally bounded* semimartingale. This class of models is already very general, as all the diffusions are locally bounded semimartingales, as well as any jump-diffusion process with bounded jumps.

The agent has an initial endowment x and there are no restrictions on the quantities he/she can buy, sell, or sell short. $H_t = (H_t^1, \dots, H_t^N)$ is the random vector with the number of shares of each risky asset that the agent holds in the infinitesimal interval $[t, t + dt]$. B_t represents the number of shares of the risk-free asset held in the same interval. $H = (H_t)_t$ and $B = (B_t)_t$ are the corresponding processes and are referred to as the *strategy* of the agent. To be technically precise, H must be a predictable process and B a semimartingale. As there is no consumption and no infusion of money in the trading period $[0, T]$, the wealth from a strategy (H, B) is the process X that solves

$$\begin{cases} dX_t = (H_t dS_t + B_t dS_t^0) = H_t dS_t \\ X_0 = x \end{cases} \quad (8)$$

or, in integral form, $X_t = x + \int_0^t H_s dS_s$. This can be equivalently stated as the strategy (H, B) is self-financing. Since $dS^0 = 0$, the self-financing condition enables a representation of the wealth X only in terms of H . This is the reason one typically refers to H only as *the strategy*.

As usual in continuous-time trading (see **Fundamental Theorem of Asset Pricing**) to avoid phenomena like doubling strategies, not every self-financing H is allowed. A self-financing strategy H is said to be *admissible* only if during the trading the losses do not exceed a finite credit line. That is, H is admissible

if there exists some constant $c > 0$ such

$$\text{for all } t \in [0, T], \quad \int_0^t H_s dS_s \geq -c \quad \mathbb{P} - a.s. \quad (9)$$

so that for any x the wealth process $X = x + \int H_s dS_s$ is also bounded from below. Maximizing expected utility from terminal wealth means, in fact, *maximizing expected utility from the set $K(x)$ of those random variables X_T that can be represented as $X_T = x + \int_0^T H_t dS_t$ with H admissible in the sense of equation (9).*

Hereafter, the notation $E[\cdot]$ indicates $\mathbb{P}$ -expectation. When considering expectation under another probability $\mathbb{Q}$, the notation is explicitly $E_{\mathbb{Q}}[\cdot]$.

As shown by Delbaen and Schachermayer [7] a financially relevant set of probabilities is $\mathcal{M}^e$, namely, the set of the equivalent (local) martingale probabilities for S . When the market is complete, this set consists of only one probability, but in the general, incomplete market case, this set is infinite. Under each probability, $\mathbb{Q} \in \mathcal{M}^e$ S is a (local) martingale and thus $\mathbb{Q}$ is a risk-neutral probability. This is the theoretical justification for the use of each of these $\mathbb{Q}$ s as a pricing measure for any derivative claim B , with (arbitrage-free) price given by the expectation $E_{\mathbb{Q}}[B]$.

However, we need the less restrictive set $\mathcal{M}$ of the absolutely continuous (local) martingale probabilities $\mathbb{Q}$ for S , as this is the set that will show up in the dual problem. The set $\mathcal{M}$ can be characterized in the following manner:

$$\mathcal{M} = \left\{ \mathbb{Q} \ll \mathbb{P} \mid E_{\mathbb{Q}} \left[\int_0^T H_t dS_t \right] \leq 0 \quad \forall \text{ adm. } H \right\} \quad (10)$$

as the set of absolutely continuous probability measures that give nonpositive expectation to the terminal wealths from admissible self-financing strategies starting with zero wealth. Therefore, given any $X_T \in K(x)$ and any $\mathbb{Q} \in \mathcal{M}$,

$$E_{\mathbb{Q}}[X_T] = E_{\mathbb{Q}} \left[x + \int_0^T H_t dS_t \right] \leq x \quad (11)$$

2. Hypothesis on U . As a case study, let us assume that U is finite valued on $\mathbb{R}$, that is, the wealth can become arbitrarily negative (the closest

references are [2, 16]). A typical example is the exponential utility. The reason we prefer the exponential utility (and all the other utilities with the properties listed below) to, for example, the logarithmic or the power utilities is that the dual problem is easier to interpret. References for the case when there are constraints on the wealth (then U is finite only on a half-line), like $U(x) = \ln x$ or $U(x) = \frac{1}{\alpha}x^\alpha$, are [11], [17], and the bibliography contained therein.

A main difficulty the reader may encounter when comparing this literature is that the language and style in the papers differ. Very recently, Biagini and Frittelli [3] proposed a unifying approach that works both for the case of U finite on all $\mathbb{R}$ and for the case of U finite only on a half-line. The result there is enabled by the choice of an innovative duality (an Orlicz space duality), naturally induced by the utility function U .

Regarding U , it is here required that

- U is strictly concave, strictly increasing, and differentiable over $(-\infty, +\infty)$ and
- $\lim_{x \downarrow -\infty} U'(x) = -\infty$ and $\lim_{x \rightarrow +\infty} U'(x) = 0$ (these are known as the *Inada condition* on the marginal utility U').

In addition, U must satisfy the reasonable asymptotic elasticity condition $RAE(U)$ introduced in [11, 16]:

$$\liminf_{x \rightarrow -\infty} \frac{xU'(x)}{U(x)} > 1, \quad \limsup_{x \rightarrow +\infty} \frac{xU'(x)}{U(x)} < 1 \quad (12)$$

In the cited references, it is also shown that this condition is necessary and sufficient for the duality to work properly if U is fixed and one considers all possible financial markets. However, *within a specific market model*, one may state more general necessary and sufficient conditions on U that enable the duality approach and ensure the existence of the optimal investment. We choose to impose $RAE(U)$ for it has the advantage that it can be easily verified. Note also that $RAE(U)$ is already satisfied by the commonly used utility functions, for example, by the classic exponential function $U(x) = -\frac{1}{\gamma}e^{-\gamma x}$.

3. The convex conjugate V and a joint condition between preferences and the market. The conjugate V of U is the function

$$V(y) = \sup_x \{U(x) - yx\} \quad (13)$$

and, apart from some minus signs, it coincides with the Fenchel conjugate of U (see **Convex Duality**). Thus, V is a convex function, which is identically equal to $+\infty$ when $y < 0$. It is also differentiable on $(0, +\infty)$ and its derivative is $V' = -(U')^{-1}$. Traditionally, the inverse of the marginal utility $(U')^{-1}$ is denoted by I . By mere definition of V , for all x, y , the Fenchel inequality holds

$$U(x) \leq xy + V(y) \quad (14)$$

and the above relation is, in fact, an equality iff $y = U'(x)$ or equivalently $x = (U')^{-1}(y) = I(y)$. Also note that

$$U(x) = \inf_y \{xy + V(y)\} = \inf_{y>0} \{xy + V(y)\} \quad (15)$$

The typical example (and most used) is the following couple (U, V) :

$$U(x) = -\frac{1}{\gamma}e^{-\gamma x}$$

$$V(y) = \begin{cases} \frac{1}{\gamma}(y \ln y - y) & y > 0 \\ 0 & y = 0 \\ +\infty & y < 0 \end{cases} \quad (16)$$

Let us recall that a probability $\mathbb{Q}$ absolutely continuous with respect to $\mathbb{P}$ is said to have *finite generalized entropy* (or, also, *finite V -divergence*), if its density $\frac{d\mathbb{Q}}{d\mathbb{P}}$ is integrable when composed with V

$$E \left[V \left(\frac{d\mathbb{Q}}{d\mathbb{P}} \right) \right] < +\infty \quad (17)$$

The joint condition required between preferences and the market is actually a condition between V and the set of probabilities $\mathcal{M}$, which is as follows:

Condition 1 There exists a $Q^0 \in \mathcal{M}$ with finite generalized entropy, that is, $E \left[V \left(\frac{dQ^0}{dP} \right) \right] < +\infty$.

Duality in Complete Market Models

Suppose that the market is complete and arbitrage free. Thus, there exists a unique equivalent martingale measure $\mathbb{Q} \in \mathcal{M}^e$, which, by Condition 1, has also

finite generalized entropy. Let us restate problem (2), the *primal* problem

$$u(x) := \sup_{X_T \in K(x)} E[U(X_T)] \quad (18)$$

where $u(x)$ denotes the optimal level of the expected utility. It is not difficult to derive an upper bound for $u(x)$. From inequality (14), in fact, for all $X_T \in K(x)$ and for all $y > 0$

$$U(X_T) \leq X_T y \frac{d\mathbb{Q}}{d\mathbb{P}} + V\left(y \frac{d\mathbb{Q}}{d\mathbb{P}}\right) \quad (19)$$

and taking $\mathbb{P}$ -expectations on both sides

$$E[U(X_T)] \leq xy + E\left[V\left(y \frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] \quad (20)$$

because $E\left[X_T \frac{d\mathbb{Q}}{d\mathbb{P}}\right] = E_{\mathbb{Q}}[X_T] \leq x$. Therefore, taking the supremum over X_T and the infimum over y ,

$$\begin{aligned} u(x) &= \sup_{X_T \in K(x)} E[U(X_T)] \leq \inf_{y>0} xy + E\left[V\left(y \frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] \\ &\quad (21) \end{aligned}$$

As noted by Merton, the above supremum is not necessarily reached over the restricted set of admissible terminal wealths $K(x)$. Following a well-known procedure in the calculus of variations, a *relaxation* of the primal problem allows to obtain the optimal terminal wealth. Here, this means enlarging $K(x)$ slightly and considering the larger set

$$K_{\mathbb{Q}}(x) := \{k \in L^1(\mathbb{Q}) \mid E_{\mathbb{Q}}[k] \leq x\} \quad (22)$$

$K_{\mathbb{Q}}(x)$ is simply the set of claims that have initial price smaller or equal to the initial endowment x . An application of the separating hyperplane theorem gives that $K_{\mathbb{Q}}(x)$ is the norm closure of $K(x) - L_+^1(\mathbb{Q})$ in $L^1(\mathbb{Q})$. Then, an approximation argument shows that the optimal expected value $u(x)$ and

$$u_{\mathbb{Q}}(x) := \sup_{k \in K_{\mathbb{Q}}(x)} E[U(k)] \quad (23)$$

are, in fact, equal. The *relaxed* maximization problem over $K_{\mathbb{Q}}(x)$ is much simpler than the original one over $K(x)$. The replication-with-admissible-strategies issue has been removed and there is just an inequality constraint, given by the pricing measure $\mathbb{Q}$. To find

out the value $u_{\mathbb{Q}}(x) = u(x)$, one can now apply the traditional Lagrange multiplier method to get

$$\begin{aligned} u_{\mathbb{Q}}(x) &= \sup_{k \in K_{\mathbb{Q}}(x)} E[U(k)] \\ &= \sup_{k \in L^1(\mathbb{Q})} \inf_{y>0} \{E[U(k)] + y(x - E_{\mathbb{Q}}[k])\} \end{aligned} \quad (24)$$

The dual problem is defined by exchanging the inf and the sup in the above expression:

$$\inf_{y>0} \sup_{k \in L^1(\mathbb{Q})} \{E[U(k)] + y(x - E_{\mathbb{Q}}[k])\} \quad (25)$$

From [15 Theorem 21] or from a direct computation, the inner sup is actually equal to

$$xy + E\left[V\left(y \frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] \quad (26)$$

so that the dual problem takes the traditional form

$$\inf_{y>0} \left\{ xy + E\left[V\left(y \frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] \right\} \quad (27)$$

which is exactly the right-hand side of equation (21).

Thanks to Condition 1, the dual problem is always finite valued and so is u . *A priori*, however, one has only the chain $u(x) = u_{\mathbb{Q}}(x) \leq \inf_{y>0} \left\{ xy + E\left[V\left(y \frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] \right\}$, but under the current assumptions there is *no duality gap*:

$$u(x) = u_{\mathbb{Q}}(x) = \inf_{y>0} \left\{ xy + E\left[V\left(y \frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] \right\} \quad (28)$$

the infimum is a minimum and the supremum over $K_{\mathbb{Q}}(x)$ is reached. In fact, the *RAE(U)* condition on the utility function U implies that $E\left[V\left(y \frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] < +\infty \forall y > 0$, so the infimum in (27) can be obtained by differentiation under the expectation sign. The dual minimizer y^* (which depends on x) is then the unique solution of

$$x + E_{\mathbb{Q}}\left[V'\left(y \frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] = 0 \quad (29)$$

or, equivalently, y^* is the unique solution of

$$E_{\mathbb{Q}}\left[I\left(y \frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] = x \quad (30)$$

Therefore, the (unique) optimal claim is $k^* = I\left(y^* \frac{dQ}{dP}\right)$ because it verifies the following:

- the balance equation $E_Q[k^*] = x$, so $k^* \in K_Q(x)$ and
- the Fenchel equality

$$U(k^*) = k^* y^* \frac{dQ}{dP} + V\left(y^* \frac{dQ}{dP}\right) \quad (31)$$

from which, by taking the $\mathbb{P}$ -expectations, we get

$$\begin{aligned} E[U(k^*)] &= y^* E\left[k^* \frac{dQ}{dP}\right] + E\left[V\left(y^* \frac{dQ}{dP}\right)\right] \\ &= y^* x + E\left[V\left(y^* \frac{dQ}{dP}\right)\right] \end{aligned} \quad (32)$$

which proves the main equality (28).

By market completeness, the martingale representation theorem applies, so that k^* can be obtained via a self-financing strategy H^* :

$$k^* = x + \int_0^T H_t^* dS_t \quad (33)$$

though H^* is not admissible in general, that is, when optimally investing, the agent can incur arbitrarily large losses.

Moreover, as a function of x , the optimal value $u(x)$ is also a utility function finite on $\mathbb{R}$, with the same properties of U . The duality equation (28) shows that u and

$$v(y) = \begin{cases} E\left[V\left(y \frac{dQ}{dP}\right)\right] & \text{if } y \geq 0 \\ +\infty & \text{otherwise} \end{cases} \quad (34)$$

are conjugate functions.

The relationship between the primal and dual optima can also be expressed as

$$\frac{dQ}{dP} = \frac{1}{y^*} U'(k^*) \quad (35)$$

which amounts to saying that $\frac{dQ}{dP}$ is proportional to one's marginal utility from the optimal investment. Therefore, in the complete market case, pricing by taking Q -expectations coincides with the *pricing by marginal utility principle*, introduced in the option pricing context by Davis [6].

Duality in Incomplete Market Models

The same methodology applies to the incomplete market framework, but the technicalities require some more effort. The main results are (more or less intuitive) generalizations of what happens in the complete case, as summarized below (see [2, 16] for the proofs).

1. The duality relation is the natural generalization of equation (28):

$$\begin{aligned} u(x) &= \sup_{X_T \in K(x)} E[U(X_T)] \\ &= \inf_{y > 0, Q \in \mathcal{M}} \left\{ xy + E\left[V\left(y \frac{dQ}{dP}\right)\right] \right\} \end{aligned} \quad (36)$$

and there exists a unique couple of dual minimizers y^*, Q^* .

2. As in the complete case, the supremum of the expected utility on $K(x)$ may be not reached. $K_V(x)$ denotes the set of $k \in L^1(Q)$ such that $E_Q[k] \leq x$ for all $Q \in \mathcal{M}$ with finite generalized entropy. Then, the supremum of the expected utility on $K_V(x)$ coincides with the value $u(x)$ and it is a maximum. The claim $k^* \in K_V$ attaining the maximum is unique and the relationship between primal and dual optima still holds

$$\frac{dQ^*}{dP} = \frac{1}{y^*} U'(k^*) \quad (37)$$

3. Q^* may be not equivalent to $\mathbb{P}$. However, in the case $Q^* \sim \mathbb{P}$, k^* can be obtained through a self-financing strategy H^* , albeit not admissible in general.
4. The optimal value u as a function of the initial endowment x is a utility function, with the same properties of U . In fact, it is finite on $\mathbb{R}$, strictly concave, strictly increasing, it verifies the Inada conditions, and $RAE(u)$ holds. The duality relation (36), rewritten as $u(x) = \inf_{y > 0} \{xy + v(y)\}$ with

$$v(y) = \begin{cases} \inf_{Q \in \mathcal{M}} E\left[V\left(y \frac{dQ}{dP}\right)\right] & \text{if } y \geq 0 \\ +\infty & \text{otherwise} \end{cases} \quad (38)$$

shows that u and v are conjugate functions (see **Convex Duality**).

As $\mathbb{Q}^*$ results from a minimax theorem, it is also known as the *minimax measure*. For the applications, it is important to know that there are easy sufficient conditions that guarantee that $\mathbb{Q}^*$ is equivalent to $\mathbb{P}$, such as the following: (i) $U(+\infty) = +\infty$ as noted in [1] or (ii) in case $U(x) = -\frac{1}{\gamma}e^{-\gamma x}$, the existence of a $\mathbb{Q} \in \mathcal{M}^e$ with finite generalized entropy (see [17] for an extensive bibliography).

When $\mathbb{Q}^*$ is indeed equivalent to $\mathbb{P}$, its selection in the class of all risk-neutral, equivalent probabilities $\mathcal{M}^e$ as the pricing measure is economically motivated by its proportionality to the marginal utility from the optimal investment.

Utility Maximization with Random Endowment

Under all the conditions stated above (on the market, on U , and on both), suppose that the agent has a random endowment B at T , in addition to the initial wealth x . For example, B can be the payoff of a European option expiring at T . The agent's goal is still maximizing of expected utility from terminal wealth, which now becomes

$$u(x, B) := \sup_{X_T \in K(x)} E[U(B + X_T)] \quad (39)$$

The duality results, in this case, are similar to the ones just shown. In fact,

$$\begin{aligned} u(x, B) \\ = \min_{y > 0, \mathbb{Q} \in \mathcal{M}} \left\{ xy + yE_{\mathbb{Q}}[B] + E \left[v \left(y \frac{d\mathbb{Q}}{d\mathbb{P}} \right) \right] \right\} \end{aligned} \quad (40)$$

Note that the maximization without the claim can be seen as a particular case of the one above, with $B = 0$: $u(x, 0) = u(x)$. The solution of a utility maximization problem with random endowment is the key step to the indifference pricing technique. The (buyer's) indifference price of B is, in fact, the unique price p_B that solves

$$u(x - p, B) = u(x, 0) \quad (41)$$

This means that the agent is indifferent, that is, he/she has the same (optimal expected) utility, between (i) paying p_B at time $t = 0$ and receiving B at T and (ii) not entering into a deal for the claim B .

References

- [1] Bellini, F. & Frittelli, M. (2002). On the existence of minimax martingale measures, *Mathematical Finance* **12/1**, 1–21.
- [2] Biagini, S. & Frittelli, M. (2005). Utility maximization in incomplete markets for unbounded processes, *Finance and Stochastics* **9**, 493–517.
- [3] Biagini, S. & Frittelli, M. (2008). A unified framework for utility maximization problems: an Orlicz space approach, *Annals of Applied Probability* **18/3**, 929–966.
- [4] Bismut, J.M. (1973). Conjugate convex functions in optimal stochastic control, *Journal of Mathematical Analysis and Applications* **44**, 384–404.
- [5] Cox, J.C. & Huang, C.F. (1989). Optimal consumption and portfolio policies when asset prices follow a diffusion process, *Journal of Economic Theory* **49**, 33–83.
- [6] Davis, M.H.A. (1997). Option pricing in incomplete markets, in *Mathematics of Derivative Securities*, M. Dempster & S.R. Pliska, eds, Cambridge University Press, pp. 216–227.
- [7] Delbaen, F. & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing, *Mathematische Annalen* **300**, 463–520.
- [8] He, H. & Pearson, N.D. (1991). Consumption and portfolio policies with incomplete markets and short-sale constraints: the infinite-dimensional case, *Journal of Economic Theory* **54**, 259–304.
- [9] Karatzas, I. & Shreve, S. (1998). *Methods of Mathematical Finance*, Springer.
- [10] Karatzas, I., Shreve, S., Lehoczky, J. & Xu, G. (1991). Martingale and duality methods for utility maximization in an incomplete market, *SIAM Journal on Control and Optimization* **29**, 702–730.
- [11] Kramkov, D. & Schachermayer, W. (1999). The asymptotic elasticity of utility function and optimal investment in incomplete markets, *Annals of Applied Probability* **9/3**, 904–950.
- [12] Merton, R.C. (1969). Lifetime portfolio selection under uncertainty: the continuous-time case, *The Review of Economics and Statistics* **51**, 247–257.
- [13] Merton, R.C. (1971). Optimum consumption and portfolio rules in a continuous-time model, *Journal of Economic Theory* **3**, 373–413.
- [14] Pliska, S.R. (1986). A stochastic calculus model of continuous trading: optimal portfolios, *Mathematics of Operations Research* **11**, 371–382.

- [15] Rockafellar, R.T. (1974). *Conjugate Duality and Optimization*, Conference Board of Math. Sciences Series, SIAM Publications, No. 16.
- [16] Schachermayer, W. (2001). Optimal investment in incomplete markets when wealth may become negative, *Annals of Applied Probability* **11/3**, 694–734.
- [17] Schachermayer, W. (2004). Portfolio Optimization in incomplete financial markets, *Notes of the Scuola Normale Superiore di Pisa, Cattedra Galileiana* downloadable at <http://www.fam.tuwien.ac.at/~wschach/pubs/>
- [18] Tobin, J. (1958). Liquidity preference as behavior towards risk, *Review of Economic Studies* **25**, 68–85.

Related Articles

Complete Markets; Convex Duality; Equivalent Martingale Measures; Expected Utility Maximization; Merton Problem; Second Fundamental Theorem of Asset Pricing; Utility Function; Utility Indifference Valuation.

SARA BIAGINI

Change of Numeraire

Consider a financial market model with nondividend paying asset price processes $(S^0, S^1, \dots, S^N)$ living on a filtered probability space $(\Omega, \mathcal{F}, \mathbf{F}, P)$, where $\mathbf{F} = \{\mathcal{F}_t\}_{t \geq 0}$ and P is the objective probability measure. For general results concerning completeness, self-financing portfolios, martingale measures, and arbitrage, (see **Arbitrage Strategy; Fundamental Theorem of Asset Pricing; Risk-neutral Pricing**).

We choose the asset S^0 as the *numeraire asset*, and we assume that $S_t^0 > 0$ with probability 1. From general theory, we know that (modulo integrability and technical conditions) the market is free of arbitrage if and only if there exists a measure $Q^0 \sim P$ such that the normalized price processes

$$\frac{S_t^0}{S_t^0}, \frac{S_t^1}{S_t^0}, \dots, \frac{S_t^N}{S_t^0}$$

are Q^0 martingales. Using the notation $Z^i = S^i/S^0$, thus we also have, apart from the nominal price system $S^0, S^1, \dots, S^N$, the normalized price system $Z^0, Z^1, \dots, Z^N$. The economic importance of the normalized system is clarified by the following standard result.

Proposition 1 *With notation as defined above the following hold.*

- A portfolio is self-financing in the S system if and only if it is self-financing in the Z system.
- A portfolio is an arbitrage opportunity in the S system if and only if it is an arbitrage in the Z system.
- The S market is complete if and only if the Z market is complete.
- In the Z market, the asset Z^0 has the property that $Z_t^0 \equiv 1$, so it represents a bank account with zero interest rate.

If $X \in \mathcal{F}_T$ is a fixed contingent claim with exercise date T , and if we denote the (not necessarily unique) arbitrage-free price process of X by $\Pi_t[X]$, then by applying the above-mentioned result to the extended market $S^0, S^1, \dots, S^N, \Pi_t[X]$ we see that $\Pi_t[X]/S_t^0$ is a Q^0 martingale, and using this fact together with the obvious fact that $\Pi_T[X] = X$ we obtain the basic

pricing formula:

$$\Pi_t[X] = S_t^0 E^0 \left[\frac{X}{S_T^0} \middle| \mathcal{F}_t \right] \quad (1)$$

where E^0 denotes integration with respect to (w.r.t.) Q^0 .

Very often one uses the bank account B with dynamics

$$dB_t = r_t B_t dt, \quad B_0 = 1 \quad (2)$$

where r is the short rate, as numeraire. The corresponding martingale measure Q^B is then often denoted by Q and referred to as “the risk neutral martingale measure”. In this case, the pricing formula becomes

$$\Pi_t[X] = E^Q \left[e^{-\int_t^T r_s ds} X \middle| \mathcal{F}_t \right] \quad (3)$$

In many concrete situations, the computational work needed for the determination of arbitrage-free prices can be drastically reduced by a clever choice of numeraire, and the purpose of this article is to analyze such changes.

To set the scene, we consider a fixed risk neutral martingale measure Q for the numeraire B , and an alternative numeraire asset S^0 with the corresponding martingale measure Q^0 . Our first task is to find the measure transformation between Q and Q^0 .

To see what Q^0 must look like, we consider a fixed time T and an arbitrarily chosen T -claim X . Assuming enough integrability we then know that, by using B as the numeraire, the arbitrage-free price of X at time $t = 0$ is given as

$$\Pi_0[X] = E^Q \left[\frac{X}{B_T} \right] \quad (4)$$

On the other hand, using S^0 as numeraire, the price is also given by the following formula:

$$\Pi_0[X] = S_0^0 E^0 \left[\frac{X}{S_T^0} \right] \quad (5)$$

Defining the likelihood process L by $L_t = dQ^0/dQ$ on $\mathcal{F}_t$, we thus have

$$E^Q \left[\frac{X}{B_T} \right] = S_0^0 E^0 \left[L_T \frac{X}{S_T^0} \right] \quad (6)$$

Since this holds for all $X \in \mathcal{F}_T$, we have the following basic result.

Proposition 2 *Under the above-mentioned assumptions, the likelihood process L , defined as*

$$L_t = \frac{dQ^0}{dQ}, \quad \text{on } \mathcal{F}_t, \quad 0 \leq t \leq T \quad (7)$$

is given by the formula

$$L_t = \frac{S_t^0}{S_0^0 \cdot B_t} \quad (8)$$

We note that since S^0/B is a Q martingale, the likelihood process L is also, as expected, a Q martingale.

As an immediate corollary we have the following.

Proposition 3 *Assume that the S^0 dynamics under the Q measure are of the form*

$$dS_t^0 = r_t S_t^0 dt + S_t^0 \sigma_t dW_t^Q \quad (9)$$

where W^Q is a d -dimensional Q Wiener process, r is the short rate, and σ is a d -dimensional optional row vector process. Then the dynamics for the likelihood process L are of the form

$$dL_t = L_t \sigma_t dW_t^Q \quad (10)$$

We can thus easily construct the relevant Girsanov transformation directly from the volatility of the S^0 -process.

We can, in a straightforward manner, extend Proposition 3 to change from one numeraire Q^0 to another numeraire Q^1 . The proof is obvious.

Proposition 4 *Let S^0 and S^1 be two strictly positive numeraire assets with the corresponding martingale measures Q^0 and Q^1 . Denote the likelihood process $L^{0,1}$ as*

$$L_t^{0,1} = \frac{dQ^1}{dQ^0}, \quad \text{on } \mathcal{F}_t \quad (11)$$

Then $L^{0,1}$ is given by

$$L_t^{0,1} = \frac{S_t^1}{S_t^0} \cdot \frac{S_0^0}{S_0^1} \quad (12)$$

Remark 1 It may perhaps seem surprising that even in the case of an incomplete market, we obtain a

unique martingale measure Q^0 . In more detail, the situation is as follows.

- If the market is incomplete, then there will exist several risk neutral measures Q .
- Each of these measures generates a different price system, defined by the pricing formula (3).
- Choosing one particular Q is thus equivalent to choosing one particular price system.
- For a given numeraire S^0 , there will also exist several different martingale measures Q^0 .
- Each of these measures generates a different price system, defined by the pricing formula (1).
- If a risk neutral measure Q and thus a price system are fixed, there exists a unique measure Q^0 such that Q^0 generates the same price system as Q .
- The measure transformations considered here are precisely those corresponding to a change of measure within a given price system.

Pricing Homogeneous Contracts

Using a numeraire S^0 is particularly useful when the claim X is of the form $X = S_T^0 \cdot Y$, since then we obtain the following simple expression:

$$\Pi_t[X] = S_t^0 E^0[Y|\mathcal{F}_t] \quad (13)$$

A typical example when this situation occurs is when dealing with derivatives defined in terms of several underlying assets. Assume, for example, that we are given two asset prices S^0 and S^1 , and that the contract X to be priced is of the form $X = \Phi(S_T^0, S_T^1)$, where Φ is a given linearly homogeneous function. Using the standard machinery, we would have to compute the price as

$$\Pi_t[X] = E^0 \left[e^{-\int_t^T r(s) ds} \Phi(S_T^0, S_T^1) \middle| \mathcal{F}_t \right] \quad (14)$$

which essentially amounts to the calculation of a triple integral. If we instead use S^0 as numeraire we have

$$\begin{aligned} \Pi_t[X] &= S_t^0 E^0 \left[\frac{1}{S_T^0} \Phi(S_T^0, S_T^1) \middle| \mathcal{F}_t \right] \\ &= S_t^0 E^0 \left[\Phi \left(1, \frac{S_T^1}{S_T^0} \right) \middle| \mathcal{F}_t \right] \end{aligned}$$

$$= S_t^0 E^0[\varphi(Z_T) \mathcal{F}_t] \quad (15)$$

where $\varphi(z) = \Phi(1, z)$ and $Z_T = S_T^1 / S_T^0$. Note that the factor S_t^0 is the price of the traded asset S^0 at time t , so this quantity does not have to be computed—it can be directly observed on the market. Thus, the computational work is reduced to computing a single integral.

As an example, assume that we have two stocks, S^0 and S^1 , with price processes of the following form under the objective probability P :

$$dS_t^0 = \alpha S_t^0 dt + \sigma S_t^0 d\tilde{W}_t^0 \quad (16)$$

$$dS_t^1 = \beta S_t^1 dt + \delta S_t^1 d\tilde{W}_t^1. \quad (17)$$

Here $\tilde{W}^0$ and $\tilde{W}^1$ are assumed to be independent P -Wiener processes, but it would also be easy to treat the case when there is a coupling between the two assets.

Under Q the price dynamics will be given as

$$dS_t^0 = r S_t^0 dt + \sigma S_t^0 dW_t^0 \quad (18)$$

$$dS_t^1 = r S_t^1 dt + \delta S_t^1 dW_t^1 \quad (19)$$

where W^0 and W^1 are Q -Wiener processes, and from Proposition 3 it follows that the Girsanov transformation from Q to Q^0 has a likelihood process with dynamics given as

$$dL_t = L_t \sigma dW_t^0 \quad (20)$$

The T -claim to be priced is an exchange option, which gives the holder the right, but not the obligation, to exchange one S^0 share for one S^1 share at time T . Formally, this means that the claim is given by $X = \max[S_T^1 - S_T^0, 0]$, and we note that we have a linearly homogeneous contract function. From equation (15), the price process is given as

$$\Pi_t[X] = S_t^0 E^0[\max[Z_T - 1, 0] | \mathcal{F}_t] \quad (21)$$

with $Z(t) = S_t^1 / S_t^0$. We are thus, in fact, valuing a European call option on Z_T , with strike price $K = 1$.

By construction, Z will be a Q^0 -martingale, and since a Girsanov transformation will not affect the volatility, it follows easily from equations (16) and (17) that the Q^0 -dynamics of Z are given by

$$dZ_t = Z_t \sqrt{\sigma^2 + \delta^2} dW_t \quad (22)$$

where W is a standard Q^0 -Wiener process. The price is thus given by the following formula:

$$\Pi_t[X] = S_t^0 \cdot c(t, Z_t) \quad (23)$$

Here $c(t, z)$ is given directly by the Black–Scholes formula as the price of a European call option, valued at t , with time of maturity T , strike price $K = 1$, short rate $r = 0$, on a stock with volatility $\sqrt{\sigma^2 + \delta^2}$ and price z .

Forward Measures

We now specialize the theory to the case when the chosen numeraire is a zero coupon bond. As can be expected, this choice of numeraire is particularly useful when dealing with interest rate derivatives.

Suppose, therefore, that we are given a specified bond market model with a fixed risk neutral martingale measure Q (always with B as numeraire). For a fixed time of maturity T , we now choose the price process $p(t, T)$, of a zero coupon bond maturing at T , as our new numeraire.

Definition 1 *The T -forward measure Q^T is defined as*

$$L_t^T = \frac{dQ^T}{dQ} \quad (24)$$

on $\mathcal{F}_t$ for $0 \leq t \leq T$ where L^T is defined as

$$L_t^T = \frac{p(t, T)}{B_t p(0, T)} \quad (25)$$

Observing that $p(T, T) = 1$ we have the following useful pricing formula as an immediate corollary of Proposition 3.

Proposition 5 *For any sufficiently integrable T -claim X , we have the pricing formula*

$$\Pi_t[X] = p(t, T) E^T[X | \mathcal{F}_t] \quad (26)$$

where E^T denotes integration w.r.t. Q^T .

Note again that the price $p(t, T)$ does not have to be computed. It can be observed directly on the market at time t .

A natural question to ask is when Q and Q^T coincide. This occurs if and only if we Q -a.s. have

$L^T(T) = 1$, that is when

$$1 = \frac{p(T, T)}{B_T p(0, T)} = \frac{e^{-\int_0^T r(s) ds}}{E^Q \left[e^{-\int_0^T r(s) ds} \right]} \quad (27)$$

that is if and only if r is deterministic.

The General Option Pricing Formula

We now present a fairly general formula for the pricing of European call options. Therefore, assume that we are given a financial market with a (possibly stochastic) short rate r and a strictly positive asset price process S . We also assume the existence of a risk neutral martingale measure Q .

Consider now a fixed time T , and a European call on S with exercise date T and strike price K . We are, thus, considering the T -claim:

$$X = \max[S_T - K, 0] \quad (28)$$

The main trick when dealing with options is to write X as

$$\begin{aligned} X &= (S_T - K) \cdot I\{S_T \geq K\} \\ &= S_T \cdot I\{S_T \geq K\} - K \cdot I\{S_T \geq K\} \end{aligned} \quad (29)$$

where I denotes an indicator function. Using the linear property of pricing we thus obtain

$$\Pi_t[X] = \Pi_t[S_T \cdot I\{S_T \geq K\}] - K \cdot \Pi_t[I\{S_T \geq K\}] \quad (30)$$

For the first term, we change to the measure Q^S having S as numeraire, and for the second term, we use the T -forward measure. Using the pricing formula (1) twice, once for each numeraire, we obtain the following basic option pricing formula, where we recognize the structure of the standard Black–Scholes formula.

Proposition 6 *Given the above-mentioned assumptions, the option price is given as*

$$\begin{aligned} \Pi_t[X] &= S_t Q^S(S_T \geq K | \mathcal{F}_t) \\ &\quad - K p(t, T) Q^T(S_T \geq K | \mathcal{F}_t) \end{aligned} \quad (31)$$

Notes

The first use of a numeraire different from the risk-free asset B was probably in [8] where, however,

the technique is not explicitly discussed. The first explicit use of a change of numeraire change was in [7], where an underlying stock was used as numeraire in order to value an exchange option. The numeraire change is also used in [4, 5] and basically in all later works on the existence of martingale measures in order to reduce the general case to the basic case of zero short rate. In these papers, the numeraire change as such is, however, not put to systematic use as an instrument for facilitating the computation of option prices in complicated models. In the context of interest rate theory, changes of numeraire were used and discussed independently in [2] and (within a Gaussian framework) in [6], where in both cases a bond maturing at a fixed time T is used as numeraire. A systematic study of general changes of numeraire can be found in [3]. For further examples of the change of numeraire technique see [1].

References

- [1] Benninga, S., Björk, T. & Wiener, Z. (2002). On the use of numeraires in option pricing, *Journal of Derivatives* 43–58.
- [2] Geman, H. (1989). *The Importance of the Forward Neutral Probability in a Stochastic Approach of Interest Rates*. Working paper, ESSEC, 10.
- [3] Geman, H., El Karoui, N. & Rochet, J.-C. (1995). Changes of numeraire, changes of probability measure and option pricing, *Journal of Applied Probability* 32, 443–458.
- [4] Harrison, J. & Kreps, J. (1979). Martingales and arbitrage in multiperiod markets, *Journal of Economic Theory* 11, 418–443.
- [5] Harrison, J. & Pliska, S. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and Applications* 11, 215–260.
- [6] Jamshidian, F. (1989). An exact bond option formula, *Journal of Finance* 44, 205–209.
- [7] Margrabe, W. (1978). The value of an option to exchange one asset for another, *Journal of Finance* 33, 177–186.
- [8] Merton, R. (1973). The theory of rational option pricing, *Bell Journal of Economics and Management Science* 4, 141–183.

Related Articles

Forward and Swap Measures.

TOMAS BJÖRK

Utility Indifference Valuation

Under market frictions like illiquidity or transaction costs, contingent claims can incorporate some inevitable intrinsic risk that cannot be completely hedged away but remains with the holder. In general, they cannot be synthesized by dynamical trading in liquid assets and hence cannot be priced by no-arbitrage arguments alone. Still, an agent (she) can determine a valuation with respect to her preferences towards risk. The utility indifference value for a variation in the quantity of illiquid assets held by the agent is defined as the compensating variation of wealth, under which her maximal expected utility remains unchanged.

Consider an agent acting in a financial market with $d + 1$ liquid assets, which can be traded at market prices in a frictionless way anytime up to horizon $T < \infty$. In addition, there are J illiquid assets providing risky payouts $(B^j)_{j=1,\dots,J}$ at T . A preference order of the agent is described by an (indirect) utility function $u_t^b(x)$, describing the maximal expected utility obtainable when holding at t a position consisting of wealth $x \in \mathbb{R}$ invested in liquid assets (at market prices) and $b \in \mathbb{R}^J$ shares of illiquid assets. At time t , the agent prefers a position (x, b) to (x', b') if $u_t^b(x) \geq u_t^{b'}(x')$. She is indifferent if $u_t^b(x) = u_t^{b'}(x')$. The agent's *utility indifference (buy) value* for adding δ illiquid assets to her current position (x, b) is defined as the *compensating variation* $\pi_t^{b,x}(\delta)$ of her present wealth that leaves her utility unchanged, that is, as the solution to

$$u_t^{b+\delta}(x - \pi_t^{b,x}(\delta)) = u_t^b(x) \quad (1)$$

The *indifference sell value* is $-\pi_t^{b,x}(-\delta)$. In comparison, the *certainty equivalent* for adding δ to her position in illiquid assets is the *equivalent variation* $c_t^{b,x}(\delta)$ of the wealth that yields the same utility, that is, it is the solution to

$$u_t^b(x + c_t^{b,x}(\delta)) = u_t^{b+\delta}(x) \quad (2)$$

Equations (1), (2) have unique solutions if the functions $x \mapsto u_t^b(x)$ are strictly increasing and have the

same range for any b . The notion *compensating variation* is classical from the economic theory of demand by John Richard Hicks [7]. Alternatively, the terms *indifference value* (“price”) and *reservation price* have been used frequently in recent literature. We use the terms synonymously, but note that the classical terminology appears more accurate in reflecting the definition (1): in general, the compensating variation $\pi_t^{b,x}(\delta)$ is not a “price” for which δ illiquid assets can be traded in the market. Also, $\pi_t^{b,x}(\delta)$ is determined only at t in dependence of the position (x, b) prevailing and the variation δ occurring at the same time t ; it should not be interpreted prematurely as a “value at t ” that could be attributed at times before t to the payoff δB at T .

Next, we introduce the setup for a market model in which a family of utility functions u_t , $t \leq T$, is to be obtained. For simplicity, we consider a finite probability space $(\Omega, \mathcal{F}, P)$. For time $t \in \{0, \dots, T\}$ being discrete, the information flow is described by a filtration $(\mathcal{F}_t)_{0 \leq t \leq T}$ (see **Filtrations**), that is defined by refining partitions $\mathcal{A}_t$ of Ω and corresponds to a nonrecombining tree (see **Arrow–Debreu Prices**, Figure 1). The smallest nonempty events of $\mathcal{F}_t$ are called atoms $A \in \mathcal{A}_t$ of $\mathcal{F}_t$. Take $\mathcal{F}_0$ as trivial, $\mathcal{F}_T = \mathcal{F} = 2^\Omega$ as the set of all events, and all probabilities $P(\omega) > 0$, $\omega \in \Omega$ to be positive. Random variables X_t known at time t are denoted by $X_t \in L_{\mathcal{F}_t}$. They are determined by their values on each atom $A \in \mathcal{A}_t$, and can be identified with elements of a suitable $\mathbb{R}^N$. A process $(X_t)_{t \leq T}$ is adapted if $X_t \in L_{\mathcal{F}_t}$ for any t . Inequalities and properties stated for random variables and functions are meant to hold for all outcomes (coordinates). Conditional expectations with respect to $\mathcal{F}_t$ are denoted by $E_t[\cdot]$. For a family $\{X_t^a\} \subset L_{\mathcal{F}_t}(\mathbb{R})$, the random variable $\text{ess sup}_a X_t^a$ takes the value $\sup_a X_t^a(A)$ on atom $A \in \mathcal{A}_t$.

The price evolution of d liquidly traded risky assets is described by an $\mathbb{R}^d$ -valued adapted process $(S_t)_{0 \leq t \leq T}$. All prices are expressed in units of a further liquid riskless asset (unit of account) whose price is constant at 1. If, for example, the unit of account is the zero-coupon bond with maturity T , all prices are expressed in T -forward units. A trading strategy $(\vartheta_t)_{t \leq T} \in \Theta$ is described by the numbers of liquid assets $\vartheta_t \in L_{\mathcal{F}_{t-1}}$ to be held over any period $[t - 1, t)$. Its gains from t until T are $\int_t^T \vartheta \, dS := \sum_{k=t+1}^T \vartheta_k \Delta S_k$, with $\Delta S_k := S_k - S_{k-1}$. Any $X_T \in L_{\mathcal{F}_T}$ of the form $X_T = x + \int_t^T \vartheta \, dS$ represents a

wealth at time T that is attainable from $x \in L_{\mathcal{F}_t}$ at t by trading. Let $\mathcal{X}_T(x, t)$ denote the set of all such X_T , and set $\mathcal{X}_t(x) := \mathcal{X}_t(x, t-1)$. In addition to the liquid assets, there exist J illiquid assets delivering payoffs $B = (B^j)_{1 \leq j \leq J} \in L_{\mathcal{F}_T}(\mathbb{R}^J)$ at T . A quantity $b \in \mathbb{R}^J$ of illiquid assets provides at T the payoff $bB := \sum_j b^j B^j$. We assume that the market is free of arbitrage in the sense that the set $\mathcal{M}^e$ of equivalent probability measures Q under which S is a martingale (see **Martingales**) is nonempty. This is equivalent to assuming that all sets $\mathcal{D}_{s,t}$, $s < t \leq T$, of conditional state-price densities are nonempty. Technically, $\mathcal{D}_{s,t}$ is the set of strictly positive $D_{s,t} \in L_{\mathcal{F}_t}$ satisfying $E_s[D_{s,t}] = 1$ and $E_s[D_{s,t} \int_s^t \vartheta dS] = 0$ for all $\vartheta \in \Theta$. For brevity, let $\mathcal{D}_t := \mathcal{D}_{t-1,t}$. State-price densities are related to the likelihood density process $Z_t = E_t[dQ/dP]$ of a $Q \in \mathcal{M}^e$ by $D_{s,t} = Z_t/Z_s$.

Conditional Utility Functions and Dual Problems

Our agent's objective is to maximize her expected utility (3) of wealth at T for a direct utility function U , which is finite, differentiable, strictly increasing, and concave on all of $\mathbb{R}$, with $\lim_{x \rightarrow -\infty} U'(x) = \infty$ and $\lim_{x \rightarrow \infty} U'(x) = 0$. Holding position $(x, b) \in \mathbb{R} \times \mathbb{R}^J$ in liquid and illiquid assets at $t \leq T$, she maximizes

$$\begin{aligned} u_t^b(x) &:= \operatorname{ess\,sup}_{X_T \in \mathcal{X}_T(x,t)} E_t[U(X_T + bB)] \\ &= \operatorname{ess\,sup}_{X_T \in \mathcal{X}_T(x,t)} E_t[u_T^b(X_T)] \end{aligned} \quad (3)$$

We call $u_t^b(x)$ *u-regular* if (for all b, ω) the function $x \mapsto u_t^b(x)$ is strictly concave, increasing, and continuously differentiable on $\mathbb{R}$ with $\lim_{x \rightarrow -\infty} u_t^b(x) = +\infty$ and $\lim_{x \rightarrow +\infty} u_t^b(x) = 0$. For $t = T$, $u_T^b(x) = U(x + bB)$ satisfies the condition

$$\begin{aligned} u_t^b(x) &\text{ is } u\text{-regular, concave, and differentiable} \\ &\text{ in } (x, b), \text{ with } u_t^b(\mathbb{R}) = U(\mathbb{R}) \end{aligned} \quad (4)$$

with $U(\mathbb{R}) = \{U(x) : x \in \mathbb{R}\}$ denoting the range of U . The primal problems (3) are related, see A1–3, to the dual problems

$$v_t^b(y) := \operatorname{ess\,inf}_{D_{t,T} \in \mathcal{D}_{t,T}} E_t[V^b(yD_{t,T})], \quad y > 0 \quad (5)$$

for the conjugate function

$V^b(y) := \operatorname{ess\,sup}_x (U(x + bB) - xy)$ ($y > 0$, $b \in \mathbb{R}^J$), with $V^b(y) = V^0(y) + ybB$ and $v_T^b(y) = V^b(y)$. For later arguments, we assume the following:

(A1) $u_t^b(x)$ satisfies condition (4) for any t, b , and the value functions are conjugate:

$$v_t^b(y) = \operatorname{ess\,sup}_{x \in \mathbb{R}} (u_t^b(x) - xy), \quad y > 0 \quad (6)$$

$$u_t^b(x) = \operatorname{ess\,inf}_{y > 0} (v_t^b(y) + xy), \quad x \in \mathbb{R} \quad (7)$$

(A2) For any $t \leq T$ and $b, x, y \in L_{\mathcal{F}_{t-1}}$, there exist unique $\widehat{X}_t^b(x)$ and $\widehat{D}_t^b(y)$ that attain the single-period optima (8) and (9)

$$\begin{aligned} u_{t-1}^b(x) &= \operatorname{ess\,sup}_{X_t \in \mathcal{X}_t(x,t-1)} E_{t-1}[u_t^b(X_t)] \\ &= E_{t-1}[u_t^b(\widehat{X}_t^b(x))] \end{aligned} \quad (8)$$

$$\begin{aligned} v_{t-1}^b(y) &= \operatorname{ess\,inf}_{D_t \in \mathcal{D}_t} E_{t-1}[v_t^b(yD_t)] \\ &= E_{t-1}[v_t^b(y\widehat{D}_t^b(y))] \end{aligned} \quad (9)$$

and satisfy $u_t^b(\widehat{X}_t^b(x)) = y\widehat{D}_t^b(y)$ for x and y being related by $u_{t-1}^b(x) = y$.

(A3) For $t \leq T$, $b, x, y \in L_{\mathcal{F}_t}$, unique optima $\widehat{X}_T^b$ and $\widehat{D}_{t,T}^b$ for the multiperiod problems (3), (5) are attained, and can be constructed by dynamical programming $\widehat{X}_k^b = \widehat{X}_k^b(\widehat{X}_{k-1}^b)$ and $\widehat{D}_{t,k}^b = \widehat{D}_{t,k-1}^b \widehat{D}_k^b(y\widehat{D}_{t,k-1}^b)$ for $t < k \leq T$, with $\widehat{D}_k^b(\cdot)$ and $\widehat{X}_k^b(\cdot)$ from A2 and $\widehat{X}_t^b := x$ and $\widehat{D}_{t,t}^b := 1$. The optima satisfy $u_k^b(\widehat{X}_k^b) = y\widehat{D}_{t,k}^b$, $t < k \leq T$, for x and y being related by $u_t^b(x) = y$.

Ω being finite, A1–3 can be shown by convex duality; by arguments as in [21] follows inductively that A1–3 hold for $t-1$ on each atom, given they hold at t . Also see **Convex Duality** and **Second Fundamental Theorem of Asset Pricing**. Let us just mention here that, under regularity, the transforms (7) and (6) are inversions of each other, and $-v_t^b(y)$ is the inverse function of the marginal utility $u_t^b(x) = \frac{\partial}{\partial x} u_t^b(x)$.

Properties of Utility Indifference Values

Concavity (Convexity)

By concavity of U , indifference buy (sell) values $\pi_t^{b,x}(\delta)$ (respectively $-\pi_t^{b,x}(-\delta)$) are concave (convex) with respect to the quantity δ of illiquid assets

that they compensate for, that is,

$$\begin{aligned} & \lambda \pi_t^{b,x}(\delta^1) + (1 - \lambda) \pi_t^{b,x}(\delta^2) \\ & \leq \pi_t^{b,x}(\lambda \delta^1 + (1 - \lambda) \delta^2) \quad \text{for } \lambda \in [0, 1] \end{aligned} \quad (10)$$

Monotonicity

Monotonicity of U implies, on any atom $A \in \mathcal{A}_t$, that

$$\begin{aligned} 1_A \pi_t^{b,x}(\delta) & \leq 1_A \pi_t^{b,x}(\delta') \quad \text{if } 1_A \delta B \leq 1_A \delta' B \\ & \text{for } \delta, \delta' \in \mathbb{R}^J \end{aligned} \quad (11)$$

and that $1_A \pi_t^{b,x}(\delta) = 0$ holds if $1_A \delta B = 0$.

Dynamic consistency with no arbitrage

So far, we took the agent to trade optimally in liquid assets, while holding a fixed position b in illiquid assets. Now, suppose that she is ready to buy (or sell) at her compensating variations shares of illiquid assets in quantities as requested by another agent (he), dynamically over time. Let $\beta_t - \beta_0 \in L_{\mathcal{F}_{t-1}}(\mathbb{R}^J)$ denote the cumulative position in illiquid assets she has accepted until date $t - 1$, when she initially has held $\beta_0 := b \in \mathbb{R}^J$. At $t - 1 < T$, he chooses to sell $\Delta \beta_t \in L_{\mathcal{F}_{t-1}}(\mathbb{R}^J)$ illiquid assets. Given $\widehat{X}_{t-1}$ is the wealth in liquid assets she arrived with at $t - 1$, paying compensating variation changes her liquid wealth to $\widetilde{X}_{t-1} := \widehat{X}_{t-1} - \pi_{t-1}^{\beta_{t-1}, \widehat{X}_{t-1}}(\Delta \beta_t)$ such that the utility of her position stays equal $u_{t-1}^{\beta_{t-1}}(\widehat{X}_{t-1}) = u_{t-1}^{\beta_t}(\widetilde{X}_{t-1})$. Investing optimally for the next period according to A2 from her new position $(\widetilde{X}_{t-1}, \beta_t)$ (without knowing his future $(\beta_{t+k})_{k \geq 1}$), she arrives at t with liquid wealth

$$\begin{aligned} \widehat{X}_t &= \widehat{X}_t^{\beta_t}(\widetilde{X}_{t-1}) \\ &=: \widehat{X}_{t-1} - \pi_{t-1}^{\beta_{t-1}, \widehat{X}_{t-1}}(\Delta \beta_t) + \int_{t-1}^t \widehat{\vartheta} dS \end{aligned} \quad (12)$$

for an optimal strategy $\widehat{\vartheta}$ over $(t - 1, t]$. Given an initial wealth $\widehat{X}_0 = x \in \mathbb{R}$, the wealth process $\widehat{X}_t$ is determined by compensating variations and A2 such that $(u_t^{\beta_t}(\widehat{X}_t))$ is a martingale. Trading against indifference valuations but not following strategy $\widehat{\vartheta}$ would result in a suboptimal wealth process X_t for which the utility process $(u_t^{\beta_t}(X_t))$ is a supermartingale, therefore decreasing in the mean. By accepting to trade illiquid assets against her indifference values, she is not offering arbitrage

opportunities to her counterparty. Indeed, a strategy $\theta \in \Theta$ would offer arbitrage profits to him, jointly with (β_t) , if his gains

$$\begin{aligned} G_T &:= \int_0^T \theta dS \\ &+ \sum_{t=1}^T \pi_{t-1}^{\beta_{t-1}, \widehat{X}_{t-1}}(\Delta \beta_t) - (\beta_T - b)B \end{aligned} \quad (13)$$

satisfy $G_T \geq 0$ and $P[G_T > 0] > 0$. Unwinding her illiquid asset position at T , leaves her with final wealth

$$\begin{aligned} \widehat{X}_T &= x + \int_0^T \widehat{\vartheta} dS \\ &- \sum_{t=1}^T \pi_{t-1}^{\beta_{t-1}, \widehat{X}_{t-1}}(\Delta \beta_t) + \beta_T B \end{aligned} \quad (14)$$

Adding equation (13) to equation (14), would imply $E[u_T^b(x + \int_0^T \theta + \widehat{\vartheta} dS)] > E[u_T^b(\widehat{X}_T)] = u_0^b(x)$, contradicting definition (3).

Static no-arbitrage bounds

In particular, there is no arbitrage from buy(sell)-and-hold strategies in illiquid assets. For $x \in \mathbb{R}$, $b, \delta \in \mathbb{R}^J$ it thus holds

$$\begin{aligned} \pi_t^{b,x}(\delta) &\leq \text{ess sup}_{Q \in \mathcal{M}^e} E_t^Q[\delta B] \\ \text{and} \quad -\pi_t^{b,x}(-\delta) &\geq \text{ess inf}_{Q \in \mathcal{M}^e} E_t^Q[\delta B] \end{aligned} \quad (15)$$

For replicable payoffs $B^j = B_t^j + \int_t^T \vartheta^{B^j} dS$ with ϑ^{B^j} in Θ for all j , $B_t \in L_{\mathcal{F}_t}(\mathbb{R}^d)$ the indifference value $\pi_t^{b,x}(\delta)$ equals the replication cost (market price) δB_t .

Marginal indifference values

In general, $\pi_t^{b,x}(\delta)$ is nonlinear in δ . Since $\varepsilon \mapsto u_t^{b+\varepsilon\delta}(x - \pi_t^{b,x}(\varepsilon\delta))$ is constant, it holds

$$\left. \frac{\partial}{\partial \varepsilon} \pi_t^{b,x}(\varepsilon\delta) \right|_{\varepsilon=0} = \frac{\text{grad}_b u_t^b(x)}{u_t^b(x)} \delta \quad (16)$$

Hence, *marginal indifference values*, that is, compensating variations for infinitesimal changes of quantities, are linear in δ and given by the ratio of the gradient of $u_t^b(x)$ with respect to b and the marginal utility of wealth $u_t^b(x)$. The principle of valuation at ratios of marginal utilities is classical in economics, see for example [4]. Marginal indifference values can be computed from optimizers of the dual problem. They coincide with prices of an arbitrage-free dynamical price process in an enlarged market, where previously illiquid assets are tradable at “shadow” price processes, which are such that the utility maximizing agent is not trading those assets. To see this for $t = 0$, fix x, b . For y and $\widehat{D}_{0,T}^b$ from A3, let $R_k := E_k^Q[B]$, $k \leq T$, for $dQ := \widehat{D}_{0,T}^b dP$. Let $\bar{u}_0^b(x)$, $\bar{v}_0^b(y)$ be the primal and dual value functions (cf. equations (3),(5)) of the market $\bar{S} = (S, R)$ that is enlarged by the additional price process R . The set of state-price densities for the enlarged market is smaller, but includes the minimizer for equation (5). Hence $\bar{v}_0^b(\cdot) \geq v_0^b(\cdot)$ and $\bar{v}_0^b(y) = v_0^b(y)$, implying $\bar{v}_0^b(y) = v_0^b(y)$ and $\bar{u}_0^b(x) = u_0^b(x)$ by A1. Thus, the optimal strategy in the enlarged market is not trading the additional asset at the shadow price process (R_t) . The agent is, in particular, indifferent to infinitesimal initial variations of his position at shadow prices. Hence, R_0 must be given by the ratio in (16) of marginal utilities at $t = 0$. If the agent is taken to be representative for the whole market, holding a net supply of b illiquid assets, then (R_t) could be interpreted as a partial equilibrium price process.

Numeraire dependence

In general, utility indifference values depend on the utility functions and the numeraire (unit of account) with respect to which they are defined. But it is possible to choose state-dependent utility functions with respect to another numeraire such that indifference values (and optimal strategies) become numeraire invariant. Let (N_t) be the price process of a tradable numeraire, that is, $N_t = N_0 + \int_0^t \vartheta^N dS$ for $t \leq T$, $\vartheta^N \in \Theta$, with $N > 0$. Then indifference values coincide, that means $\pi_{t,N}^{b,x}(\delta) = \pi_t^{b,xN_t}(\delta)/N_t$ holds, if utilities and payoffs with respect to N satisfy the relations $u_{t,N}^b(x) = u_t^b(xN_t)$ (for $t = T$, hence all t) and $B_N := B/N_T$. Likewise, for numeraires $N, \tilde{N}$, the relations should be $u_{t,\tilde{N}}^b(x) = u_t^b(x\tilde{N}_t/N_t)$ and $B_{\tilde{N}} = B_N N_T / \tilde{N}_T$.

Partial hedging

Compensating variations can be associated with a utility-based hedging strategy, which, for an aggregate position (x, b) at $t = 0$, is defined as the strategy whose wealth process is $\widehat{X}_t^b(x) - \widehat{X}_t^0(x + c_0^{0,x}(b))$, for optimal wealth processes $\widehat{X}^b, \widehat{X}^0$ from A3 and $c_0^{0,x}(b)$ from equation (2). The risk that remains under partial hedging can be substantial, see the example below.

Case of Exponential Utility

Much of the literature on indifference pricing deals with exponential utility $U(x) = -\frac{1}{\alpha} \exp(-\alpha x)$ of constant absolute risk aversion $\alpha > 0$. Because U factorizes, the utility functions are of the form $u_t^b(x) = -\frac{1}{\alpha} e^{-\alpha(x+C_t^b)}$, $t \leq T$, for random variables $C_t^b \in L_{\mathcal{F}_t}$ not depending on x , with $C_T^b = bB$. Clearly, $\pi_t^{b,x}(\delta) = C_t^{b+\delta} - C_t^b$ does not depend on x , and the compensating variation (1) and the equivalent variation (2) coincide for exponential utility. From the dual value functions $v_t^b(y) = \frac{y}{\alpha} (\log y - 1 + \alpha C_t^b)$ from equation (5), one obtains a general formula

$$\begin{aligned} \pi_t^{0,x}(\delta) = & \operatorname{ess\,inf}_{D \in \mathcal{D}_{t,T}} \left\{ E_t[D(\delta B)] \right. \\ & \left. + \frac{1}{\alpha} (E_t[D \log D] - E_t[\widehat{D}_{t,T}^0 \log \widehat{D}_{t,T}^0]) \right\} \end{aligned} \quad (17)$$

for the indifference value $\pi_t^{b,x}(\delta) = \pi_t^{0,x}(\delta + b) - \pi_t^{0,x}(b)$, where $\widehat{D}_{t,T}^0$ is the minimizer of equation (5) for $b = 0$ that satisfies $E_t[\widehat{D}_{t,T}^0 \log \widehat{D}_{t,T}^0] = \operatorname{ess\,inf}_{D_{t,T}} E_t[D_{t,T} \log D_{t,T}]$. By equation (17), utility indifference sell values $\delta B \mapsto -\pi_t^{b,x}(-\delta)$ are monotonic in α , and satisfy the properties of convexity, translation invariance, and monotonicity, that constitute a *convex risk measure* (see **Convex Risk Measures**).

Under particular model assumptions, indifference values $\pi_t^{0,x}(\delta)$ can be computed by a backward induction scheme

$$\begin{aligned} & -\pi_{t-1}^{0,x}(\delta) \\ & = E_{t-1}^{Q^0} \left[\frac{1}{\alpha} \log E_{\mathcal{G}_t} [\exp(-\alpha \pi_t^{0,x}(\delta))] \right] \end{aligned} \quad (18)$$

starting from $\pi_T^{0,x}(\delta) = \delta B$. Roughly speaking, the assumptions needed comprise certain independence conditions plus semicompleteness of the market at each period. The scheme (18) has intuitive appeal, in showing that the indifference valuation is computed here by intertwining two well-known valuation methods: First, one takes an exponential certainty equivalent with respect to nontradable risk at the inner expectation (with $\mathcal{F}_{t-1} \subset \mathcal{G}_t \subset \mathcal{F}_t$); after that one takes a risk-neutral expectation of this certainty equivalent at the outer expectation (under the **minimal entropy martingale measure**), where, $\mathcal{G}_t$ -risk is taken as replicable from $t - 1$. See [1, 18] for precise technical assumptions and examples.

Example in Continuous Time

For an instructive example, consider a (nonfinite) filtered probability space $(\Omega, (\mathcal{F}_t)_{t \leq T}, P)$ with Brownian motions (W_t) and $(\tilde{W}_t) = (\rho W_t + \sqrt{1 - \rho^2} W_t^\perp)$ correlated by $\rho \in [-1, 1]$. The price process of a single risky asset is $S_t = S_0 + \sigma(W_t + \lambda t)$ with $\lambda, S_0 \in \mathbb{R}$, $\sigma > 0$, as in the model by **Louis Bachelier**. The illiquid asset's payout $B := Y_T$ for $Y_t = Y_0 + \tilde{\sigma}(\tilde{W}_t + \lambda t)$ can be interpreted as position in a nontraded but correlated asset. Trading strategies $\vartheta \in \Theta$ are taken to be adapted and bounded. The maximal expected exponential utilities

$$u_t^b(x) = -\frac{1}{\alpha} \exp \left(-\alpha \left(x + \frac{1}{2} \frac{\lambda^2}{\alpha} (T - t) + \pi_t^{0,x}(b) \right) \right), \quad x, b \in \mathbb{R} \quad (19)$$

are then attained by the optimal strategies $\hat{\vartheta}^b = \frac{\lambda}{\sigma\alpha} - b \frac{\sigma}{\sigma} \rho$, with

$$\pi_t^{0,x}(b) = bY_t + b\tilde{\sigma}(\tilde{\lambda} - \lambda\rho)(T - t) - \frac{1}{2} \alpha b^2 \tilde{\sigma}^2 (1 - \rho^2)(T - t) \quad (20)$$

Indifference values $\pi_t^{b,x}(\delta) = \pi_t^{0,x}(b + \delta) - \pi_t^{0,x}(b)$ for exponential utility do not depend on wealth x . Optimality of equation (19) and $\hat{\vartheta}^b$ follows by noting that $u_t^b(\int_0^t \vartheta \, dS)$ is a martingale for $\vartheta = \hat{\vartheta}^b$ and a supermartingale for any other $\vartheta \in \Theta$. Clearly, indifference buy (sell) values $\pi_t^{b,x}(\delta)$ (respectively $-\pi_t^{b,x}(-\delta)$) are decreasing (increasing) in the risk aversion α . They are linear in the quantity δ only if correlation is perfect ($|\rho| = 1$). Then, they coincide

with the replication cost of δB . Marginal utility indifference values are given by

$$\frac{\partial}{\partial \delta} \pi_t^{b,x}(\delta) = Y_t + \tilde{\sigma}(\tilde{\lambda} - \lambda\rho)(T - t) - \alpha(b + \delta)\tilde{\sigma}^2(1 - \rho^2)(T - t) \quad (21)$$

Under the (minimal entropy) martingale measure $dQ^0 = \exp(-\lambda W_T - \lambda^2 T/2) dP$ we have $S_t = S_0 + \sigma W_t^0$ for independent Q^0 -Brownian motions (W_t^0) and $(W_t^\perp)$. Indifference values can be expressed by

$$-\pi_t^{0,x}(b) = \frac{1}{\alpha(1 - \rho^2)} \log E_t^{Q^0} [\exp(\alpha(1 - \rho^2)(-bB))] \quad (22)$$

Formulas like equation (22) have also been obtained for different models, including the case where the price processes of the risky and the nontraded asset (underlying of B) are given by correlated geometric Brownian motions, see [6, 16].

To discuss the possible size of the partial hedging error, let $S_0, Y_0, \lambda, \tilde{\lambda}$ be zero, $T = 1$, and $\sigma = \tilde{\sigma}$. We assume that the agent has accepted initially an illiquid position b at her indifference valuation. Her utility-based partial hedging strategy when holding b illiquid assets is $\hat{\vartheta}^b - \vartheta^0 = -b\rho$. Her hedging error $H = -\pi_0^{0,x}(b) + bB - b\sigma\rho W_1^0 = \frac{1}{2}\alpha b^2 \sigma^2 (1 - \rho^2) + b\sigma\sqrt{1 - \rho^2} W_1^\perp$ is normally distributed. Its standard deviation accounts for $\sqrt{1 - \rho^2} \times 100\%$ of that for the unhedged payoff $bB = bY_T$. For correlation $\rho = 80\%$, for example, the error size is still substantial at a ratio of 60%. Even for $\rho = 99\%$, it is still above 14%. To be compensated for the remaining risk in terms of her expected utility, the agent requires $-\pi_0^{0,x}(b) = \frac{1}{2}\alpha b^2 \sigma^2 (1 - \rho^2)$ at $t = 0$. Her compensating variation of wealth is proportional to the variance of H and to her risk aversion α .

Further Reading

To value options under transaction costs, indifference valuation was applied in [8]. The method is not limited to European payoffs. For payoffs with optimal exercise features, see [10, 17]. Indifference values for payoff streams could be defined by equation (1) for utilities that reflect preferences on future payment streams, like in [22]. For results on nonexponential

utilities, see [5, 12, 14]. For performance of utility-based hedging strategies, see [15]. Besides dynamical programming and convex duality, solutions have been obtained by backward stochastic differential equations (see **Backward Stochastic Differential Equations**) [10, 13, 20], also for non-convex closed constraints [9] and jumps [2]. For asymptotic results on valuation and hedging for small volumes, see [2, 5, 12, 13]. A Paretian equilibrium formulation for indifference pricing has been presented in [11]. Being nonlinear, indifference values can reflect diversification or accumulation of risk for applications areas like real options or insurance, see [6, 14, 19, 22]; but modeling and computation are more demanding, since a portfolio of assets cannot be valued by parts in general. Instead, each component is to be judged by its contribution to the overall portfolio. More comprehensive references are given in [1, 3, 4, 6, 16].

References

- [1] Becherer, D. (2003). Rational hedging and valuation of integrated risks under constant absolute risk aversion, *Insurance: Mathematics and Economics* **33**, 1–28.
- [2] Becherer, D. (2006). Bounded solutions to backward SDEs with jumps for utility optimization and indifference hedging, *Annals of Applied Probability* **16**, 2027–2054.
- [3] Davis, M.H.A. (2006). Optimal hedging with basis risk, in *From Stochastic Calculus to Mathematical Finance*, Y. Kabanov, R. Liptser & J. Stoyanov, eds, Springer, Berlin, pp. 169–188.
- [4] Foldes, L. (2000). Valuation and martingale properties of shadow prices: an exposition, *Journal of Economic Dynamics and Control* **24**, 1641–1701.
- [5] Henderson, V. (2002). Valuation of claims on non-traded assets using utility maximization, *Mathematical Finance* **12**, 351–373.
- [6] Henderson, V. & Hobson, D. (2008). Utility indifference pricing—an overview, in *Indifference Pricing*, R. Carmona, ed, Princeton University Press, pp. 44–74.
- [7] Hicks, J.R. (1956). *A Revision of Demand Theory*, Oxford University Press, Oxford.
- [8] Hodges, S.D. & Neuberger, A. (1989). Optimal replication of contingent claims under transaction costs, *Review of Futures Markets* **8**, 222–239.
- [9] Hu, Y., Imkeller, P. & Müller, M. (2005). Utility maximization in incomplete markets, *Annals of Applied Probability* **15**, 1691–1712.
- [10] Kobylanski, M., Lepeltier, J., Quenez, M. & Torres, S. (2002). Reflected backward SDE with super-linear quadratic coefficient, *Probability and Mathematical Statistics* **22**, 51–83.
- [11] Kramkov, D. & Bank, P. (2007). *A model for large investor, where she trades at utility indifference prices of market makers*, ICMS, Edinburgh, Presentation, www.icms.org.uk/downloads/quantfin/Kramkov.pdf
- [12] Kramkov, D. & Sirbu, M. (2007). Asymptotic analysis of utility-based hedging strategies for small number of contingent claims, *Stochastic Processes and Applications* **117**, 1606–1620.
- [13] Mania, M. & Schweizer, M. (2005). Dynamic exponential utility indifference valuation, *Annals of Applied Probability* **15**, 2113–2143.
- [14] Møller, T. (2003). Indifference pricing of insurance contracts: applications, *Insurance: Mathematics and Economics* **32**, 295–315.
- [15] Monoyios, M. (2004). Performance of utility-based strategies for hedging basis risk, *Quantitative Finance* **4**, 245–255.
- [16] Musiela, M. & Zariphopoulou, T. (2004). An example of indifference pricing under exponential preferences, *Finance and Stochastics* **8**, 229–239.
- [17] Musiela, M. & Zariphopoulou, T. (2004). Indifference prices of early exercise claims, in *Mathematics of Finance*, G. Yin & Q. Zhang, eds, *Contemporary Mathematics*, AMS, Vol. 351, pp. 259–273.
- [18] Musiela, M. & Zariphopoulou, T. (2004). A valuation algorithm for indifference prices in incomplete markets, *Finance and Stochastics* **8**, 399–414.
- [19] Porchet, A., Touzi, N. & Warin, X. (2008). Valuation of power plants by utility indifference and numerical computation, *Mathematical Methods of Operations Research* [Online], DOI: 10.007/s00186-008-0231-z.
- [20] Rouge, R. & El Karoui, N. (2000). Pricing via utility maximization and entropy, *Mathematical Finance* **10**, 259–276.
- [21] Schachermayer, W. (2002). Optimal investment in incomplete financial markets, in *Mathematical Finance: Bachelier Congress 2000*, H. Geman, D. Madan & S.R. Pliska, eds, Springer, Berlin, pp. 427–462.
- [22] Smith, J.E. & Nau, R.F. (1995). Valuing risky projects: option pricing theory and decision analysis, *Management Science* **41**, 795–816.

Related Articles

Complete Markets; Expected Utility Maximization; Duality Methods; Good-deal Bounds; Hedging; Minimal Entropy Martingale Measure; Utility Theory; Historical Perspectives; Utility Function.

DIRK BECHERER

Superhedging

Pricing and hedging of contingent claims are the two main problems of mathematical finance. They both have a clear and transparent solution when the underlying market model is complete, that is, for each contingent claim with promised payoff H there exists a self-financing admissible trading strategy whose wealth at maturity equals H (see **Complete Markets**). Such a strategy is called the *hedging strategy* of the contingent claim H . The smallest initial wealth that allows to reach H at maturity via admissible trading is called the *hedging price* of H .

Under a suitable no-arbitrage assumption (see **Fundamental Theorem of Asset Pricing**), the second fundamental theorem of asset pricing (see **Second Fundamental Theorem of Asset Pricing**) states that replicability of every contingent claim is equivalent to the uniqueness of the equivalent martingale measure $\mathbb{Q}$ (see **Equivalent Martingale Measures**).

It turns out that in a complete market (see **Complete Markets**), the hedging price at time $t = 0$ of a contingent claim H , denoted by $p(H)$, coincides with the expectation of discounted H under the unique equivalent martingale measure $\mathbb{Q}$, that is, $p(H) = \mathbb{E}_{\mathbb{Q}}[D_T H]$ where D_T is a discounting factor over $[0, T]$.

If the market model is incomplete, there exist contingent claims that are not perfectly replicable via admissible trading strategies. In other words, in such financial models, contingent claims are not redundant assets. Therefore, since perfect replicability cannot be always achieved, this requirement has to be relaxed. One way of doing this consists in introducing the concept of *superhedging*.

Given a contingent claim H with maturity $T > 0$, a superhedging strategy for H is an admissible trading strategy such that its terminal wealth V_T super-replicates H , that is, $V_T \geq H$. The superhedging price of H is the smallest initial endowment that allows an investor to super-replicate H at maturity; in other words, it is the initial value V_0 of the superhedging strategy of H .

Superhedging was introduced and investigated first by El Karoui and Quenez [13, 14] in a continuous-time setting where the risky assets follow a multidimensional diffusion process. Independently, Naik and Uppal [25] studied the same problem in a discrete-time model with finite set of scenarios and

noticed that, in the presence of leverage constraints, superhedging may be cheaper than perfect hedging. The same phenomenon has been observed by Benais *et al.* [1] in the presence of transaction costs.

The characterization of superhedging strategies and prices is the object of a family of results called *superhedging theorems*.

Superhedging Theorems

A large literature has been devoted to characterizing the set of all initial endowments that allows to superhedge a contingent claim H as a first crucial step to compute the superhedging price, the infimum of that set. In this article, we focus essentially on continuous-time hedging of European options, that is, with a fixed exercise time T , and distinguish between two cases: frictionless incomplete markets and markets with frictions. For superhedging in discrete-time models and for American options, the interested reader could see respectively Föllmer and Schied's book [17] and **American Options**.

Frictionless Incomplete Markets

To facilitate the discussion, let us fix the notation first. We consider a market model composed of $d \geq 1$ risky assets whose discounted price dynamics is described by a càdlàg and locally bounded semimartingale $S = (S_t)_{t \in [0, T]}$, where $T > 0$ is a given finite time horizon. S is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and adapted to a filtration $(\mathcal{F}_t)_{t \in [0, T]}$ with $\mathcal{F}_t \subset \mathcal{F}$ for all $t \leq T$ satisfying usual conditions. Notice that prices S are already discounted; this is equivalent to assuming that the spot interest rate $r = 0$. This model is, in general, incomplete, that is, it may admit infinitely many equivalent martingale measures (see **Equivalent Martingale Measures**).

Let H be a positive $\mathcal{F}_T$ -measurable random variable, modeling the final payoff of a given contingent claim, for example, $H = (S_T - K)^+$, a European call option written on S , with maturity T and strike price $K > 0$.

An admissible trading strategy is a couple (x, θ) where $x \in \mathbb{R}$ is an initial endowment and $\theta = (\theta_t)_{t \in [0, T]}$ a predictable S -integrable process, such that the corresponding wealth $V_t^{x, \theta} = x + \int_0^t \theta_u dS_u \geq -a$ for every $t \in [0, T]$ and for some threshold $a > 0$. We denote $\mathcal{A}$ as the set of all admissible strategies.

Definition 1 Let $H \geq 0$ be a given contingent claim. $(x, \theta) \in \mathcal{A}$ is a superhedging strategy for H if $V_T^{x, \theta} \geq H$ $\mathbb{P}$ -a.s. (almost surely). Moreover, the superhedging price $\bar{p}(H)$ of H is given by

$$\bar{p}(H) = \inf \left\{ x \in \mathbb{R} : \exists (x, \theta) \in \mathcal{A}, V_T^{x, \theta} \geq H \text{ a.s.} \right\} \quad (1)$$

The fundamental result in the literature on superhedging is the dual characterization of the set $\mathcal{D}^H$ of all initial endowments $x \in \mathbb{R}$ leading to superhedge H . In an incomplete frictionless market, the relevant dual variables are the densities of all equivalent martingale measures $d\mathbb{Q}/d\mathbb{P}$. We denote $\mathcal{M}^e$ as the set of all equivalent (local) martingale measures for S . In this setting, the superhedging theorem states that

$$\mathcal{D}^H = \{x \in \mathbb{R} : \mathbb{E}_{\mathbb{Q}}[H] \leq x, \forall \mathbb{Q} \in \mathcal{M}^e\} \quad (2)$$

An important consequence of equation (2) is that the superhedging price $\bar{p}(H)$ satisfies

$$\bar{p}(H) = \sup_{\mathbb{Q} \in \mathcal{M}^e} \mathbb{E}_{\mathbb{Q}}[H] \quad (3)$$

While an advantage of superhedging is that it is preference free, from the previous characterization of $\bar{p}(H)$ as the biggest expectation $\mathbb{E}_{\mathbb{Q}}[H]$ over all equivalent martingale measures, it becomes apparent that pursuing a superhedging strategy can be too expensive, depending on the financial model and on the constraints on portfolios. This is the main disadvantage of such a criterion, which is, nonetheless, of great interest as a benchmark. Moreover, for an agent with a large risk aversion and under transaction costs (see the section Markets with Frictions), the reservation price approaches the superhedging price, as established in [2].

El Karoui and Quenez [13, 14] first proved the superhedging theorem in an Itô's diffusion setting and Delbaen and Schachermayer [10, 11] generalized it to, respectively, a locally bounded and unbounded semimartingale model, using a Hahn–Banach separation argument.

The superhedging theorem can be extended in order to characterize the *dynamics* of the minimal superhedging portfolio of a contingent claim H , that is, the cheapest at any time t of all superhedging portfolios of H with the same initial

wealth. This extension is a consequence of the so-called *optional decomposition of supermartingales*. The optional decomposition was first proved in [13, 14] for diffusions and then extended to general semimartingales by Kramkov [24], Föllmer and Kabanov [15], and Delbaen and Schachermayer [12]. This is a very deep result of the general theory of stochastic processes and roughly states that any càdlàg-positive $\mathbb{Q}$ -supermartingale X , for any $\mathbb{Q} \in \mathcal{M}^e$, can be decomposed as follows:

$$X_t = X_0 + \int_0^t \theta_u dS_u - C_t, \quad t \in [0, T] \quad (4)$$

where θ is a predictable, S -integrable process and C an increasing optional process, to be interpreted as a cumulative consumption process. What is remarkable is that the local martingale part can be represented as a stochastic integral with respect to S so that it is a local martingale under any equivalent martingale measure $\mathbb{Q}$. In this sense, decomposition (4) is universal. The price to pay is that the increasing process C is, in general, not predictable as in the Doob–Meyer decomposition (see **Doob–Meyer Decomposition**) but only optional. The process C has the economic interpretation of cumulative consumption.

The decomposition (4) implies that the wealth dynamics of the minimal superhedging portfolio for a contingent claim H is given by

$$V_t = \text{ess sup}_{\mathbb{Q} \in \mathcal{M}^e} \mathbb{E}_{\mathbb{Q}}[H | \mathcal{F}_t], \quad t \in [0, T] \quad (5)$$

An analogous result holds for American contingent claims too (see [13–15, 24] for details at increasing levels of generality).

Finally, in the more specific setting of stochastic volatility models, Cvitanic *et al.* [8] compute the superhedging strategy and price for a contingent claim $H = g(S_T)$, yielding that the former is a buy-and-hold strategy and so the latter is just S_0 . The same study is carried over under portfolio constraints.

Markets with Frictions

In the previous section, we made the implicit assumption that investors can trade in continuous time and without frictions. This is clearly a strong idealization of the real world; that is why during the last 15 years much effort has been devoted to the superhedging approach under various types of trading constraints.

Transaction Costs. Financial models with proportional transaction costs were studied first by Jouini and Kallal [19] and then generalized in a series of papers by Kabanov and his coauthors [20–22].

For the reader's convenience, we briefly introduce the model, following the bid–ask matrix formalism introduced by Schachermayer [27], which is only one of many equivalent convenient ways of describing it (see, e.g., [22] and **Transaction Costs** for more details).

We consider an economy with $d \geq 1$ risky assets (e.g., foreign currencies); $\pi_t^{ij}(\omega)$ denotes the number of physical units of asset i that can be exchanged with 1 unit of asset j at time $t \in [0, T]$. All of them are assumed to be adapted to some filtration and càdlàg. An important role is played by the so-called *solvency region* $K_t(\omega)$, the cone generated by the unit vectors e^i and $\pi_t^{ij}e^i - e^j$ for $1 \leq i, j \leq d$. Elements of $K_t(\omega)$ are all the positions that can be liquidated into a portfolio with a nonnegative quantity of each currency. We denote $K_t^*(\omega)$ as the positive polar of $K_t(\omega)$.

A self-financing portfolio process is modeled by a d -dimensional finite variation process $V = (V_t)_{t \in [0, T]}$ such that each infinitesimal change $dV_t(\omega)$ lies in $-K_t(\omega)$, that is, a portfolio change at time t has to be done according to the trading terms described by the solvency cone K_t .

In this setting, so-called *strictly consistent price systems* play the same role as the equivalent martingale measures. A strictly consistent price system Z is a positive non-null d -dimensional martingale such that each $Z_t(\omega)$ belongs to the relative interior of $K_t^*(\omega)$ almost surely for all $t \in [0, T]$. We denote $\mathcal{Z}^s$ as the set of all strictly consistent price systems. A standard assumption is that there exists at least one of such Z 's, that is, $\mathcal{Z}^s \neq \emptyset$, which is equivalent to some kind of no-arbitrage condition (see **Transaction Costs** for details).

Let $H = (H^1, \dots, H^d)$ be a d -dimensional contingent claim such that $H + a\mathbf{1} \in K_T$ for some $a \in \mathbb{R}$. We say that an admissible^a portfolio V superhedges H if $V_T - H \in K_T$. Consider the set $\mathcal{D}^H$ of all initial endowment $x \in \mathbb{R}^d$ such that there exists an admissible portfolio V , $V_0 = x$, that superhedges H .

In this model, the superhedging theorem states that

$$\mathcal{D}^H = \{x \in \mathbb{R}^d : \mathbb{E}[Z_T H] \leq \langle x, Z_0 \rangle, \forall Z \in \mathcal{Z}^s\} \quad (6)$$

where $\langle \cdot, \cdot \rangle$ denotes the usual scalar product in $\mathbb{R}^d$. This theorem has been proven with increasing degree of generality by Cvitanić and Karatzas [7], Kabanov [20], and Kabanov and Last [21] for continuous bid–ask processes $(\pi_t)_{t \in [0, T]}$ and constant proportional transaction costs, by Kabanov and Stricker [22] under slightly more general assumptions and finally, motivated by a counterexample constructed by Rásonyi [26], Campi and Schachermayer [5] extend it to discontinuous π .

Explicit computations of the superhedging price have been performed in [3, 9, 18] for a European-type contingent claim $H = g(S_T)$, where S_T is the price at time T of a given asset in terms of some fixed numéraire. Under different assumptions, the superhedging strategy is a buy-and-hold one, so that the corresponding superhedging price is the price at time $t = 0$ of the underlying S_0 .

Finally, duality methods for American options under proportional transaction costs are briefly treated in **Transaction Costs**.

Other Types of Market Frictions. Superhedging has also been studied under other types of constraints on, for example, shortselling and/or borrowing (see, e.g., Cvitanić and Karatzas' paper [6] and Karatzas and Shreve's book [23], Chapter 5, for more details). Very often, an agent willing to superhedge a contingent claim H has to choose a strategy fulfilling a given set of constraints. Let us denote $\mathcal{A}_c$ as the class of constrained trading strategies. In this case, the constrained superhedging price $\bar{p}_c(H)$ is given by

$$\bar{p}_c(H) = \inf \left\{ x \in \mathbb{R}^d : \exists (x, \theta) \in \mathcal{A}_c, V_T^{x, \theta} \geq H \right\} \quad (7)$$

Cvitanić and Karatzas [6] gave the first dual characterization of $\bar{p}_c(H)$ in a diffusions setting, which was further generalized to general semimartingales by Föllmer and Kramkov [16] via a constrained version of the optional decomposition theorem, whose original version we already discussed at the end of the section Frictionless Incomplete Markets.

We conclude by mentioning a recent series of papers by Broadie *et al.* [4] and by Soner and Touzi [28, 29] on superhedging under *gamma constraints*, where an agent is allowed to hedge H , having at the same time a control on the gamma of his or her portfolios.

End Notes

^aWe remark *en passant* that the notion of admissibility in the presence of transaction costs, that we do not give here, is a subtle one. The interested reader could look at [5] for a short discussion.

References

- [1] Benais, B., Lesne, J.P., Pagès, H. & Scheinkman, J. (1992). Derivative asset pricing with transaction costs, *Mathematical Finance* **2**, 63–86.
- [2] Bouchard, B., Kabanov, Yu.M. & Touzi, N. (2001). Option pricing by large risk aversion utility under transaction costs, *Decisions in Economics and Finance* **24**, 127–136.
- [3] Bouchard, B. & Touzi, N. (2000). Explicit solution of the multivariate super-replication problem under transaction costs, *Annals of Applied Probability* **10**, 685–708.
- [4] Broadie, M., Cvitanic, J. & Soner, H.M. (1998). Optimal replication of contingent claims under portfolio constraints, *The Review of Financial Studies* **11**, 59–79.
- [5] Campi, L. & Schachermayer, W. (2006). A super-replication theorem in Kabanov's model of transaction costs, *Finance and Stochastics* **10**(4), 579–596.
- [6] Cvitanic, J. & Karatzas, I. (1993). Hedging contingent claims with constrained portfolios, *The Annals of Applied Probability* **3**(3), 652–681.
- [7] Cvitanic, J. & Karatzas, I. (1996). Hedging and portfolio optimization under transaction costs: a martingale approach, *Mathematical Finance* **6**(2), 133–165.
- [8] Cvitanic, J., Pham, H. & Touzi, N. (1999). Super-replication in stochastic volatility models under portfolio constraints, *Journal of Applied Probability* **36**(2), 523–545.
- [9] Cvitanic, J., Pham, H. & Touzi, N. (1999). A closed form solution to the problem of super-replication under transaction costs, *Finance and Stochastics* **3**, 35–54.
- [10] Delbaen, F. & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing, *Mathematische Annalen* **300**, 463–520.
- [11] Delbaen, F. & Schachermayer, W. (1998). The fundamental theorem of asset pricing for unbounded stochastic processes, *Mathematische Annalen* **312**, 215–250.
- [12] Delbaen, F. & Schachermayer, W. (1999). A compactness principle for bounded sequences of martingales with applications, *Proceedings of the Seminar of Stochastic Analysis, Random Fields and Applications, Progress in Probability* **45**, 137–173.
- [13] El Karoui, N. & Quenez, M.-C. (1991). Programmation dynamique et évaluation des actifs contingents en marché incomplet. (French) [Dynamic programming and pricing of contingent claims in an incomplete market], *Comptes Rendus de l'Académie des Sciences Série Mathématiques* **313**(12), 851–854.
- [14] El Karoui, N. & Quenez, M.-C. (1995). Dynamic programming and pricing of contingent claims in an incomplete market, *SIAM Journal of Control and Optimization* **33**(1), 27–66.
- [15] Föllmer, H. & Kabanov, Yu.M. (1998). Optional decomposition and Lagrange multipliers, *Finance and Stochastics* **2**(1), 69–81.
- [16] Föllmer, H. & Kramkov, D. (1997). Optional decompositions under constraints, *Probability Theory and Related Fields* **109**, 1–25.
- [17] Föllmer, H. & Schied, A. (2004). *Stochastic Finance: An Introduction in Discrete Time*, 2nd Edition, de Gruyter Studies in Mathematics, Berlin, P. 27.
- [18] Guasoni, P., Rásonyi, M. & Schachermayer, W. (2007). Consistent price systems and face-lifting under transaction costs, *Annals of Applied Probability* **18**(2), 491–520.
- [19] Jouini, E. & Kallal, H. (1995). Martingales and arbitrage in securities markets with transaction costs, *Journal of Economic Theory* **66**, 178–197.
- [20] Kabanov, Yu.M. (1999). Hedging and liquidation under transaction costs in currency markets, *Finance and Stochastics* **3**(2), 237–248.
- [21] Kabanov, Yu.M. & Last, G. (2002). Hedging under transaction costs in currency markets: a continuous-time model, *Mathematical Finance* **12**(1), 63–70.
- [22] Kabanov, Yu. & Stricker, Ch. (2002). Hedging of contingent claims under transaction costs, in *Advances in Finance and Stochastics. Essays in Honour of Dieter Sondermann*, K. Sandmann & Ph. Schonbucher, eds, Springer, Berlin, Heidelberg, New York.
- [23] Karatzas, I. & Shreve, S. (1998). *Methods of Mathematical Finance*, Springer.
- [24] Kramkov, D. (1996). Optional decomposition of supermartingales and hedging contingent claims in incomplete security markets, *Probability Theory and Related Fields* **105**, 459–479.
- [25] Naik, V. & Uppal, R. (1994). Leverage constraints and the optimal hedging of stock and bond options, *Journal of Financial and Quantitative Analysis* **29**(2), 199–222.
- [26] Rásonyi, M. (2003). *A Remark on the Superhedging Theorem Under Transaction Costs. Séminaires de Probabilités XXXVII*, Lecture Notes in Mathematics, 1832, Springer, pp. 394–398.
- [27] Schachermayer, W. (2004). The fundamental theorem of asset pricing under proportional transaction costs in finite discrete time, *Mathematical Finance* **14**(1), 19–48.
- [28] Soner, H.M. & Touzi, N. (2000). Super-replication under gamma constraints, *SIAM Journal of Control and Optimization* **39**(1), 73–96.
- [29] Soner, M. & Touzi, N. (2007). Hedging under gamma constraints by optimal stopping and face-lifting, *Mathematical Finance* **17**(1), 59–80.

LUCIANO CAMPI

Free Lunch

In the process of building realistic mathematical models of financial markets, absence of opportunities for riskless profit is considered to be a *minimal* normative assumption in order for the market to be in equilibrium state. The reason is quite obvious. If opportunities for riskless profit were present in the market, every economic agent would try to reap them. Prices would then instantaneously move in response to an imbalance between supply and demand. This sudden price movement would continue as long as opportunities for riskless profit are still present in the market. Therefore, in market equilibrium, no such opportunities should be possible.

The aforementioned simple and very natural idea has proved very fruitful and has lead to great mathematical as well as economical insight in the theory of quantitative finance. A rigorous formulation of the exact definition of “absence of opportunities for riskless profit” turned out to be a highly nontrivial fact that troubled mathematicians and economists for at least two decades.^a As the road unfolded, the valuable input of the theory of stochastic analysis in financial theory was obvious; in the other direction, the development of the theory of stochastic processes benefited immensely from problems that emerged purely from these financial considerations.

Since the late 1970s, there has been a notion that there is a deep connection between the absence of opportunities for riskless profit and the existence of a risk-neutral measure,^b that is, a probability that is equivalent to the original one under which the discounted asset price processes have some kind of martingale property. Existence of such measures are of major practical importance, since they open the road to pricing illiquid assets or contingent claims in the market (see **Risk-neutral Pricing**). The result of the above notion has been called the *fundamental theorem of asset pricing (FTAP)*; for a detailed account, see **Fundamental Theorem of Asset Pricing**.

The easiest and most classical way to formulate the notion of riskless profit is *via* the so-called *arbitrage strategy* (see **Arbitrage Strategy**). An arbitrage is a combination of positions in the traded assets that requires zero initial capital and results in non-negative outcome with a strictly positive probability of the wealth being strictly positive at a fixed time

point in the future (after liquidation has taken place). Naturally, the previous formulation of an arbitrage presupposes that a probabilistic model for the random movement of liquid asset prices has been set up. In [5], a discrete state space, multiperiod discrete-time financial market was considered. For this model, the authors showed the equivalence between the economical “no arbitrage” (NA) condition and the mathematical stipulation of existence of an equivalent probability that makes the discounted asset price processes martingales.

Crucial in the proof of the result in [5] was the separating hyperplane theorem in finite-dimensional Euclidean spaces. One of the convex sets to be separated is the class of all terminal outcomes resulting from trading and possible consumption starting from zero capital; the other is the positive orthant. The NA condition is basically the statement that the intersection of these two convex sets consists of only the zero vector.

After the publication of [5], a saga of papers followed that were aimed, one way or another, at strengthening the conclusion by considering more complicated market models. It quickly became obvious that the previous NA condition is no longer sufficient to imply the existence of a risk-neutral measure; it is too weak. In infinite-dimensional spaces, separation of hyperplanes, made possible by means of the geometric version of the Hahn–Banach theorem, requires the closedness of the set C of all terminal outcomes resulting from trading and possible consumption starting from zero capital. The simple NA condition does not imply this, in general. This has lead Kreps [7] to define a *free lunch* as a generalized, asymptotic form of an arbitrage.

Essentially, a free lunch is a possibly infinite-valued random variable f with $\mathbb{P}[f \geq 0] = 1$ and $\mathbb{P}[f > 0] > 0$ that belongs to the *closure* of C . Once an appropriate topology is defined on $\mathbf{L}^0$, the space of all random variables, in order for the last closure (call it $\tilde{C}$) to make sense, the “no-free-lunch” (NFL) condition states that^c $\tilde{C} \cap \mathbf{L}_+^0 = \{0\}$. Kreps [7] used this idea with a very weak topology on locally convex spaces and showed the existence of a *separating measure*.^d However, apart from trivial cases, this topology does not stem from a metric, which means that closedness cannot be described in terms of convergence of sequences. This makes the definition of a free lunch quite nonintuitive.

After [7], there were lots of attempts to introduce a condition closely related to NFL that would be more economically plausible, albeit still equivalent to NFL, and would prove equivalent to the existence of a risk-neutral measure. In general finite-horizon, discrete-time markets, it was shown in^e [1] that the plain NA condition is equivalent to NFL. This seemed to suggest the possibility of a nice counterpart of the NFL condition for more complicated models. Delbaen [2] treated the case of continuous time, bounded, and continuous asset prices and used a neat condition, equivalent to NFL, called^f *no free lunch with bounded risk* (NFLBR) that can be stated in terms of sequence convergence. Essentially, the NFLBR condition precludes asymptotic arbitrage at some fixed point in time, when the overall downside risk of all the wealth processes involved is bounded. Later, [8] treated the case of infinite-horizon discrete-time models, where the NFLBR condition was once again used. At this point, with the continuous-path and infinite-horizon discrete-time cases resolved, there seemed to be one more “gluing” step to reach a general version of the FTAP for semimartingale models. Not only did Delbaen and Schachermayer make this step for semimartingale models, they actually further weakened the NFLBR condition to the *no free lunch with vanishing risk* (NFLVR) condition, where the previous asymptotic arbitrage at some fixed point in time is precluded and the overall downside risk of all the wealth processes *tends to zero* in the limit. In more precise mathematical terms, the NFLVR condition can be stated as $\overline{C} \cap \mathbf{L}_+^0 = \{0\}$, where $\overline{C}$ is the closure in the very strong $\mathbf{L}^\infty$ -topology of (almost sure) uniform convergence.

The NFLVR condition was finally the one that proved itself to be the most fruitful in obtaining a general version of the FTAP; see [3] and [4] (also see **Fundamental Theorem of Asset Pricing**). It is both economically plausible and mathematically convenient. Needless to say, and like many great results in science, the final simplicity and clarity of the result’s statement came with the price that the corresponding proof was extremely technical.

End Notes

^aThe exact market viability definition is still sometimes the source of debate.

^bAlso called an *equivalent martingale measure*—see **Equivalent Martingale Measures** for an account of the different notions that the previous appellation encompasses.

^c $\mathbf{L}_+^0$ is the subset of $\mathbf{L}^0$ consisting of nonnegative random variables.

^dA separating measure is a probability $\mathbb{Q}$ equivalent to the original one such that all elements of C have nonpositive expectation with respect to $\mathbb{Q}$. In this context, also see **Fundamental Theorem of Asset Pricing**. Note that in the case of a continuous-time market model with locally bounded asset prices, a separating measure automatically makes the discounted asset prices *local* martingales. This was proved in [3].

^eFor a compact and rather elementary proof of this result, see [6].

^fThe appellation to this condition was actually coined by W. Schachermayer in [8].

References

- [1] Dalang, R.C., Morton, A. & Willinger, W. (1990). Equivalent martingale measures and no-arbitrage in stochastic securities market models. *Stochastics and Stochastics Reports* **29**, 185–201.
- [2] Delbaen, F. (1992). Representing martingale measures when asset prices are continuous and bounded. *Mathematical Finance* **2**, 107–130.
- [3] Delbaen, F. & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing, *Mathematische Annalen* **300**, 463–520.
- [4] Delbaen, F. & Schachermayer, W. (1998). The fundamental theorem of asset pricing for unbounded stochastic processes. *Mathematische Annalen* **312**, 215–250.
- [5] Harrison, J.M. & Kreps, D.M. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- [6] Kabanov, Y. & Stricker, C. (2001). A teachers’ note on no-arbitrage criteria, in *Seminaire de Probabilités, XXXV*, Lecture Notes in Mathematics, Springer, Berlin, Vol. 1755, 149–152.
- [7] Kreps, D.M. (1981). Arbitrage and equilibrium in economies with infinitely many commodities, *Journal of Mathematical Economics* **8**, 15–35.
- [8] Schachermayer, W. (1994). Martingale measures for discrete-time processes with infinite horizon, *Mathematical Finance* **4**, 25–55.

CONSTANTINOS KARDARAS

Minimal Entropy Martingale Measure

Consider a stochastic process $S = (S_t)_{t \geq 0}$ on a probability space $(\Omega, \mathcal{F}, P)$ and adapted to a filtration $\mathcal{F} = (\mathcal{F}_t)_{t \geq 0}$. Each S_t takes values in $\mathbb{R}^d$ and models the discounted prices at time t of d basic assets traded in a financial market. An equivalent local martingale measure (ELMM) for S , possibly on $[0, T]$ for a time horizon $T < \infty$, is a probability measure Q equivalent to the original (historical, real-world) measure P (on $\mathcal{F}_T$, if there is a T) such that S is a local Q -martingale (on $[0, T]$, respectively); *see Equivalent Martingale Measures*. If S is a nonnegative P -semimartingale, the fundamental theorem of asset pricing says that the existence of an ELMM Q for S is equivalent to the absence-of-arbitrage condition (NFLVR) that S admits no free lunch with vanishing risk; *see Fundamental Theorem of Asset Pricing*.

Definition 1 Fix a time horizon $T < \infty$. An ELMM Q^E for S on $[0, T]$ is called *minimal entropy martingale measure (MEMM)* if Q^E minimizes the relative entropy $H(Q|P)$ over all ELMMs Q for S on $[0, T]$.

Recall that the *relative entropy* is defined as

$$H(Q|P) := \begin{cases} E_P \left[\frac{dQ}{dP} \log \frac{dQ}{dP} \right] & \text{if } Q \ll P \\ +\infty & \text{otherwise} \end{cases} \quad (1)$$

This is an example of the general concept of an *f-divergence* of the form

$$D_f(Q|P) := \begin{cases} E_P \left[f \left(\frac{dQ}{dP} \right) \right] & \text{if } Q \ll P \\ +\infty & \text{otherwise} \end{cases} \quad (2)$$

where f is a convex function on $[0, \infty)$; *see* [26, 49], or [22] for a number of examples. The minimizer $Q^{*,f}$ of $D_f(\cdot|P)$ is then called *f-optimal ELMM*.

In many situations arising in mathematical finance, *f*-optimal ELMMs come up *via* duality from expected utility maximization problems; *see Expected Utility Maximization: Duality Methods; Expected Utility Maximization*. One starts with a utility function U (*see Utility Function*) and obtains f (up to an affine function) as the convex conjugate

of U , that is,

$$f(y) - \alpha y - \beta = \sup_x (U(x) - xy) \quad (3)$$

Finding $Q^{*,f}$ is then the dual to the primal problem of maximizing the expected utility

$$\vartheta \mapsto E \left[U \left(x_0 + \int_0^T \vartheta_r dS_r \right) \right] \quad (4)$$

from terminal wealth over allowed investment strategies ϑ . Moreover, under suitable conditions, the solutions $Q^{*,f}$ and $\vartheta^{*,U}$ are related by

$$\frac{dQ^{*,f}}{dP} = \text{const. } U' \left(x_0 + \int_0^T \vartheta_r^{*,U} dS_r \right) \quad (5)$$

More details can, for instance, be found in [26, 41, 46, 67, 68]. Relative entropy comes up with $f_E(y) = y \log y$ when one starts with the exponential utility functions $U_\alpha(x) = -e^{-\alpha x}$ with risk aversion $\alpha > 0$. The duality in this special case has been studied in detail in [8, 18, 40].

Since f_E is strictly convex, the minimal entropy martingale measure is always unique. If S is locally bounded, the MEMM (on $[0, T]$) exists if and only if there is at least one ELMM Q for S on $[0, T]$ with $H(Q|P) < \infty$ [21]. For general unbounded S , the MEMM need not exist; [21] contains a counterexample, and [1] shows how the duality above will then fail. In [21], it is also shown that the MEMM is automatically equivalent to P , even if it is defined as the minimizer of $H(Q|P)$ over all P -absolutely continuous local martingale measures for S on $[0, T]$, provided that there exists some ELMM Q for S on $[0, T]$ with $H(Q|P) < \infty$. Moreover, the density of Q^E with respect to P on $\mathcal{F}_T$ has a very specific form; it is given by

$$\frac{dQ^E}{dP} \Big|_{\mathcal{F}_T} = Z_T^E = Z_0 \exp \left(\int_0^T \vartheta_r^E dS_r \right) \quad (6)$$

for some constant $Z_0 > 0$ and some predictable S -integrable process ϑ^E . This has been proved in [21] for models in finite discrete time and in [26, 28] in general; *see also* [23] for an application to finding optimal strategies in a Lévy process setting. Note,

however, that representation (2) holds only at the time horizon, T ; the density process

$$Z_t^E = \frac{dQ^E}{dP} \Big|_{\mathcal{F}_t} = E_P[Z_T^E | \mathcal{F}_t], \quad 0 \leq t \leq T \quad (7)$$

is usually quite difficult to find. We remark that the above results on both the equivalence to P and the structure of the f_E -optimal Q^E have versions for more general f -divergences [26]. (Essentially, equation (2) is relation (1) in the case of exponential utility, but it can also be proved directly without using general duality.)

The history of the minimal entropy martingale measure Q^E is not straightforward to trace. A general definition and an authoritative exposition are given by Frittelli [21]. However, the idea of the so-called mini-max measures to link martingale measures *via* duality to utility maximization already appears, for instance, in [30, 31, 41]; see also [8]. Other early contributors include Miyahara [53], who used the term “canonical martingale measure”, and Stutzer [70]; some more historical comments and references are contained in [71]. Even before, in [20], it was shown that the property defining the MEMM is satisfied by the so-called minimal martingale measure if S is continuous and the so-called mean-variance trade-off of S has constant expectation over all ELMMs for S ; see also **Minimal Martingale Measure**. The most prominent example for this occurs when S is a Markovian diffusion [53].

After the initial foundations, work on the MEMM has mainly concentrated on three major areas. The first aims to determine or describe the MEMM and, in particular, its density process Z^E more explicitly in specific models. This has been done, among others, for the following:

- stochastic volatility models: see [9, 10, 35, 62, 63], and compare also **Volatility; Barndorff-Nielsen and Shephard (BNS) Models**;
- jump-diffusions [54]; and
- Lévy processes (see **Lévy Processes**), both in general and in special settings: see [36] for an overview and [42, 43] for some examples. In particular, many studies have considered *exponential Lévy models* (see **Exponential Lévy Models**) where $S = S_0 \mathcal{E}(L)$ and L is a Lévy process under P . There, the existence of the MEMM Q^E reduces to an analytical condition on the Lévy triplet of L . Moreover, Q^E is then

given by an Esscher transform (see **Esscher Transform**) and L is again a Lévy process under Q^E ; see, for instance, [13, 19, 24, 39].

For continuous semimartingales S , an alternative approach is to characterize Z^E *via* semimartingale backward equations or backward stochastic differential equations [50, 52]. The results in [56, 57] use a mixture of the above ideas in a specific class of models.

The second major area is concerned with convergence questions. Several authors have proved, in several settings and with various techniques, that the minimal entropy martingale measure Q^E is the limit, as $p \searrow 1$, of the so-called p -optimal martingale measures obtained by minimizing the f -divergence associated to the function $f(y) = y^p$. This line of research was initiated in [27, 28], and later contributions include [39, 52, 65]. In [45, 60], this convergence is combined with the general duality (1) from utility maximization in order to obtain convergence results for optimal wealths and strategies as well.

The third, and by far the most important area of research on the MEMM, is centered on its link to the exponential utility maximization problem; see [8, 18] for a detailed exposition of this issue. More specifically, the MEMM is very useful when one studies the valuation of contingent claims by (exponential) *utility indifference valuation*; see **Utility Indifference Valuation**. To explain this, we fix an initial capital x_0 and a random payoff H due at time T . The maximal expected utility one can obtain by trading in S *via* some strategy ϑ , if one starts with x_0 and has to pay out H in T , is

$$\sup_{\vartheta} E \left[U \left(x_0 + \int_0^T \vartheta_r dS_r - H \right) \right] =: u(x_0; -H) \quad (8)$$

and the *utility indifference value* x_H is then implicitly defined by

$$u(x_0 + x_H; -H) = u(x_0; 0) \quad (9)$$

Hence, x_H represents the monetary compensation required for selling H if one wants to achieve utility indifference at the optimal investment behavior. If $U = U_\alpha$ is exponential, its multiplicative structure

makes the analysis of the utility indifference value x_H tractable, in remarkable contrast to all other classical utility functions. Moreover, $u(x_0; -H)$ as well as x_H and the optimal strategy ϑ_H^* can be described with the help of a minimal entropy martingale measure (defined here with respect to a new, H -dependent reference measure P_H instead of P). This topic has first been studied in [4, 58, 59, 64]; later work has examined intertemporally dynamic extensions [5, 51], descriptions *via* backward stochastic differential equations (BSDEs) in specific models [6, 51], extensions to more general payoff structures [38, 47, 48, 61], and so on [29, 37, 69].

Apart from the above, there are a number of other areas where the minimal entropy martingale measure has come up; these include the following:

- option price comparisons [7, 11, 32–34, 55];
- generalizations or connections to other optimal ELMMs [2, 14, 15, 66]); *see also* **Minimal Martingale Measure** and [20];
- utility maximization with a random time horizon [12];
- good deal bounds [44]; *see also* **Good-deal Bounds**; and
- a calibration game [25].

There are also many papers that simply choose the MEMM as pricing measure for option pricing applications; especially in papers from the actuarial literature, this approach is often motivated by the connections between the MEMM and the Esscher transformation. Finally, we mention that the idea of looking for a martingale measure subject to a constraint on relative entropy also naturally comes up in calibration problems; *see, for instance*, [3, 16, 17] and **Model Calibration**.

References

- [1] Acciaio, B. (2005). Absolutely continuous optimal martingale measures, *Statistics and Decisions* **23**, 81–100.
- [2] Arai, T. (2001). The relations between minimal martingale measure and minimal entropy martingale measure, *Asia-Pacific Financial Markets* **8**, 137–177.
- [3] Avellaneda, M. (1998). Minimum-relative-entropy calibration of asset pricing models, *International Journal of Theoretical and Applied Finance* **1**, 447–472.
- [4] Becherer, D. (2003). Rational hedging and valuation of integrated risks under constant absolute risk aversion, *Insurance: Mathematics and Economics* **33**, 1–28.
- [5] Becherer, D. (2004). Utility-indifference hedging and valuation via reaction-diffusion systems, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **460**, 27–51.
- [6] Becherer, D. (2006). Bounded solutions to backward SDEs with jumps for utility optimization and indifference hedging, *Annals of Applied Probability* **16**, 2027–2054.
- [7] Bellamy, N. (2001). Wealth optimization in an incomplete market driven by a jump-diffusion process, *Journal of Mathematical Economics* **35**, 259–287.
- [8] Bellini, F. & Frittelli, M. (2002). On the existence of minimax martingale measures, *Mathematical Finance* **12**, 1–21.
- [9] Benth, F.E. & Karlsen, K.H. (2005). A PDE representation of the density of the minimal entropy martingale measure in stochastic volatility markets, *Stochastics* **77**, 109–137.
- [10] Benth, F.E. & Meyer-Brandis, T. (2005). The density process of the minimal entropy martingale measure in a stochastic volatility model with jumps, *Finance and Stochastics* **9**, 563–575.
- [11] Bergenthum, J. & Rüschendorf, L. (2007). Convex ordering criteria for Lévy processes, *Advances in Data Analysis and Classification* **1**, 143–173.
- [12] Blanchet-Scalliet, C., El Karoui, N. & Martellini, L. (2005). Dynamic asset pricing theory with uncertain time-horizon, *Journal of Economic Dynamics and Control* **29**, 1737–1764.
- [13] Chan, T. (1999). Pricing contingent claims on stocks driven by Lévy processes, *Annals of Applied Probability* **9**, 504–528.
- [14] Choulli, T. & Stricker, C. (2005). Minimal entropy-Hellinger martingale measure in incomplete markets, *Mathematical Finance* **15**, 465–490.
- [15] Choulli, T. & Stricker, C. (2006). More on minimal entropy-Hellinger martingale measure, *Mathematical Finance* **16**, 1–19.
- [16] Cont, R. & Tankov, P. (2004). Nonparametric calibration of jump-diffusion option pricing models, *Journal of Computational Finance* **7**, 1–49.
- [17] Cont, R. & Tankov, P. (2006). Retrieving Lévy processes from option prices: regularization of an ill-posed inverse problem, *SIAM Journal on Control and Optimization* **45**, 1–25.
- [18] Delbaen, F., Grandits, P., Rheinländer, T., Samperi, D., Schweizer, M. & Stricker, C. (2002). Exponential hedging and entropic penalties, *Mathematical Finance* **12**, 99–123.
- [19] Esche, F. & Schweizer, M. (2005). Minimal entropy preserves the Lévy property: how and why, *Stochastic Processes and their Applications* **115**, 299–327.
- [20] Föllmer, H. & Schweizer, M. (1991). Hedging of contingent claims under incomplete information, in M.H.A. Davis & R.J. Elliott, eds, *Applied Stochastic Analysis*,

- Stochastics Monographs, Gordon and Breach, London, Vol. 5, pp. 389–414.
- [21] Frittelli, M. (2000). The minimal entropy martingale measure and the valuation problem in incomplete markets, *Mathematical Finance* **10**, 39–52.
- [22] Frittelli, M. (2000). Introduction to a theory of value coherent with the no-arbitrage principle, *Finance and Stochastics* **4**, 275–297.
- [23] Fujiwara, T. (2004). From the minimal entropy martingale measures to the optimal strategies for the exponential utility maximization: the case of geometric Lévy processes, *Asia-Pacific Financial Markets* **11**, 367–391.
- [24] Fujiwara, T. & Miyahara, Y. (2003). The minimal entropy martingale measures for geometric Lévy processes, *Finance and Stochastics* **7**, 509–531.
- [25] Glonti, O., Harremoes, P., Khechinashvili, Z., Topsøe, F. & Tbilisi, G. (2007). Nash equilibrium in a game of calibration, *Theory of Probability and its Applications* **51**, 415–426.
- [26] Goll, T. & Rüschendorf, L. (2001). Minimax and minimal distance martingale measures and their relationship to portfolio optimization, *Finance and Stochastics* **5**, 557–581.
- [27] Grandits, P. (1999). The p -optimal martingale measure and its asymptotic relation with the minimal entropy martingale measure, *Bernoulli* **5**, 225–247.
- [28] Grandits, P. & Rheinländer, T. (2002). On the minimal entropy martingale measure, *Annals of Probability* **30**, 1003–1038.
- [29] Grasselli, M. (2007). Indifference pricing and hedging for volatility derivatives, *Applied Mathematical Finance* **14**, 303–317.
- [30] He, H. & Pearson, N.D. (1991). Consumption and portfolio policies with incomplete markets and short-sale constraints: the finite-dimensional case, *Mathematical Finance* **1**(3), 1–10.
- [31] He, H. & Pearson, N.D. (1991). Consumption and portfolio policies with incomplete markets and short-sale constraints: the infinite dimensional case, *Journal of Economic Theory* **54**, 259–304.
- [32] Henderson, V. (2005). Analytical comparisons of option prices in stochastic volatility models, *Mathematical Finance* **15**, 49–59.
- [33] Henderson, V. & Hobson, D.G. (2003). Coupling and option price comparisons in a jump-diffusion model, *Stochastics and Stochastics Reports* **75**, 79–101.
- [34] Henderson, V., Hobson, D., Howison, S. & Kluge, T. (2005). A comparison of option prices under different pricing measures in a stochastic volatility model with correlation, *Review of Derivatives Research* **8**, 5–25.
- [35] Hobson, D. (2004). Stochastic volatility models, correlation, and the q -optimal measure, *Mathematical Finance* **14**, 537–556.
- [36] Hubalek, F. & Sgarra, C. (2006). Esscher transforms and the minimal entropy martingale measure for exponential Lévy models, *Quantitative Finance* **6**, 125–145.
- [37] İlhan, A., Jonsson, M. & Sircar, R. (2005). Optimal investment with derivative securities, *Finance and Stochastics* **9**, 585–595.
- [38] İlhan, A. & Sircar, R. (2006). Optimal static-dynamic hedges for barrier options, *Mathematical Finance* **16**, 359–385.
- [39] Jeanblanc, M., Klöppel, S. & Miyahara, Y. (2007). Minimal f^q -martingale measures for exponential Lévy processes, *Annals of Applied Probability* **17**, 1615–1638.
- [40] Kabanov, Y.M. & Stricker, C. (2002). On the optimal portfolio for the exponential utility maximization: remarks to the six-author paper, *Mathematical Finance* **12**, 125–134.
- [41] Karatzas, I., Lehoczky, J.P., Shreve, S.E. & Xu, G.L. (1991). Martingale and duality methods for utility maximization in an incomplete market, *SIAM Journal on Control and Optimization* **29**, 702–730.
- [42] Kassberger, S. & Liebmann, T. (2008). *Minimal q -entropy Martingale Measures for Exponential Time-changed Lévy Processes and within Parametric Classes*, preprint, University of Ulm, <http://www.uniu-lm.de/mawi/finmath/people/kassberger.html>
- [43] Kim, Y.S. & Lee, J.H. (2007). The relative entropy in CGMY processes and its applications to finance, *Mathematical Methods of Operations Research* **66**, 327–338.
- [44] Klöppel, S. & Schweizer, M. (2007). Dynamic utility-based good deal bounds, *Statistics and Decisions* **25**, 285–309.
- [45] Kohlmann, M. & Niethammer, C.R. (2007). On convergence to the exponential utility problem, *Stochastic Processes and their Applications* **117**, 1813–1834.
- [46] Kramkov, D. & Schachermayer, W. (1999). The asymptotic elasticity of utility functions and optimal investment in incomplete markets, *Annals of Applied Probability* **9**, 904–950.
- [47] Leung, T. & Sircar, R. (2008). *Exponential Hedging with Optimal Stopping and Application to ESO Valuation*, preprint, Princeton University, <http://ssrn.com/abstract=1111993>
- [48] Leung, T. & Sircar, R. (2009). Accounting for risk aversion, vesting, job termination risk and multiple exercises in valuation of employee stock options, *Mathematical Finance* **19**, 99–128.
- [49] Liese, F. & Vajda, I. (1987). *Convex Statistical Distances*, Teubner.
- [50] Mania, M., Santacrose, M. & Tevzadze, R. (2003). A semimartingale BSDE related to the minimal entropy martingale measure, *Finance and Stochastics* **7**, 385–402.
- [51] Mania, M. & Schweizer, M. (2005). Dynamic exponential utility indifference valuation, *Annals of Applied Probability* **15**, 2113–2143.
- [52] Mania, M. & Tevzadze, R. (2003). A unified characterization of q -optimal and minimal entropy martingale measures by semimartingale backward equations, *Georgian Mathematical Journal* **10**, 289–310.

- [53] Miyahara, Y. (1995). Canonical martingale measures of incomplete assets markets, *Probability Theory and Mathematical Statistics: Proceedings of the Seventh Japan-Russia Symposium*, Tokyo, pp. 343–352.
- [54] Miyahara, Y. (1999). Minimal entropy martingale measures of jump type price processes in incomplete assets markets, *Asia-Pacific Financial Markets* **6**, 97–113.
- [55] Møller, T. (2004). Stochastic orders in dynamic reinsurance markets, *Finance and Stochastics* **8**, 479–499.
- [56] Monoyios, M. (2006). Characterisation of optimal dual measures via distortion, *Decisions in Economics and Finance* **29**, 95–119.
- [57] Monoyios, M. (2007). The minimal entropy measure and an Esscher transform in an incomplete market, *Statistics and Probability Letters* **77**, 1070–1076.
- [58] Musiela, M. & Zariphopoulou, T. (2004). An example of indifference prices under exponential preferences, *Finance and Stochastics* **8**, 229–239.
- [59] Musiela, M. & Zariphopoulou, T. (2004). A valuation algorithm for indifference prices in incomplete markets, *Finance and Stochastics* **8**, 399–414.
- [60] Niethammer, C.R. (2008). On convergence to the exponential utility problem with jumps, *Stochastic Analysis and Applications* **26**, 169–196.
- [61] Oberman, A. & Zariphopoulou, T. (2003). Pricing early exercise contracts in incomplete markets, *Computational Management Science* **1**, 75–107.
- [62] Rheinländer, T. (2005). An entropy approach to the Stein and Stein model with correlation, *Finance and Stochastics* **9**, 399–413.
- [63] Rheinländer, T. & Steiger, G. (2006). The minimal entropy martingale measure for general Barndorff-Nielsen/Shephard models, *Annals of Applied Probability* **16**, 1319–1351.
- [64] Rouge, R. & El Karoui, N. (2000). Pricing via utility maximization and entropy, *Mathematical Finance* **10**, 259–276.
- [65] Santacrose, M. (2005). On the convergence of the p -optimal martingale measures to the minimal entropy martingale measure, *Stochastic Analysis and Applications* **23**, 31–54.
- [66] Santacrose, M. (2006). Derivatives pricing via p -optimal martingale measures: some extreme cases, *Journal of Applied Probability* **43**, 634–651.
- [67] Schachermayer, W. (2001). Optimal investment in incomplete markets when wealth may become negative, *Annals of Applied Probability* **11**, 694–734.
- [68] Schäl, M. (2000). Portfolio optimization and martingale measures, *Mathematical Finance* **10**, 289–303.
- [69] Stoikov, S. (2006). Pricing options from the point of view of a trader, *International Journal of Theoretical and Applied Finance* **9**, 1245–1266.
- [70] Stutzer, M. (1996). A simple nonparametric approach to derivative security valuation, *Journal of Finance* **51**, 1633–1652.
- [71] Stutzer, M.J. (2000). Simple entropic derivation of a generalized Black-Scholes option pricing model, *Entropy* **2**, 70–77.

Related Articles

Entropy-based Estimation; Exponential Lévy Models; Minimal Martingale Measure; Risk-neutral Pricing; Semimartingale.

MARTIN SCHWEIZER

Minimal Martingale Measure

Let $S = (S_t)$ be a stochastic process on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), P)$ that models the discounted prices of primary traded assets in a financial market. An equivalent local martingale measure (ELMM) for S is a probability measure Q equivalent to the original (historical) measure P such that S is a local Q -martingale (see **Equivalent Martingale Measures**). If S is a nonnegative P -semimartingale, the fundamental theorem of asset pricing says that an ELMM Q for S exists if and only if S satisfies the no-arbitrage condition (NFLVR), that is, admits no free lunch with vanishing risk (see **Fundamental Theorem of Asset Pricing**). By Girsanov's theorem, S is then under P a semimartingale with a decomposition $S = S_0 + M + A$ into a local P -martingale M and an adapted process A of finite variation. If S is special under P , then A can be chosen predictable and the resulting canonical decomposition of S is unique. We say that S satisfies the structure condition (SC) if M is locally P -square-integrable and A has the form $A = \int d\langle M \rangle \lambda$ for a predictable process λ such that the increasing process $\int \lambda' d\langle M \rangle \lambda$ is finite-valued. In an Itô process model where S is given by a stochastic differential equation $dS_t = S_t((\mu_t - r_t)dt + \sigma_t dW_t)$, the latter process is given by $\int ((\mu_t - r_t)/\sigma_t)^2 dt$, the integrated squared instantaneous Sharpe ratio of S (see **Sharpe Ratio**).

Definition 1 Suppose S satisfies (SC). An ELMM $\hat{P}$ for S with P -square-integrable density $d\hat{P}/dP$ is called minimal martingale measure (MMM) (for S) if $\hat{P} = P$ on $\mathcal{F}_0$ and if every local P -martingale L that is locally P -square integrable and strongly P -orthogonal to M is also a local $\hat{P}$ -martingale. We call $\hat{P}$ orthogonality preserving if L is also strongly $\hat{P}$ -orthogonal to S .

The basic idea for the MMM first appeared in [46] in a more specific model, where it was used as an auxiliary technical tool in the context of local risk-minimization (see also **Hedging** for an overview of key ideas on hedging and **Mean-Variance Hedging** for an alternative quadratic approach). More precisely, the so-called

locally risk-minimizing strategy for a given contingent claim H was obtained there (under some specific assumptions) as the integrand from the classical Galtchouk–Kunita–Watanabe decomposition of H under $\hat{P}$. However, the introduction of $\hat{P}$ in [46] and also in [47] was still somewhat *ad hoc*. The above definition was given in [18] where the main results presented here can also be found. In particular, [18] showed that for continuous S , the Galtchouk–Kunita–Watanabe decomposition of H under the MMM $\hat{P}$ provides (under very mild integrability conditions) the so-called Föllmer–Schweizer decomposition of H under the original measure P , and this in turn immediately gives the locally risk-minimizing strategy for H . We emphasize that this is no longer true, in general, if S has jumps. The MMM subsequently found various other applications and uses and has become fairly popular, especially in models with continuous price processes.

Suppose now S satisfies (SC). For every ELMM Q for S with $dQ/dP \in L^2(P)$, the density process then takes the form

$$Z^Q := \frac{dQ}{dP} \Big|_{\mathcal{F}} = Z_0^Q \mathcal{E} \left(- \int \lambda dM + L^Q \right) \quad (1)$$

with some locally P -square-integrable local P -martingale L^Q . If the MMM $\hat{P}$ exists, then it has $\hat{Z}_0 = 1$ and $L^{\hat{P}} \equiv 0$, and its density process is thus given by the stochastic exponential (see **Stochastic Exponential**)

$$\begin{aligned} \hat{Z} &= \mathcal{E} \left(- \int \lambda dM \right) \\ &= \exp \left(- \int \lambda dM - \frac{1}{2} \int \lambda' d[M] \lambda \right) \\ &\quad \times \prod (1 - \lambda' \Delta M) \exp \left(\lambda' \Delta M + \frac{1}{2} (\lambda' \Delta M)^2 \right) \end{aligned} \quad (2)$$

The advantage of this explicit representation is that it allows to determine the MMM $\hat{P}$ and its density process $\hat{Z}$ directly from the ingredients M and λ of the canonical decomposition of S . Conversely, one can start with the above expression for $\hat{Z}$ to define a candidate for the density process of the MMM. This gives existence of the MMM under the following conditions:

1. $\hat{Z}$ is strictly positive; this happens if and only if $\lambda' \Delta M < 1$, that is, all the jumps of $\int \lambda dM$ are strictly below 1.
2. The local P -martingale $\hat{Z}$ is a true P -martingale.
3. $\hat{Z}$ is P -square-integrable.

Condition 1 automatically holds (on any finite time interval) if S , hence also M , is continuous; it typically fails in models where S has jumps. Conditions 2 and 3 can fail even if 1 holds and even if there exists some ELMM for S with P -square-integrable density; see [45] or [15] for a counterexample.

The above explicit formula for $\hat{Z}$ shows that $\hat{P}$ is minimal in the sense that its density process contains the smallest number of symbols among all ELMMs Q . More seriously, the original idea was that $\hat{P}$ should turn S into a (local) martingale while having a minimal impact on the overall martingale structure of our setting. This is captured and made precise by the definition. If S is continuous, one can show that $\hat{P}$ is even orthogonality preserving; see [18] for this, and note that this usually fails if S has jumps.

To some extent, the naming of the “minimal” martingale measure is misleading since $\hat{P}$ was not originally defined as the minimizer of a particular functional on ELMMs. However, if S is continuous, Föllmer and Schweizer [18] have proved that $\hat{P}$ minimizes

$$Q \mapsto H(Q|P) - E_Q \left[\int_0^\infty \lambda'_u d\langle M \rangle_u \lambda u \right] \quad (3)$$

over all ELMMs Q for S ; see also [49]. Moreover, Schweizer [50] has shown that if S is continuous, then $\hat{P}$ minimizes the reverse relative entropy $H(P|Q)$ over all ELMMs Q for S ; this no longer holds if S has jumps. Under more restrictive assumptions, other minimality properties for $\hat{P}$ have been obtained by several authors. However, a general result under the sole assumption (SC) is not available so far.

There is a large amount of literature related to the MMM. In fact, a Google Scholar search for “minimal martingale measure” (enclosed in quotation marks) produced in April 2008 a list of well over 400 hits. As a first category, this contains papers where the MMM is studied *per se* or used as in the original approach of local risk-minimization. In terms of topics, the following areas of related work can be found in that category:

- Properties, characterization results, and generalizations for the MMM: [1, 4, 9–11, 14, 19, 33, 36, 37, 49, 51];
- Convergence results for option prices (computed under the MMM): [25, 32, 42, 44];
- Applications to hedging: [7, 39, 47, 48] (*see also Hedging*).
- Uses for option pricing: [8, 13, 55], to name only a very few; comparison results for option prices are given in [22, 24, 34] (*see also Risk-neutral Pricing*).
- Problems and counterexamples: [15, 16, 43, 45, 52].
- Equilibrium justifications for using the MMM: [26, 40].

A second category of papers contains those where the MMM has (sometimes unexpectedly) come up in connection with various other problems and topics in mathematical finance. Examples include the following:

- Classical utility maximization and utility indifference valuation [3, 20, 21, 23, 35, 41, 53, 54]: the MMM here often appears because the special structure of a given model implies that $\hat{P}$ has a particular optimality property (*see also Expected Utility Maximization; Expected Utility Maximization: Duality Methods; Utility Indifference Valuation; and Minimal Entropy Martingale Measure*).
- The numeraire portfolio and growth-optimal investment [2, 12]: this is related to the minimization of the reverse relative entropy $H(P|\cdot)$ over ELMMs (*see also Kelly Problem*).
- The concept of value preservation [28–30]: here the link seems to come up because value preservation is, like local risk-minimization, a local optimality criterion.
- Good deal bounds in incomplete markets [5, 6]: the MMM naturally shows up here because good deal bounds are formulated *via* instantaneous quadratic restrictions on the pricing kernel (ELMM) to be chosen (*see also Good-deal Bounds; Sharpe Ratio; Market Price of Risk*).
- Local utility maximization [27]; again, the link here is due to the local nature of the criterion that is used.
- Risk-sensitive control [17, 31, 38]; this is an area where the connection to the MMM seems

not yet well understood. See also **Risk-sensitive Asset Management**.

References

- [1] Arai, T. (2001). The relations between minimal martingale measure and minimal entropy martingale measure, *Asia-Pacific Financial Markets* **8**, 137–177.
- [2] Becherer, D. (2001). The numeraire portfolio for unbounded semimartingales, *Finance and Stochastics* **5**, 327–341.
- [3] Berrier, F., Rogers, L.C.G. & Tehranchi, M. (2008). *A Characterization of Forward Utility Functions*, preprint, <http://www.statslab.cam.ac.uk/~mike/forward-utilities.pdf>.
- [4] Biagini, F. & Pratelli, M. (1999). Local risk minimization and numeraire, *Journal of Applied Probability* **36**, 1126–1139.
- [5] Björk, T. & Slinko, I. (2006). Towards a general theory of good-deal bounds, *The Review of Finance* **10**, 221–260.
- [6] Černý, A. (2003). Generalised Sharpe ratios and asset pricing in incomplete markets, *European Finance Review* **7**, 191–233.
- [7] Černý, A. & Kallsen, J. (2007). On the structure of general mean-variance hedging strategies, *The Annals of Probability* **35**, 1479–1531.
- [8] Chan, T. (1999). Pricing contingent claims on stocks driven by Lévy processes, *The Annals of Applied Probability* **9**, 504–528.
- [9] Choulli, T. & Stricker, C. (2005). Minimal entropy-Hellinger martingale measure in incomplete markets, *Mathematical Finance* **15**, 465–490.
- [10] Choulli, T. & Stricker, C. (2006). More on minimal entropy-Hellinger martingale measure, *Mathematical Finance* **16**, 1–19.
- [11] Choulli, T., Stricker, C. & Li, J. (2007). Minimal Hellinger martingale measures of order q , *Finance and Stochastics* **11**, 399–427.
- [12] Christensen, M.M. & Larsen, K. (2007). No arbitrage and the growth optimal portfolio, *Stochastic Analysis and Applications* **25**, 255–280.
- [13] Colwell, D.B. & Elliott, R.J. (1993). Discontinuous asset prices and non-attainable contingent claims, *Mathematical Finance* **3**, 295–308.
- [14] Delbaen, F., Grandits, P., Rheinländer, T., Samperi, D., Schweizer, M. & Stricker, C. (2002). Exponential hedging and entropic penalties, *Mathematical Finance* **12**, 99–123.
- [15] Delbaen, F. & Schachermayer, W. (1998). A simple counterexample to several problems in the theory of asset pricing, *Mathematical Finance* **8**, 1–11.
- [16] Elliott, R.J. & Madan, D.B. (1998). A discrete time equivalent martingale measure, *Mathematical Finance* **8**, 127–152.
- [17] Fleming, W.H. & Sheu, S.J. (2002). Risk-sensitive control and an optimal investment model II, *The Annals of Applied Probability* **12**, 730–767.
- [18] Föllmer, H. & Schweizer, M. (1991). Hedging of contingent claims under incomplete information, in *Applied Stochastic Analysis, Stochastics Monographs*, M.H.A. Davis & R.J. Elliott eds, Gordon and Breach, London, Vol. 5, pp. 389–414.
- [19] Grandits, P. (2000). On martingale measures for stochastic processes with independent increments, *Theory of Probability and its Applications* **44**, 39–50.
- [20] Grasselli, M. (2007). Indifference pricing and hedging for volatility derivatives, *Applied Mathematical Finance* **14**, 303–317.
- [21] Henderson, V. (2002). Valuation of claims on nontraded assets using utility maximization, *Mathematical Finance* **12**, 351–373.
- [22] Henderson, V. (2005). Analytical comparisons of option prices in stochastic volatility models, *Mathematical Finance* **15**, 49–59.
- [23] Henderson, V. & Hobson, D.G. (2002). Real options with constant relative risk aversion, *Journal of Economic Dynamics and Control* **27**, 329–355.
- [24] Henderson, V. & Hobson, D.G. (2003). Coupling and option price comparisons in a jump-diffusion model, *Stochastics and Stochastics Reports* **75**, 79–101.
- [25] Hong, D. & Wee, I.S. (2003). Convergence of jump-diffusion models to the Black-Scholes model, *Stochastic Analysis and Applications* **21**, 141–160.
- [26] Jouini, E. & Napp, C. (1999). *Continuous Time Equilibrium Pricing of Nonredundant Assets*, Leonard N. Stern School Finance Department Working Paper 99-008, New York University, http://w4.stern.nyu.edu/finance/research.cfm?doc_id=1216, <http://www.stern.nyu.edu/fin/workpapers/papers99/wpa99008.pdf>.
- [27] Kallsen, J. (2002). Utility-based derivative pricing in incomplete markets, in *Mathematical Finance—Bachelier Congress 2000*, H. Geman, D. Madan, S.R. Pliska & T. Vorst, eds, Springer-Verlag, Berlin, Heidelberg, New York, pp. 313–338.
- [28] Korn, R. (1998). Value preserving portfolio strategies and the minimal martingale measure, *Mathematical Methods of Operations Research* **47**, 169–179.
- [29] Korn, R. (2000). Value preserving strategies and a general framework for local approaches to optimal portfolios, *Mathematical Finance* **10**, 227–241.
- [30] Korn, R. & Schäl, M. (1999). On value preserving and growth optimal portfolios, *Mathematical Methods of Operations Research* **50**, 189–218.
- [31] Kuroda, K. & Nagai, H. (2002). Risk-sensitive portfolio optimization on infinite time horizon, *Stochastics and Stochastics Reports* **73**, 309–331.
- [32] Lesne, J.-P., Prigent, J.-L. & Scaillet, O. (2000). Convergence of discrete time option pricing models under stochastic interest rates, *Finance and Stochastics* **4**, 81–93.
- [33] Mania, M. & Tevzadze, R. (2003). A unified characterization of q -optimal and minimal entropy martingale

- measures by semimartingale backward equation, *The Georgian Mathematical Journal* **10**, 289–310.
- [34] Møller, T. (2004). Stochastic orders in dynamic reinsurance markets, *Finance and Stochastics* **8**, 479–499.
- [35] Monoyios, M. (2004). Performance of utility-based strategies for hedging basis risk, *Quantitative Finance* **4**, 245–255.
- [36] Monoyios, M. (2006). Characterisation of optimal dual measures via distortion, *Decisions in Economics and Finance* **29**, 95–119.
- [37] Monoyios, M. (2007). The minimal entropy measure and an Esscher transform in an incomplete market, *Statistics and Probability Letters* **77**, 1070–1076.
- [38] Nagai, H. & Peng, S. (2002). Risk-sensitive portfolio optimization with partial information on infinite time horizon, *The Annals of Applied Probability* **12**, 173–195.
- [39] Pham, H., Rheinländer, T. & Schweizer, M. (1998). Mean-variance hedging for continuous processes: new results and examples, *Finance and Stochastics* **2**, 173–198.
- [40] Pham, H. & Touzi, N. (1996). Equilibrium state prices in a stochastic volatility model, *Mathematical Finance* **6**, 215–236.
- [41] Pirvu, T.A. & Haussmann, U.G. (2007). *On Robust Utility Maximization*, University of British Columbia, arXiv:math/0702727, preprint.
- [42] Prigent, J.-L. (1999). Incomplete markets: convergence of options values under the minimal martingale measure, *Advances in Applied Probability* **31**, 1058–1077.
- [43] Rheinländer, T. (2005). An entropy approach to the Stein and Stein model with correlation, *Finance and Stochastics* **9**, 399–413.
- [44] Runggaldier, W.J. & Schweizer, M. (1995). Convergence of option values under incompleteness, in *Seminar on Stochastic Analysis, Random Fields and Applications*, E. Bolthausen, M. Dozzi & F. Russo, eds, Birkhäuser Verlag, Basel, pp. 365–384.
- [45] Schachermayer, W. (1993). A counterexample to several problems in the theory of asset pricing, *Mathematical Finance* **3**, 217–229.
- [46] Schweizer, M. (1988). *Hedging of options in a general semimartingale model*, Dissertation ETH Zürich 8615.
- [47] Schweizer, M. (1991). Option hedging for semimartingales, *Stochastic Processes and their Applications* **37**, 339–363.
- [48] Schweizer, M. (1992). Mean-variance hedging for general claims, *The Annals of Applied Probability* **2**, 171–179.
- [49] Schweizer, M. (1995). On the minimal martingale measure and the Föllmer-Schweizer decomposition, *Stochastic Analysis and Applications* **13**, 573–599.
- [50] Schweizer, M. (1999). A minimality property of the minimal martingale measure, *Statistics and Probability Letters* **42**, 27–31.
- [51] Schweizer, M. (2001). A guided tour through quadratic hedging approaches, in *Option Pricing, Interest Rates and Risk Management*, E. Jouini, J. Cvitanic & M. Musiela, eds, Cambridge University Press, Cambridge, pp. 538–574.
- [52] Sin, C.A. (1998). Complications with stochastic volatility models, *Advances in Applied Probability* **30**, 256–268.
- [53] Stoikov, S. & Zariphopoulou, T. (2004). Optimal investments in the presence of unhedgeable risks and under CARA preferences, in *IMA Volume in Mathematics and its Applications*, in press.
- [54] Tehranchi, M. (2004). Explicit solutions of some utility maximization problems in incomplete markets, *Stochastic Processes and their Applications* **114**, 109–125.
- [55] Zhang, X. (1997). Numerical analysis of American option pricing in a jump-diffusion model, *Mathematics of Operations Research* **22**, 668–690.

HANS FÖLLMER & MARTIN SCHWEIZER

Good-deal Bounds

Most contingent claims valuation is based, at least notionally, on the concept of exact replication. The difficulties of exactly replicating derivative positions suggest that in many cases we should, instead, put bounds around the value of an instrument. These bounds ought to depend on model assumptions and on the prices of securities that would be used to exploit mispricing. No-arbitrage bounds are often very weak, so good-deal bounds provide an attractive alternative.

Good-deal bounds provide a range of prices within which an instrument must trade if it is not to offer a surprisingly good reward-for-risk opportunity. This is illustrated in Figure 1, where the horizontal axis represents the distribution of future payoffs (or values) after zero cost hedging. In an incomplete market setting, rather strong assumptions are needed to arrive at a unique forward value, such as p^* in the figure. Conversely, risk-free arbitrage typically allows a rather wide band of prices, as between the upper and lower bounds b_+ , b_- . We can hope to obtain a much narrower band without the need for strong assumptions if we simply preclude profitable opportunities. This gives the good-deal bounds p_+ and p_- . These bounds have two alternative interpretations: we can think of them as establishing normative bid and ask forward prices for a particular trader or as predicting a range in which we expect the market price to lie.

This line of valuation analysis now has an interesting history and it has inspired a quite significant literature, much of it very mathematical. There are a great many different variations by which the philosophy just described can be implemented.

This article aims to cover the main issues without going too deeply into mathematical technicalities. We begin by considering a simple illustrative example to provide intuitive insights into the nature of the analysis, including the use of duality in the solutions. We then sketch the history of this topic, including the generalized Sharpe ratio. Finally, there is a discussion of the role of the utility function (*see Utility Function*) in the analysis, of applications, and of the more recent literature.

Illustration

Consider the problem faced by a financial intermediary in determining reservation bid and ask prices

for some derivative which can, at best, only be partly hedged. There is no chance of replicating this claim exactly, and super-replication bounds may be too loose to be practically helpful. The company expects to trade using some kind of statistical arbitrage, for which each transaction passes a minimum reward-for-risk threshold and overall to obtain a portfolio that performs much better than that minimum.

More specifically, reservation forward bid and ask prices $p_- < p_+$ are to be determined at time zero for a derivative that will pay a random amount $\tilde{C}_T$ at later date T . We suppose a von Neumann–Morgenstern utility function $U(\cdot)$ for date T wealth and a forward wealth endowment of W_0 . The reservation prices are constructed so that trade will provide a level of expected utility at a predetermined level U_R that exceeds the expected utility that could be reached without it by $A > 0$.

Figure 2 illustrates the construction. The horizontal axis represents the price of the contingent claim. The vertical axis represent the expected utility obtained from buying (or selling) the optimal quantity of the claim. Outside the super-replication bounds, b_- , b_+ , unbounded wealth can be obtained.

In the case where no hedging will be undertaken and the forward price of the claim is p , we simply have the optimization of the quantity θ bought or sold as

$$\underset{\theta}{Max} E \left[U \left(W_0 + \theta \left(\tilde{C}_T - p \right) \right) \right] \quad (1)$$

If p is low enough we will expect to buy the claim, and if p is high enough we will want to sell it. Intuitively the good-deal lower bound, p_- , is the highest price at which we can buy the claim and obtain expected utility of U_R , and the good-deal upper bound, p_+ , is the lowest price at which we can sell the claim and obtain expected utility of U_R .

Now consider the first-order conditions from the optimization:

$$E \left[\left(\tilde{C}_T - p \right) U' \left(W_0 + \theta \left(\tilde{C}_T - p \right) \right) \right] = 0, \text{ so } \quad (2)$$

$$p = \frac{E \left[\tilde{C}_T \left[U' \left(\tilde{W}_T \right) \right] \right]}{E \left[U' \left(\tilde{W}_T \right) \right]},$$

$$\text{where } \tilde{W}_T = W_0 + \theta \left(\tilde{C}_T - p \right) \quad (3)$$

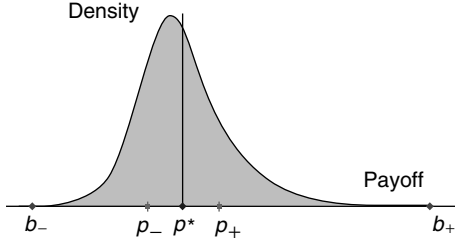


Figure 1 Good-deal bounds. Alternative forward prices and the distribution of future values after zero cost hedging: b_- , b_+ : super-replication bounds; p_- , p_+ : good-deal bounds; p^* : unique (indifference) price

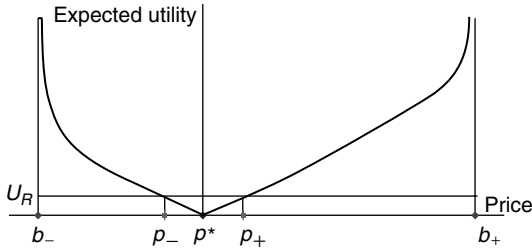


Figure 2 Expected utility against price. The good-deal bounds, p_- , p_+ , are defined as the prices at or beyond which expected utility of U_R can be obtained

Thus, the reservation price corresponds to pricing with stochastic discount factors induced by the marginal utility at the optimal wealth levels corresponding to this price.

In principle, the extension to hedging is straightforward. The gains or losses from a self-financing strategy with zero initial cost are simply added into the date T wealth. If at date t the strategy involves holdings x_t at prices P_t , the expression for wealth at date T becomes

$$\tilde{W}_T = W_0 + \theta (\tilde{C}_T - p) + \int_0^T x_t dP_t \quad (4)$$

Ideally, we would like to find and use the optimum hedging strategies, but any strategies that enhance expected utility will provide tighter reservation prices. Note that if the claim can be replicated exactly, then both the good-deal bounds will tend to the replication cost. Similarly, the good-deal bounds will always be at least as tight as any super-replication bounds that can be based on the same assumptions.

So far, we have only described the primal view of this problem. Well-known duality results provide an alternative viewpoint that provides both insights and alternative computational schemes. The good-deal lower bound is characterized as the infimum of values over nonnegative changes of measure, m , that price all reference assets and have insufficient dispersion to provide higher levels of expected utility. For example,

$$p_- = \inf_{m \geq 0} \left\{ E \left[m \tilde{C}_T \right] \right\} \text{ subject to} \\ E[V(m)] \leq A, \quad E[mS_T] = S_0 \quad (5)$$

where $V(m)$ is the conjugate function of U , defined by

$$V(m) = \sup_{S_T} \{U(S_T) - mS_T\} \text{ for } m > 0 \quad (6)$$

and the final constraint in equation (5) represents the correct pricing of reference assets.

Note that this only differs from the formulation for super-replication (no-arbitrage) bounds by the addition of the inequality constraint in equation (5), which precludes extreme changes of measure that would generate expected utility greater than U_R . In both cases, the more assets we hedge with, the more the change of measure is constrained and the tighter the valuation bounds.

Early Literature

Finding bounds on the values of derivatives has a long history. Merton [15] summarizes the conventional upper and lower bounds on vanilla options and how they are enforced by arbitrage. The subsequent contributions of Harrison and Kreps [10], Dybvig and Ross [8], and others have shown the pricing implications of no arbitrage more generally. Later papers by Perrakis and Ryan [16] and Levy [14] obtained slightly more general bounds on the prices of options, for example, based on stochastic dominance by adding some additional stronger assumptions. More recently a number of papers, such as [11], have considered super-replication bounds on exotic options when vanilla options can be used to engineer the hedge (*see Arbitrage Bounds*). Interest in this topic has further intensified with the growth of the literature on Levy processes (*see Exponential Lévy Models*), exotic options, and incomplete markets.

Much of the work in the incomplete markets literature focuses on ways to obtain a particular pricing measure and hence unique prices (for example, see **Minimal Entropy Martingale Measure**; **Minimal Martingale Measure** and Schweizer [17]), but it is not clear why a particular agent would be prepared to trade at these prices.

The good-deal literature represents an important alternative between these two paths. Hansen and Jagannathan [9] provide a crucial stepping stone. They showed that the Sharpe ratio on any security is bounded by the coefficient of variation of the stochastic discount factor (see **Stochastic Discount Factors**). The Sharpe ratio provides a very natural benchmark (see [18]) and Cochrane and Saá-Requejo [6] subsequently used this to limit the volatility of the stochastic discount factor and infer the first no-good-deal prices, conditional on the absence of high Sharpe ratios. At about the same time, a related paper by Bernardo and Ledoit [2] showed how similar bounds could be obtained relative to a maximum gain–loss ratio for the economy as a whole. These papers have their disadvantages. Cochrane and Saá-Requejo work with quadratic utility (and sometimes truncated quadratic utility), whereas Bernardo and Ledoit use Domar–Musgrave utility (i.e., two linear segments). This led Hodges [12] to investigate bounds based on the more conventional choice of exponential utility and to thereby introduce the idea of a generalized Sharpe ratio.

This concept was extended by Černý and Hodges [5] into the more general framework of good-deal pricing mostly used today. By then, it was already clear that these prices satisfied the criteria for coherent risk measures of Artzner *et al.*, [1], namely, the linearity, subadditivity, and monotonicity properties. This includes the representation of the lower good-deal price as an infimum over values from alternative pricing measures. Nevertheless, Jaschke and Küchler [13] provided an important clarification and unification of these ideas.

General Framework

The general framework of “no-good-deal” pricing (first described by Černý and Hodges [5]) places no-arbitrage and representative agent equilibrium at the two ends of a spectrum of possibilities. They define a *desirable claim* as one which provides a specific

level of von Neumann–Morgenstern expected utility and a *good-deal* as a desirable claim with zero or negative price. Within the analysis, it is assumed that any quantity of any claim may be bought or sold.

The economy contains a collection of claims with predetermined prices, so called basis assets. These claims generate the marketed subspace M and their prices define a price correspondence on this subspace. In an incomplete market, it is often convenient to suppose that the market is augmented in such a way that the resulting complete market contains no arbitrages. Instead, we can more powerfully augment the market so that the complete market contains no good-deals. We obtain a set of pricing functionals that form a subset of those that simply preclude arbitrage.

The link between no arbitrage and strictly positive pricing rules carries over to good-deals and enables price restrictions to be placed on nonmarketed claims. Under suitable technical assumptions, the no-good-deal price region for a set of claims is a convex set, and redundant assets have unique good-deal prices.

With an *acceptance set* of deals, K , typically defined in terms of expected utility, the upper and lower good-deal bounds can be defined simply as

$$p_+ = \inf_{p, x_t} \left\{ p | -\tilde{C}_T + p + \int_0^T x_t dS_t \in K \right\} \text{ and } (7)$$

$$p_+ = \inf_{p, x_t} \left\{ p | \tilde{C}_T - p + \int_0^T x_t dS_t \in K \right\} \quad (8)$$

For a given utility function, the positions of the good-deal bounds naturally depend on the required expected utility premium, A . The higher this level, the further apart the bounds will be. Coherent risk measures, well into the tails of the final distribution, can be obtained if high levels are employed for A . Except for the case of exponential utility, the bounds also depend on the initial wealth level.

Generalized Sharpe Ratios

One method for setting the required premium comes from the Sharpe ratio available on a market opportunity. This give rise to what are called generalized Sharpe ratio bounds (see [12] or [4]). The idea is to first compute the level of expected utility U_R attainable from a market opportunity offering a specific annualized Sharpe ratio, such as 0.25, and without any investment in the derivative. The good-deal

bounds that are supported by this level of expected utility (but without this market opportunity) are then said to correspond to a generalized Sharpe ratio of 0.25. In the case of negative exponential utility, the wealth level and the risk aversion parameter play the same role and become irrelevant since the opportunity can be accepted at any scale. This provides a particularly simple implementation with minimal parameter requirements.

Subsequent analysis by Černý [4] further expands both the notions and the analysis of generalized Sharpe ratios. The analysis provides details of the dual formulations for alternative standard utility functions. For example, the dual constraints on the change of measure m for different utility functions are as given in Table 1.

The various properties of the utility affect the details of the mathematical analysis considerably. For some features to work cleanly we need unbounded utility, whereas for others the behavior for low wealth levels is critical. Exponential utility precludes any delta hedge that gives a short lognormal position over finite time—even though it would have a smaller standard deviation than the fully covered position. Capping such a liability at a finite level can therefore have a big effect on the good-deal price resulting from such an analysis. Depending on the context, this may or may not be desirable. While exponential utility precludes fat negative tails, such as the short lognormal, power and log utility preclude the possibility of any negative future wealth, and even stronger effects can, in principle, derive from this.

With constant absolute risk aversion (CARA) utility, changing the scale of investment is equivalent to changing the level of risk aversion. With constant relative risk aversion (CRRA), it is equivalent to scaling the initial wealth, W_0 . The CRRA-based good-deal bound thus searches across measures with the same exponent, but different wealth levels. There may be some advantages to finding alternative utility functions that have properties intermediate between

the Domar–Musgrave function used by Bernardo and Ledoit and the negative exponential one.

Coherent Risk Measures

Jaschke and Küchler [13] expand the link between good-deal bounds and coherent risk measures. They show that there is a one-to-one correspondence between

1. “coherent risk measures” (*see* **Convex Risk Measures**)
2. cones of “desirable claims”
3. partial orderings
4. good-deal valuation bounds
5. sets of “admissible” price systems.

It should be noted from this analysis that it is sufficient but not necessary to use expected utility to define all the abstract measures considered in their paper. In other words, acceptance sets must be consistent with coherence, but not necessarily with expected utility.

It is clear from the foregoing that good-deal analysis can easily be applied as the basis of risk measurement and will satisfy the axioms of coherent risk measures (*see* **Convex Risk Measures**). They can also be applied as a method of risk adjustment for performance measurement. For example, a utility-based generalized Sharpe ratio, when applied to an empirical distribution, provides a method of adjusting for skewness in the distribution. In doing so, it makes sense to apply a negative sign to situations where a short position would have been optimal.

Recent and Prospective Literature

Important new papers continue to appear quite regularly; a few recent ones are mentioned here. Staum [19] provides much of the background, treating good-deals from the perspective of convex optimization. Bjork and Slinko [3] provide extensions to Cochrane and Saá-Requejo in a multidimensional jump-diffusion setting. There are further papers that expand on the dynamic aspects of this analysis, apply it to settings with stochastic volatility, or implement similar optimizations using mathematical programming. There are also a number of papers, which although not directly within the framework developed here, deal with related ideas in different ways.

Table 1 Stochastic discount factor constraints for various utility functions

Utility function	Constraint
Quadratic: Cochrane <i>et al.</i>	$E[m^2] \leq 1 + A^2$
Exponential	$E[m \ln m] \leq A$
Power, RRA = γ	$E[m^{1-1/\gamma}] \leq (1 + \frac{A}{\gamma})^{1/\gamma-1}$
Logarithmic	$-E[\ln m] \leq \ln(1 + A)$

The apparently simple concept of good-deal bounds has turned out to provide a great deal of richness for mathematicians to analyze, and there are now many variations on this theme in the published literature. Although the theory stems from a practical desire, very few of the papers have an applied flavor. Rather little algorithmic or numerical work has been reported, and most of that uses only somewhat simplified models, seldom calibrated to the market. The good-deal bounds approach could easily be adapted to deal with model risk, something which is hinted at in Cont [7]. The literature needs more real applications, and, perhaps, the balance will have changed when the next survey of this area comes to be written.

References

- [1] Artzner, P., Delbaen, F., Eber, J. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**(3), 203–228.
- [2] Bernardo, A. & Ledoit, O. (1996). Gain, loss and asset pricing, *Journal of Political Economy* **108**(1), 144–172.
- [3] Bjork, T. & Slinko, I. (2006). Towards a general theory of good-deal bounds, *Review of Finance* **10**, 221–260.
- [4] Černý, A. (2003). Generalised sharpe ratios and asset pricing in incomplete markets, *European Finance Review* **7**, 191–233.
- [5] Černý, A. and Hodges, S.D. (2001). The theory of good-deal pricing in financial markets, in *Selected Proceedings of the First Bachalier Congress Held in Paris, 2000*, H. Geman, D. Madan, S.R. Pliska & T. Vorst, eds, Springer Verlag.
- [6] Cochrane, J.H. & Saá-Requejo, J. (2000). Beyond arbitrage: ‘Good-Deal’ asset price bounds in incomplete markets, *Journal of Political Economy* **108**(1), 79–119.
- [7] Cont, R. (2006). Model uncertainty and its impact on the pricing of derivative instruments, *Mathematical Finance* **16**(3), 519–547.
- [8] Dybvig, P.H. & Ross, S.A. (1987). Arbitrage, in *The New Palgrave: A Dictionary of Economics*, J. Eatwell, M. Milgate & P. Newman., eds, Macmillan, London, Vol. 1, pp. 100–106.
- [9] Hansen, L.P. & Jagannathan, R. (1991). Implications of security market data for models of dynamic economies, *Journal of Political Economy* **99**, 225–262.
- [10] Harrison, J. & Kreps, J. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **11**, 215–260.
- [11] Hobson, D.G. (1998). Robust hedging of the lookback option, *Finance and Stochastics* **2**, 329–347.
- [12] Hodges, S.D. (1998). *A Generalization of the Sharpe ratio and its Applications to Valuation Bounds and Risk Measures*. FORC Preprint 1998/88, University of Warwick.
- [13] Jaschke, S. & Küchler, U. (2001). Coherent risk measures and good-deal bounds, *Finance and Stochastics* **5**, 181–200.
- [14] Levy, H. (1985). Upper and lower bounds of put and call option values: stochastic dominance approach, *Journal of Finance* **40**, 1197–1218.
- [15] Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics* **4**, 141–183.
- [16] Perrakis, S. & Ryan, P.J. (1984). Option pricing bounds in discrete time, *Journal of Finance* **39**, 519–525.
- [17] Schweizer, M. (1995). On the minimal martingale measure and the Föllmer-Schweizer decomposition, *Stochastic Analysis and its Applications* **13**, 573–599.
- [18] Sharpe, W.F. (1994). The sharpe ratio, *Journal of Portfolio Management* **21**, 49–59.
- [19] Staum, J. (2004). Pricing and hedging in incomplete markets: fundamental theorems and robust utility maximization, *Mathematical Finance* **14**(2), 141–161.

Related Articles

Arbitrage Strategy; Convex Risk Measures; Stochastic Discount Factors; Sharpe Ratio; Superhedging; Utility Function.

STEWART D. HODGES

Arrow–Debreu Prices

Arrow–Debreu prices are the prices of “atomic” time and state-contingent claims, which deliver one unit of a specific consumption good if a specific uncertain state realizes at a specific future date. For instance, claims on the good “ice cream tomorrow” are split into different commodities depending on whether the weather will be good or bad, so that good-weather and bad-weather ice cream tomorrow can be traded separately. Such claims were introduced by Arrow and Debreu in their work on general equilibrium theory under uncertainty, to allow agents to exchange state and time contingent claims on goods. Thereby the general equilibrium problem with uncertainty can be reduced to a conventional one without uncertainty. In finite-state financial models, Arrow–Debreu securities delivering one unit of the numeraire good can be viewed as natural atomic building blocks for all other state–time contingent financial claims; their prices determine a unique arbitrage-free price system.

Arrow–Debreu Equilibrium Prices

This section explains Arrow–Debreu prices in an equilibrium context, where they originated, see [1, 3]. We first consider a single-period model with uncertain states that will be extended to multiple periods later. For this exposition, we restrict ourselves to a single consumption good only, and consider a pure exchange economy without production.

Let $(\Omega, \mathcal{F})$ be a measurable space of finitely many outcomes $\omega \in \Omega = \{1, 2, \dots, m\}$, where the σ -field $\mathcal{F} = 2^\Omega$ is the power set of all events $A \subset \Omega$. There is a finite set of agents, each seeking to maximize the utility $u^a(c^a)$ from his or her consumption $c^a = (c_0^a, c_1^a(\omega)_{\omega \in \Omega})$ at present and future dates 0 and 1, given some endowment that is denoted by a vector $(e_0^a, e_1^a(\omega)) \in \mathbb{R}_{++}^{1+m}$. For simplicity, let consumption preferences of agent a be of the expected utility form

$$u^a(c^a) = U_0^a(c_0) + \sum_{\omega=1}^m P^a(\omega) U_\omega^a(c_1(\omega)) \quad (1)$$

where $P^a(\omega) > 0$ are subjective probability weights, and the direct utility functions U_ω^a and U_0^a are, for present purposes, taken to be of the form $U_i^a(c) =$

$d_i^a c^\gamma / \gamma$ with relative risk aversion coefficient $\gamma = \gamma^a \in (0, 1)$ and discount factors $d_i^a > 0$. This example for preferences satisfies the general requirements (insaturation, continuity and convexity) on preferences for state-contingent consumption in [3], which need not be of the separable subjective expected utility form above. The only way for agents to allocate their consumption is by exchanging state-contingent claims for the delivery of some units of the (perishable) consumption good at a specific future state. Let q_ω denote the price at time 0 for the state-contingent claim that pays $q_0 > 0$ units if and only if state $\omega \in \Omega$ is realized. Given the endowments and utility preferences of the agents, an equilibrium is given by consumption allocations c^{a*} and a linear price system $(q_\omega)_{\omega \in \Omega} \in \mathbb{R}_+^m$ such that,

1. for any agent a , his or her consumption c^{a*} maximizes $u^a(c^a)$ over all c^a subject to budget constraint $(c_0^a - e_0^a)q_0 + \sum_\omega (c_1^a - e_1^a)(\omega)q_\omega \leq 0$, and
2. markets clear, that is $\sum_a (c_t^a - e_t^a)(\omega) = 0$ for all dates $t = 0, 1$ and states ω .

An equilibrium exists and yields a *Pareto optimal* allocation; see [3], Chapter 7, or the references below. Relative equilibrium prices q_ω/q_0 of the Arrow securities are determined by first-order conditions from the ratio of marginal utilities evaluated at optimal consumption: for any a ,

$$\frac{q_\omega}{q_0} = P^a(\omega) \frac{\partial}{\partial c_1^a} U_\omega^a(c_1^{a*}(\omega)) \Big/ \frac{\partial}{\partial c_0^a} U_0^a(c_0^{a*}) \quad (2)$$

To demonstrate the existence of equilibrium, the classical approach is to show that excess demand vanishes, that is markets clear, by using a fixed point argument (see Chapter 17 in [8]). To this end, it is convenient to consider c^a , e^a and $q = (q_0, q_1, \dots, q_m)$ as vectors in $\mathbb{R}^{1+m}$. Since only relative prices matter, we may and shall suppose that prices are normalized so that $\sum_0^m q_i = 1$, that is, the vector q lies in the unit simplex $\Delta = \{q \in \mathbb{R}_+^{1+m} \mid \sum_0^m q_i = 1\}$. The budget condition 1 then reads compactly as $(c^a - e^a)q \leq 0$, where the left-hand side is the inner product in $\mathbb{R}^{1+m}$. For given prices q , the optimal consumption of agent a is given by the inverse of the marginal utility, evaluated at a multiple of the state price density (see equation (12) for the general definition in the multiperiod case), as

$$c_0^{a*} = c_0^{a*}(q) = (U_0^{a'})^{-1}(\lambda^a q_0) \text{ and}$$

$$c_{1,\omega}^{a*} = c_{1,\omega}^{a*}(q) = (U_\omega^{a'})^{-1}(\lambda^a q_\omega / P^a(\omega)), \quad \omega \in \Omega \quad (3)$$

where $\lambda^a = \lambda^a(q) > 0$ is determined by the budget constraint $(c^{a*} - e^a)q = 0$ as the Lagrange multiplier associated to the constrained optimization problem 1. Equilibrium is attained at prices q^* where the aggregate excess demand

$$z(q) := \sum_a (c^{a*}(q) - e^a) \quad (4)$$

vanishes, that is $z(q^*) = 0$. One can check that $z : \Delta \rightarrow \mathbb{R}^{1+m}$ is continuous in the (relative) interior $\Delta^{\text{int}} := \Delta \cap \mathbb{R}_{++}^{1+m}$ of the simplex, and that $|z(q^n)|$ goes to ∞ when q^n tends to a point on the boundary of Δ . Since each agent exhausts his or her budget constraint 1. with equality, Walras' law $z(q)q = 0$ holds for any $q \in \Delta^{\text{int}}$. Let Δ^n be an increasing sequence of compact sets exhausting the simplex interior: $\Delta^{\text{int}} = \bigcup_n \Delta^n$. Set $v^n(z) := \{q \in \Delta^n \mid zq \geq zp \ \forall p \in \Delta^n\}$, and consider the correspondence (a multivalued mapping)

$$\Pi^n : (q, z) \mapsto (v^n(z), z(q)) \quad (5)$$

that can be shown to be convex, nonempty valued, and maps the compact convex set $\Delta^n \times z(\Delta^n)$ into itself. Hence, by Kakutani's fixed point theorem, it has a fixed point $(q^{n*}, z^{n*}) \in \Pi^n(q^{n*}, z^{n*})$. This implies that

$$z(q^{n*})q \leq z(q^{n*})q^{n*} = 0 \quad \text{for all } q \in \Delta^n \quad (6)$$

using Walras' law. A subsequence of q^{n*} converges to a limit $q^* \in \Delta$. Provided one can show that q^* is in the interior simplex Δ^{int} , existence of equilibrium follows. Indeed, it follows that $z(q^*)q \leq 0$ for all $q \in \Delta^{\text{int}}$, implying that $z(q^*) = 0$ since $z(q^*)q^* = 0$ by Walras' law. To show that any limit point of q^{n*} is indeed in Δ^{int} , it suffices to show that $|z(q^{n*})|$ is bounded in n , recalling that z explodes at the simplex boundary. Indeed, $z = \sum_a z^a$ is bounded from below since each agent's excess demand satisfies $z^a = c^a - e^a \geq -e^a$. This lower bound implies also an upper bound, by using equation (6) applied with some $q \in \Delta^1 \subset \Delta^n$, since $0 < \epsilon \leq q_i \leq 1$ uniformly in i . This establishes existence of equilibrium. To ensure uniqueness of equilibrium, a sufficient condition is that all agents' risk aversions are less than or equal to 1, that is $\gamma^a \in (0, 1]$ for all a , see [2].

For multiple consumption goods, the above ideas generalize if one considers consumption bundles and state-contingent claims of every good. Arrow [1] showed that in the case of multiple consumption goods, all possible consumption allocations are spanned if agents could trade as securities solely state-contingent claims on the unit of account (so-called *Arrow securities*), provided that spot markets with anticipated prices for all other goods exists in all future states. In the sequel, we only deal with Arrow securities in financial models with a single numeraire good that serves as unit of account, and could for simplicity be considered as money ("euro"). If the set of outcomes Ω were (uncountably) infinite, the natural notion of atomic securities is lost, although a state price density (stochastic discount factor, deflator) may still exist, which could be interpreted intuitively as an Arrow–Debreu state price per unit probability.

Multiple Period Extension and No-arbitrage Implications

The one-period setting with finitely many states is easily extended to finitely many periods with dates $t \in \{0, \dots, T\}$ by considering an enlarged state space of suitable date–event pairs (see Chapter 7 in [3]). To this end, it is mathematically convenient to describe the information flow by a filtration $(\mathcal{F}_t)$ that is generated by a stochastic process $X = (X_t(\omega))_{0 \leq t \leq T}$ (abstract, at this stage) on the finite probability space $(\Omega, \mathcal{F}, P_0)$. Let $\mathcal{F}_0$ be trivial, $\mathcal{F}_T = \mathcal{F} = 2^\Omega$, and assume $P_0(\{\omega\}) > 0$, $\omega \in \Omega$. The σ -field $\mathcal{F}_t$ contains all events that are based on information from observing paths of X up to time t , and is defined by a partition of Ω . The smallest nonempty events in $\mathcal{F}_t$ are “ t -atomic” events $A \in \mathcal{A}_t$ of the type $A = [x_0 \cdots x_t] := \{X_0 = x_0, \dots, X_t = x_t\}$, and constitute a partition of Ω . Figure 1 illustrates the partitions $\mathcal{A}_t$ corresponding to the filtration $(\mathcal{F}_t)_{t=0,1,2}$ in a five-element space Ω , as generated by a process X_t taking values $a, \dots, f$. It shows that a filtration can be represented by a nonrecombining tree. There are eight (atomic) date–event pairs (t, A) , $A \in \mathcal{A}_t$. An adapted process $(c_t)_{t \geq 0}$, describing, for instance, a consumption allocation, has the property that c_t is constant on each atom A of partition $\mathcal{A}_t$, and hence is determined by specifying its value $c_t(A)$ at each node point $A \in \mathcal{A}_t$ of the tree. Arrow–Debreu prices

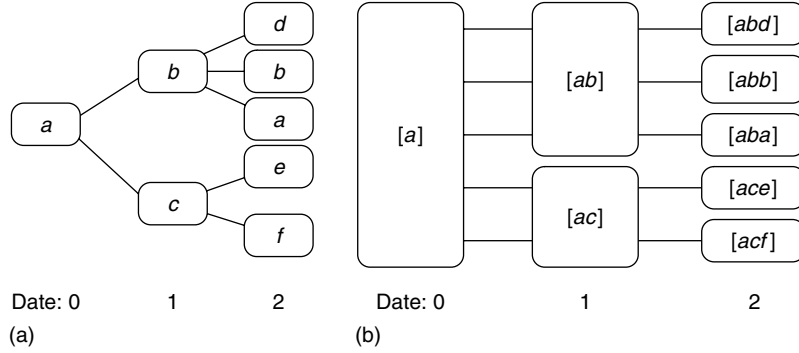


Figure 1 Equivalent representations of the multiperiod case. (a) Tree of the filtration generating process X_t . (b) Partitions $\mathcal{A}_t$ of filtration $(\mathcal{F}_t)_{t=0,1,2}$

$q(t, A)$ are specified for each node of the tree and represent the value at time 0 of one unit of account at date t in node $A \in \mathcal{A}_t$.

Technically, this is easily embedded in the previous single-period setting by passing to an extended space $\Omega' := \{1, \dots, T\} \times \Omega$ with σ -field $\mathcal{F}'$ generated by all sets $\{t\} \times A$ with A being an (atomic) event of $\mathcal{F}_t$, and $P'_0(\{t\} \times A) := \mu(t)P_0(A)$ for a (strictly positive) probability measure μ on $\{1, \dots, T\}$.

For the common no-arbitrage pricing approach in finance, the focus is to price contingent claims solely in relation to prices of other claims, that are taken as exogenously given. In doing so, the aim of the model shifts from the fundamental economic equilibrium task to explain all prices, toward a “financial engineering” task to determine prices from already given other prices solely by no-arbitrage conditions, which are a necessary prerequisite for equilibrium. From this point of view, the (atomic) Arrow–Debreu securities span a *complete market*, as every contingent payoff c , paying $c_t(A)$ at time t in atomic event $A \in \mathcal{A}_t$, can be decomposed by $c = \sum_{t, A \in \mathcal{A}_t} c_t(A)1_{(t, A)}$ into a portfolio of atomic Arrow securities, paying one euro at date t in event A . Hence the no-arbitrage price of the claim must be $\sum_{t, A \in \mathcal{A}_t} c_t(A)q(t, A)$. Given that all atomic Arrow–Debreu securities are traded at initial time 0, the market is statically complete in that any state-contingent cash flow c can be *replicated* by a portfolio of Arrow–Debreu securities that is formed *statically* at initial time, without the need for any dynamic trading. The no-arbitrage price for c simply equals the cost of replication by Arrow–Debreu securities. It is easy to check that, if all prices are determined

like this and trading takes place only at time 0, the market is free of arbitrage, given all Arrow–Debreu prices are strictly positive. To give some examples, the price at time 0 of a zero coupon bond paying one euro at date t equals $ZCB^t = \sum_{A \in \mathcal{A}_t} q(t, A)$. For the absence of arbitrage, the *t-forward prices* $q^f(t, A')$, $A' \in \mathcal{A}_t$, must be related to *spot prices* of Arrow–Debreu securities by

$$\begin{aligned} q^f(t, A') &= \frac{q(t, A')}{\sum_{A \in \mathcal{A}_t} q(t, A)} \\ &= \frac{1}{ZCB^t} q(t, A'), \quad A' \in \mathcal{A}_t \end{aligned} \quad (7)$$

Hence, the forward prices $q^f(t, A')$ are normalized Arrow–Debreu prices and constitute a probability measure Q^t on $\mathcal{F}_t$, which is the *t-forward measure* associated to the t – ZCB as numeraire, and yields $q^f(t, A) = E^t[1_A]$ for $A \in \mathcal{A}_t$, with E^t denoting expectation under Q^t . Below, we also consider “non-atomic” state-contingent claims with payoffs $c_k(\omega) = 1_{(t, B)}(k, \omega)$, $k \leq T$, for $B \in \mathcal{F}_t$, whose Arrow–Debreu prices are denoted by $q(t, B) = \sum_{A \in \mathcal{A}_t, A \subset B} q(t, A)$.

Arrow–Debreu Prices in Dynamic Arbitrage-free Markets

In the above setting, information is revealed dynamically over time, but trading decisions are static in that they are entirely made at initial time 0. To discuss relations between initial and intertemporal

Arrow–Debreu prices in arbitrage-free models with dynamic trading, this section extends the above setting, assuming that all Arrow–Debreu securities are tradable dynamically over time.

Let $q_s(t, A_t)$, $s \leq t$, denote the price process of the Arrow–Debreu security paying one euro at t in state $A_t \in \mathcal{A}_t$. At maturity t , $q_t(t, A_t) = 1_{A_t}$ takes value 1 on A_t and is 0 otherwise. For the absence of arbitrage, it is clearly necessary that Arrow–Debreu prices are nonnegative, and that $q_s(t, A_t)(A_s) > 0$ holds for $s < t$ at $A_s \in \mathcal{A}_s$ if and only if $A_s \supset A_t$. Further, for $s < t$ it must hold that

$$q_s(t, A_t)(A_s) = q_s(s+1, A_{s+1})(A_s) \times q_{s+1}(t, A_t)(A_{s+1}) \quad (8)$$

for $A_s \in \mathcal{A}_s$, $A_{s+1} \in \mathcal{A}_{s+1}$ such that $A_s \supset A_{s+1} \supset A_t$. In fact, the above conditions are also sufficient to ensure that the market model is free of arbitrage: At any date t , the Arrow–Debreu prices for the next date define the interest rate R_{t+1} for the next period $(t, t+1)$ of length $\Delta t > 0$ of a savings account $B_t = \exp(\sum_{s=1}^t R_s \Delta t)$ by

$$\begin{aligned} & \exp(-R_{t+1}(A_t)\Delta t) \\ &= \sum_{A_{t+1} \in \mathcal{A}_{t+1}(A_t), A_{t+1} \subset A_t} q_t(t+1, A_{t+1}), \\ & \text{for } A_t \in \mathcal{A}_t \end{aligned} \quad (9)$$

which is locally riskless, in that R_{t+1} is known at time t , that means it is $\mathcal{F}_t$ -measurable. They define an equivalent *risk neutral probability* Q^B by determining its transition probabilities from any $A_t \in \mathcal{A}_t$ to $A_{t+1} \in \mathcal{A}_{t+1}$ with $A_t \supset A_{t+1}$ as

$$Q^B(A_{t+1}|A_t) = \frac{q_t(t+1, A_{t+1})(A_t)}{\sum_{A \in \mathcal{A}_{t+1}, A \subset A_t} q_t(t+1, A)(A_t)} \quad (10)$$

The transition probability (10) can be interpreted as one-period forward price when being at A_t at time t , for one euro at date $t+1$ in event A_{t+1} , cf. (7). Since all B -discounted Arrow–Debreu price processes $q_s(t, A_t)/B_s$, $s \leq t$, are martingales under Q^B thanks to equations (8, 10), the model is free of arbitrage by the fundamental theorem of asset pricing, see [6]. For initial Arrow–Debreu prices, denoted

by $q_0(t, A_t) \equiv q(t, A_t)$, the martingale property and equations (8, 10) imply that

$$q(t+1, A_{t+1}) = e^{-R_{t+1}(A_t)\Delta t} Q^B(A_{t+1}|A_t)q(t, A_t) \quad (11)$$

Hence $q(t, A_t) = Q^B(A_t)/B_t(A_t)$ for $A_t \in \mathcal{A}_t$.

The *deflator* or *state price density* for agent a is the adapted process ζ_t^a defined by

$$\zeta_t^a(A_t) := \frac{q(t, A_t)}{P^a(A_t)} = \frac{Q^B(A_t)}{B_t(A_t)P^a(A_t)}, \quad A_t \in \mathcal{A}_t \quad (12)$$

so that $\zeta_t^a S_t$ is a P^a -martingale for any security price process S , e.g. $S_t = q_t(T, A_T)$, $t \leq T$, with $A_T \in \mathcal{A}_T$. If one chooses, instead of B_t , another security $N_t = \sum_{A_T \in \mathcal{A}_T} N_T(A_T)q_t(T, A_T)$ with $N_T > 0$ as the numeraire asset for discounting, one can define an equivalent measure Q^N by

$$\frac{Q^N(A)}{P^a(A)} = \frac{N_T(A)}{N_0(A)} \zeta_T^a(A), \quad A \in \mathcal{A}_T \quad (13)$$

which has the property that S_t/N_t is a Q^N -martingale for any security price process S . Taking $N = (ZCB_t^T)_{t \leq T}$ as the T -zero-coupon bond yields the T -forward measure Q^T .

If X is a Q^B -Markov process, the conditional probability $Q^B(A_{t+1}|A_t)$ in equation (11) is a transition probability $p_t(x_{t+1}|x_t) := Q^B(X_{t+1} = x_{t+1}|X_t = x_t)$, where $A_k = [x_1 \dots x_k]$ for $k = t, t+1$. By summation of suitable atomic events

$$\begin{aligned} & q(t+1, X_{t+1} = x_{t+1}) \\ &= \sum_{x_t} e^{-R(x_t)\Delta t} p_t(x_{t+1}|x_t)q(t, X_t = x_t) \end{aligned} \quad (14)$$

where the sum is over all x_t from the range of X_t .

Application Examples: Calibration of Pricing Models

The role of Arrow–Debreu securities as “atomic building blocks” is theoretical, in that there exist no corresponding securities in real financial markets. Nonetheless, they are of practical use in the calibration of pricing models. For this section, X is taken to be a Q^B -Markov process, possibly time-inhomogeneous.

The first example concerns the calibration of a short rate model to some given term structure of zero coupon prices $(ZCB_t)_{t \leq T}$, implied by market quotes. For such models, a common calibration procedure relies on a suitable time-dependent shift of the state space for the short rate (see [7], Chapter 28.7). Let suitable functions R_t^* be given such that the variations of $R_t^*(X_t)$ already reflect the desired volatility and mean-reversion behavior of the (discretized) short rate. Making an ansatz $R_t(X_t) := R_t^*(X_t) + \alpha_t$ for the short rate, the calibration task is to determine the parameters α_t , $1 \leq t \leq T$, such that

$$ZCB_t = E^B \left[\exp \left(- \sum_{k \leq t} (R_k^*(X_k) + \alpha_k) \Delta t \right) \right] \quad (15)$$

with the expectation being taken under the risk neutral measure Q^B . It is obvious that this determines all the α_t uniquely. When computing this expectation to obtain the α_t by forward induction, it is efficient to use Arrow–Debreu prices $q(t, X_t = x_t)$, since X usually can be implemented by a recombining tree. Summing over the range of states x_t of X_t is more efficient than summing over all paths of X . Suppose that α_k , $k \leq t$, and $q(t, X_t = x_t)$ for all values x_t have been computed already. Using equation (14), one can then compute α_{t+1} from equation

$$ZCB_{t+1} = \sum_{x_{t+1}} \sum_{x_t} q(t, X_t = x_t) \times e^{(R_{t+1}^*(x_t) + \alpha_{t+1}) \Delta t} p_t(x_{t+1} | x_t) \quad (16)$$

where the number of summand in the double sum is typically bounded or grows at most linearly in t . Then Arrow–Debreu prices $q(t+1, X_{t+1} = x_{t+1})$ for the next date $t+1$ are computed using equation (14), while those for t can be discarded.

The second example concerns the calibration to an implied volatility surface. Let X denote the discounted stock price $X_t = S_t \exp(-rt)$ in a trinomial tree model with constant interest $R_t := r$ and $\Delta t = 1$. Each X_{t+1}/X_t can attain three possible values $\{m, u, d\} := \{1, e^{\pm \sigma \sqrt{2}}\}$ with positive probability, for $\sigma > 0$. The example is motivated by the task to calibrate the model to given prices of European calls and puts by a suitable choice of the (nonunique) risk neutral Markov transition probabilities for X_t . We focus here on the main step for this task, which is to show that the Arrow–Debreu

prices of all state-contingent claims, which pay one unit at some t if $X_t = x_t$ for some x_t , already determine the risk neutral transition probabilities of X . It is easy to see that these prices are determined by those of calls and puts for sufficiently many strikes and maturities. Indeed, strikes at all tree levels of the stock for each maturity date t are sufficient, since Arrow–Debreu payoffs are equal to those of suitable butterfly options that are combinations of such calls and puts. From given Arrow–Debreu prices $q(t, X_t = x_t)$ for all t, x_t , the transition probabilities $p_t(x_{t+1} | x_t)$ are computed as follows: starting from the highest stock level x_t at some date t , one obtains $p_t(x_t u | x_t)$ by equation (14) with $R_t(x_t) = r$ and $\Delta t = 1$. The remaining transition probabilities $p_t(x_t m | x_t)$, $p_t(x_t d | x_t)$ from (t, x_t) are determined from

$$p_t(x_t u | x_t)u + p_t(x_t m | x_t)m + p_t(x_t d | x_t)d = 1 \quad (17)$$

and $p_t(x_t u | x_t) + p_t(x_t m | x_t) + p_t(x_t d | x_t) = 1$. Using these results, the transition probabilities from the second highest (and subsequent) stock level(s) are implied by equation (14) in a similar way. This yields all transition probabilities for any t .

To apply this in practice, the call and put prices for the maturities and strikes required would be obtained from real market quotes, using suitable interpolation, and the trinomial state space (i.e., $\sigma, r, \Delta t$) has to be chosen appropriately to ensure positivity of all p_t , see [4, 5].

References

- [1] Arrow, K.J. (1964). The role of securities in the optimal allocation of risk-bearing, As translated and reprinted in 1964, *Review of Financial Studies* **31**, 91–96.
- [2] Dana, R.A. (1993). Existence and uniqueness of equilibria when preferences are additively separable, *Econometrica* **61**, 953–957.
- [3] Debreu, G. (1959). *Theory of Value: An Axiomatic Analysis of Economic Equilibrium*, Yale University Press, New Haven.
- [4] Derman, E., Kani, I. & Chriss, N. (1996). Implied trinomial trees of the volatility smile, *Journal of Derivatives* **3**, 7–22.
- [5] Dupire, B. (1997). Pricing and hedging with smiles, in *Mathematics of Derivative Securities*, M.A.H. Dempster & S.R. Pliska, eds, Cambridge University Press, Cambridge, pp. 227–254.

- [6] Harrison, J. & Kreps, D. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- [7] Hull, J. (2006). *Options, Futures and Other Derivative Securities*, Prentice Hall, Upper Saddle River, New Jersey.
- [8] Mas-Colell, A., Whinston, M.D. & Green, J.R. (1995). *Microeconomic Theory*, Oxford University Press, Oxford.

Related Articles

Arrow, Kenneth; Complete Markets; Dupire Equation; Fundamental Theorem of Asset Pricing; Model Calibration; Pricing Kernels; Risk-neutral Pricing; Stochastic Discount Factors.

DIRK BECHERER & MARK H.A. DAVIS

Options: Basic Definitions

A financial *option* is a contract conferring on the holder the right, but not the obligation, to engage in some transaction, on precisely specified terms, at some time in the future. When the holder does decide to engage in the transaction, he/she is said to *exercise the option*. There are two parties to an option contract, generally known as the *writer* and the *buyer* or *holder*. The “optionality” accrues to the buyer; the writer has the obligation to carry out his/her side of the transaction, should the buyer exercise the option. An option is *vulnerable* if there is considered to be nonnegligible risk that the writer will fail to do this, that is, he/she will default on his/her side of the contract.

The classic contracts are European call and put options. A **call option** entitles the holder to purchase a specified number N of units of a security at a fixed price K per unit, at a specified time T . If S_T is the market price of the security at time T , the holder will exercise the option if and only if $S_T > K$ (otherwise, it would be cheaper to buy in the market). Since the holder has now acquired for K per unit something that is worth S_T , he/she makes a profit of $N(S_T - K)$. Thus, in general, the profit is $N[S_T - K]^+ = N \max(S_T - K, 0)$. Similarly, the profit on a contract, entitling the holder to *sell* at K , is $N[K - S_T]^+$. The asset on which the option is written is called the *underlying asset*. A call option is *in the money* if $S_T > K$, *at the money* (ATM) if $S_T = K$ and *out of the money* if $S_T < K$. These terms are also used at earlier times $t < T$ (e.g., ATM if $S_t = K$ etc.) even though the option cannot be exercised then. An option is *ATM forward* at time t if $K = F(t, T)$ where $F(t, T)$ is the **forward price** at time t for purchase at time T .

Option Contracts

In general, several assets may underlie a given contract, as for example in an **exchange option** where the holder has the right to exchange one asset for another. Options are sometimes called *contingent claims* or *derivative securities*. These two synonymous terms include options, but refer more generally to any contract whose value is a function of the values of some collection of underlying assets.

There are several binary classifications that help define an option contract.

European/American

An option is European if it must be exercised, if at all, on a specified date, or a specified sequence of dates (for example, a *cap* is an **interest-rate option** consisting of a sequence of call options on the Libor rate; each of these options is exercised when they fall due if in the money). In an **American option**, by contrast, the time of exercise is at the holder's discretion. The classic American option involves a fixed final time T and allows the holder to exercise at any time $T' \leq T$, leaving the holder with the problem of determining an exercise strategy that will maximize the value to him. This immediately implies three things: (i) the value of an American option that has not already been exercised at any time t can never be less than the *intrinsic value* ($[K - S_t]^+$ for a put option), since one possible strategy is always to exercise *now*, (ii) the value can never be less than the value of the corresponding European option, since another possible strategy is never to exercise before T , and (iii) the value is a nondecreasing function of the final maturity time, since for $T_1 < T_2$ the T_1 option is just the T_2 option with the additional restriction that it never be exercised beyond T_1 . The difference between the American and European values for the same contract is called the *early exercise premium*. Sometimes, American options have some restriction on the set of allowable exercise times; for example, the conversion option in **convertible bonds** often prohibits the investor from converting before a certain minimum time. Of course, any such restrictions reduce the value since they reduce the class of exercise strategies. A particular case is the *Bermuda option*, which can only be exercised at one of a finite number of times. This is the normal situation in interest-rate options, where there is a natural sequence of *coupon dates* at which exercise decisions may be taken. Bermuda options are presumably so called because Bermuda is somewhere between America and Europe—but much closer to America.

Traded/OTC

Traded options are those where the parties trade through the medium of an organized exchange,

while “over the counter” (OTC) options are bilateral agreements between market counterparties. Option exchanges have become increasingly globalized in recent years. They include the US–European consortium NYSE Euronext, Chicago Mercantile Exchange (CME), Eurex and EDX, all of which offer a range of financial contracts, and a number of specialist commodity exchanges such as NYMEX (oil), ICE, and the London Metal Exchange (LME). An exchange offers contracts on an underlying asset such as an individual stock or a stock index such as the S&P500, with a range of maturity times and strike values. New options are added as the old ones roll off, and the strikes offered are in a range around the spot price of the underlying asset at the time the contract is initiated (the options may turn out to be far in or out of the money at later times, of course). In a traded options market, prices are determined by supply and demand. If the exercise times are T_i and the strike values K_j then the matrix $V = [\hat{\sigma}_{ij}]$, where $\hat{\sigma}_{ij}$ is the **implied volatility** corresponding to the (T_i, K_j) contract, defines the so-called **volatility surface** that plays a key role in option risk management.

All interest-rate options and most FX (foreign exchange) options are OTC, but many are, nevertheless, very liquidly traded and market information on implied volatilities is readily available.

Physical Settlement/Cash Settlement

Many single-stock options, and **commodity options** are *physically settled*, that is, at exercise the holder pays the strike value and takes delivery of a share certificate or a barrel of oil. (One can, however, avoid physical delivery by selling the option shortly before final maturity.) The alternative is *cash settlement*, where the holder is simply paid a cash amount, such as $[S_T - K]^+$ for a call option, at exercise. When the underlying is an index like the S&P500 this is the only way—one cannot deliver the index! In this case, the amount paid is $c \times [I_T - K]^+$ where I_T is the value of the index and c is the contractually specified dollar value of one index point.

Liquid/Illiquid

Like any other traded asset, an option contract is **liquid** if there is large market depth, that is, there are a significant number of active traders in the market,

none of whom controls a significant proportion of the total supply. In these circumstances, the price is well established, since the last trade was never very long ago; bid/ask spreads will be tight, buyers and sellers can enter the market at will, and there is little room for price manipulation. By contrast, in an illiquid market, it may be hard to establish a market price when actual trades are infrequent and bid/ask spreads are wide. The liquid/illiquid classification is not immutable: a liquid market can suddenly become illiquid if there is some shock that forces everybody onto the same side of the market. Several well-recorded disasters in the derivatives market have been due to this phenomenon.

(Plain) Vanilla/Exotic

The simplest, most standard, and most widely traded options are often referred to as *plain vanilla* options. This would certainly include all exchange-traded options. An “exotic” option is an OTC option with nonstandard features of some kind, which requires significant modeling effort to value, and where different analysts could well come up with significantly different valuations. Exotic options often involve several underlying assets and complicated payment streams, but even a simple call option can be exotic if it poses significant hedging difficulties, as for example do long-dated equity options. On the other hand, barrier options, for example, which once would have been considered exotic, have now become vanilla in some markets such as FX, because they are so widely traded.

Path-dependent/Path-independent

An option is path-dependent if its exercise value depends on the value of the underlying asset at more than one time. Examples are barrier and Asian options, and any American option. The exercise value of a path-independent option is a function only of the underlying price, say S_T , at the maturity time T , as for example in **Black–Scholes**. Valuation then only requires specifying the one-dimensional risk-neutral distribution of S_T , whereas for a path-dependent option a distribution in path space is required, making the valuation more computationally intensive and model dependent.

Option Definitions

The purpose of this section is to collect together introductory definitions of various option contracts, or features of option contracts, found in the market. We refer to specialist articles in this encyclopaedia for a detailed treatment, and to standard textbooks such as Hull [1]. In these definitions, we use S_t to denote a generic underlying asset price, on which is written an option that starts at time 0 with final maturity at time T .

Asian

An **Asian option** is one whose exercise value depends on the *average price* over some range of times. Generally, this is an arithmetic average of the form $\bar{S} = (1/n) \sum_{i=1}^n S_{t_i}$ and, of course, in reality, it is always a finite sum, although for purposes of analysis it is often convenient to consider continuous averaging $\bar{S}^c = (1/T) \int_0^T S_t dt$. Averaging may be over the entire length of the contract or some much shorter period. For example, in **commodity options** call option values are generally based on the 10-day or one-month average price immediately prior to maturity, rather than on the spot price, to deter market manipulation.

Barrier

A **barrier option** involves one or both of two prices, a lower barrier $L < S_0$ and an upper barrier $U > S_0$, a strike K , and a maturity time T . Let $\tau_L = \inf\{t : S_t \leq L\}$, $\tau_U = \inf\{t : S_t \geq U\}$ and $\tau_{LU} = \min(\tau_L, \tau_U)$. A *knockout* option expires worthless if a specified one of these times occurs before T , while a *knock-in* option expires worthless *unless* this time occurs before T . An *up-and-out* call is a knockout call option based on τ_U , so formally its exercise value is $\mathbf{1}_{(\tau_U > T)}[S_T - K]^+$. Similarly, an up-and-in call has exercise value $\mathbf{1}_{(\tau_U \leq T)}[S_T - K]^+$. The sum is an ordinary call option. There are analogous definitions for down-and-out and down-and-in options based on τ_L . Normally, these would be put options. A double barrier option knocks out or in at time τ_{LU} . In the Black–Scholes model, there is an analytic formula for single-barrier options, based on the reflection principle for Brownian motion, but double barrier options require numerical methods.

Knockout options are cheaper than their plain vanilla counterparts because the exercise value is strictly less with positive probability. Essentially, the buyer of a vanilla option pays a premium for events he/she may regard as overwhelmingly unlikely. By buying a barrier option instead, he/she avoids paying this premium.

Basket

Consider a portfolio containing w_i units of asset i for $i = 1, \dots, n$. A **basket** call option then has exercise value $[X - K]^+$ where $X = \sum_i w_i S_i(T)$ is the portfolio value at time T . The main problem in valuing basket options is the enormous number of correlation coefficients involved, for even moderate portfolio size n .

Bermuda

These were already mentioned above. Probably the most common example is the **Bermuda swaption**, entitling the holder to enter a swap at an agreed fixed-side rate at any one of a list of coupon dates (or, equivalently, the right to walk away from an existing swap contract).

Chooser

A chooser option involves three times 0, T_1 , T_2 , and a strike K . The option is entered and the strike set at time 0, and at time T_1 , the holder selects whether it is to be a put or a call. The appropriate exercise value is then evaluated and paid at T_2 . Thus the value at T_1 is the maximum of the put and call values at that time. Given this fact valuation is straightforward in the Black–Scholes model.

Digital

A digital option pays a fixed cash amount if some condition is realized. For example, an up-and-in digital barrier option with maturity T and barrier level U will pay a fixed amount X if $\tau_U < T$. Payment might be made at τ_U or at T .

Exchange

An **exchange option** has exercise value $[aS_2(T) - S_1(T)]^+$, that is, the holder has the right to exchange

one unit of asset 1 for a units of asset 2 at the maturity time T . Exchange options can be priced by the Margrabe formula, originally introduced in [2].

Forward-starts, Ratchets, and Cliquets

A involves two times $0 < T_1 < T_2$. The premium is paid at time 0 and the exercise value at final maturity T_2 is $[S_{T_2} - mS_{T_1}]^+$, where m is a contractually specified “moneyness” factor. If, for example, $m = 1.05$, then effectively the strike is set at T_1 in such a way that the option is 5% out of the money at that time. This is a pure volatility play in that the value essentially only depends on the forward volatility between T_1 and T_2 . A ratchet or **cliquet** is a string of forward start options over times T_i, T_{i+1} , so T_{i+1} is the maturity date for option i and the start date for option $i + 1$.

Lookback

Let $S_{\max}(t) = \max_{u \leq t} S(u)$ and $S_{\min}(t) = \min_{u \leq t} S(u)$. The exercise values of a **lookback** call and a lookback put are $[S(T) - S_{\min}(T)]$ and $[S_{\max}(T) - S(T)]$, respectively. The holders of these options can essentially buy at the minimum price and sell at the maximum price. Black–Scholes valuation of these options uses the reflection principle for Brownian motion, in the same way as for barrier options.

Passport

A passport option is a call option where the underlying asset is a traded portfolio, and the holder has the right to choose the trading strategy in this portfolio.

Quanto

A **quanto** or *cross-currency* option is written on an underlying asset denominated in currency A, but the exercise value is paid in currency B. For example, we could have an option on the USD-denominated S&P index $I(t)$ where the exercise value is GBP $c[I(T) - K]^+$. We can write the constant as $c = c_1 c_2$ where c_1 is the conventional dollar value of an index point, while c_2 is an exchange rate (the number of pounds per dollar). Thus a quanto option is a combination of a “foreign”-denominated option plus an exchange-rate guarantee. Valuation amounts to deriving the state-price density applicable to a market model including foreign as well as domestic assets.

Russian

A Russian option is a perpetual lookback option, that is, an American lookback option with no final maturity time.

References

- [1] Hull, J.C. (2000). *Options, Futures and Other Derivatives*, 4th Edition, Prentice Hall.
- [2] Margrabe, W. (1978). The value of an option to exchange one asset for another, *Journal of Finance* **33**, 177–186.

MARK H.A. DAVIS

Option Pricing: General Principles

Option contracts are financial assets that involve an element of choice for the owner. Depending on an event, the holder of an option contract can exercise his/her options stated in the contract, that is, to undertake certain specified actions. The typical example of an option contract gives the holder the right to buy a specific stock at a contracted price and time in the future. The contracted price is called the *strike price*, whereas the exercise time is when the option may be executed. Such contracts are known as *call options*, and the event that triggers the execution of the option is that the underlying stock price is above the strike. There is a plethora of different options traded in today's modern financial markets, where the financial events may include **credit**, **weather** related situations, and so on. One usually refers to derivatives or claims as being financial assets whose values are dependent on other financial assets.

There are two fundamental questions that the option pricing theory tries to answer. First, what is the fair price of a claim, and second, how can one replicate the claim. The second question immediately implies the answer to the first, since if we can find an investment strategy in the market that replicates the claim, the cost of this replication should be the fair price. This replication strategy is frequently called the **hedging strategy** of the claim. The key financial concepts in pricing and replication are **arbitrage** (or rather the absence of such) and **completeness**. A mathematical concept related to these is the **equivalent martingale measure**, also known as the *risk-neutral probability*.

Explaining the Basic Concepts

To understand the concepts used, it is informative to consider a very simple (and highly unrealistic) one-period **binomial model**. Suppose that we have a stock with value \$100 today and two possible outcomes in one year. Either the stock price can increase to \$110 or it can remain unchanged. The interest rate earned on bank deposits is set to 5% yearly and considered the risk-free investment in the market.

Suppose that we wish to find a fair price of a call option with strike \$105 in one year. This option will effectively pay out \$5 if the stock increases, whereas the holder will not exercise it if the stock value is \$100. Consider now an investment today in $a = 0.5$ number of stocks and $b = -\$50/1.05 \approx -\47.62 deposited in the bank. A simple calculation reveals that this investment yields exactly the same as holding the option. In fact, this is the only investment in the stock and bank that perfectly replicates the option payoff and we, therefore, call it the *replicating strategy of the option*. The cost of replication is $P = \$50/21 \approx \2.38 .

We argue that the fair price of the option should be the same as the costs P of buying the replicating strategy. If the price would be higher, say $\tilde{P} > P$, then one could do the following. Sell n options for that price and buy n of the replicating strategy. At exercise, any claims from the options sold will be covered exactly by the replicating strategies bought. However, we have received the cash amount of $n \times \tilde{P}$ for selling the options and paid out the amount $n \times P$ for replication, thus leaving us with a profit. There is no risk attached with this investment proposition, and we can make the profit arbitrarily high by simply increasing n . This is what is known as an **arbitrage opportunity**, and in **efficient markets**, this should not be possible (or at least be ruled out quickly). If $\tilde{P} < P$, we reverse the positions above to create an arbitrage. The definition of a fair price is the price for which no arbitrage possibility exists. Thus, the option price in our example must be $P = \$50/21 \approx \2.38 .

We note that the probability of a stock price increase did not enter into our analysis. The fair price is unaffected by this probability, since the hedging strategy is the same no matter how likely the stock price is to increase to \$110. The price of the option does not depend on the expected return of the stock, but only on the spread in the two possible outcomes of the stock price at exercise time, or, in other words, the **volatility**.

One may ask if the price of an option can be stated as the present expected value of the payoff at exercise. From the above derivations, we see that this is, in general, not the case since the price is not a function of the probability of a stock price increase. Hence, a present value price of the option would lead to arbitrage possibilities. However, we may rephrase the question and ask whether *there exists* a probability q for price increase such that the fair price can be

expressed as a present value? Letting $q = 0.5$, we can easily convince ourselves that

$$\begin{aligned} P &= \frac{1}{1.05} \{q \times 5 + (1 - q) \times 0\} \\ &= \frac{1}{1.05} \mathbb{E}_q[\text{option payoff}] \end{aligned} \quad (1)$$

where $\mathbb{E}_q$ is the expectation with respect to the probability q . This probability of a stock price increase is *not* the probability for a price increase observed in the market, but a *constructed* probability for which the option price can be expressed as a present expected value.

The probability q has an interesting property that actually defines it. The present expected value of the stock price is equal to today's value,

$$100 = \frac{1}{1.05} \mathbb{E}_q[\text{stock price}] \quad (2)$$

Hence, the discounted stock price is a **martingale** with respect to the probability q . Further, the return on an investment in the stock coincides with the risk-free rate under q , defending the name “risk-neutral probability” often assigned to q .

Option Pricing in Continuous Time

Our binomial one-period example basically contains the main concepts for pricing of options and claims in more general and realistic market models. Moving to a stock price that evolves dynamically in time with stochastic marginal changes, the principles of option pricing remain basically the same, however, introducing interesting technical challenges. We now look at the case when the stock price follows a **geometric Brownian motion** (GBM), that is,

$$\frac{dS(t)}{S(t)} = \mu dt + \sigma dB(t) \quad (3)$$

defined on a probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ with the filtration $\mathcal{F}_t$ generated by the Brownian motion modeling the information flow. The GBM model indicates that returns (or more precisely, logarithmic returns) are independent and normally distributed, with mean μdt and volatility $\sigma\sqrt{dt}$. The model was first proposed for stock price dynamics by **Samuelson** [7] and later used by **Black and Scholes** [1] and **Merton** [6] in their derivation of the

famous **option pricing formula**. We suppose that the market is frictionless in the sense that there are no transaction costs incurred when we trade in the stock or the bank, and there are no restrictions on short or long positions. Further, the interest rate is the same whether we borrow or lend money, and the market is perfectly liquid.

The main difference from the one-period model is that we can invest in the underlying stock at all times up to maturity of the claim. Obviously, we can also do the same with the bank deposit, which is now assumed to yield a continuously compounding interest rate r . An investment strategy will consist of $a(t)$ shares of the stock and $b(t)$ invested in the bank at time t . Since investors cannot foresee the future, the investment decisions at time t can only be based upon the available market information, which is contained in the filtration $\mathcal{F}_t$. The value at time t of the portfolio is

$$V(t) = a(t)S(t) + b(t)R(t) \quad (4)$$

where $R(t) = \exp(rt)$, the value of an initial bank deposit of 1. Further, since we are interested in creating strategies that are replicating an option, we wish to rule out any external funding or withdrawal of money in the portfolio we are setting up. This leads to the so-called **self-financing** hypothesis, saying that any change in portfolio value comes from a change in the underlying stock price and bank deposit. Mathematically, we can formulate this condition as

$$dV(t) = a(t) dS(t) + b(t) dR(t) \quad (5)$$

Note that **Itô's formula** implies a dynamics $V(t)$ where the differentials of $a(t)$ and $b(t)$ appear. The self-financing hypothesis indicates that these differentials are zero.

For the one-period binomial model, we recall the existence of an **equivalent martingale measure** for which the discounted stock price is a martingale. Applying the Girsanov theorem, we find a probability measure $\mathbb{Q}$ **equivalent** to the market probability $\mathbb{P}$, for which the process $W(t)$ with differential

$$dW(t) = \frac{\mu - r}{\sigma} dt + dB(t) \quad (6)$$

is a Brownian motion. By a direct calculation, we find

$$d(e^{-rt} S(t)) = \sigma(e^{-rt} S(t)) dW(t) \quad (7)$$

which is a **martingale** under $\mathbb{Q}$. Furthermore, by discounting the portfolio process $V(t)$ and applying the self-financing hypothesis, we find

$$d(e^{-rt} V(t)) = \sigma a(t)(e^{-rt} S(t)) dW(t) \quad (8)$$

Hence, the discounted portfolio process is also a martingale under $\mathbb{Q}$.

Consider a claim with maturity at time T and a payoff represented by the random variable X , where $\mathcal{F}_T$ is measurable and integrable with respect to $\mathbb{Q}$. The (for the moment unknown) price at time t of the claim is denoted by $P(t)$. Suppose that the discounted price of the claim is a martingale with respect to $\mathbb{Q}$ and that we have a self-financing portfolio consisting of investments in the stock, the bank notes, and the claim. Further, we construct the investment such that the initial price is zero. The discounted value process of this portfolio will then (by the same reasoning as above) give a martingale process under $\mathbb{Q}$, and hence the expectation with respect to $\mathbb{Q}$ of the portfolio value at any future time must be the same as the initial investment, namely, zero. Thus, under $\mathbb{Q}$, there is a positive probability of having a negative portfolio value, which implies by equivalence of $\mathbb{Q}$ with $\mathbb{P}$ that we cannot have any arbitrage opportunities in this market. On the other hand, if the market does not allow for any arbitrage, one can show that $e^{-rt} P(t)$ must be a $\mathbb{Q}$ -martingale. We refer to [3] for the connection between no-arbitrage and existence of equivalent martingale measures. It is a financially reasonable condition to assume that the market is arbitrage free.

By the **martingale representation theorem**, there exists an **adapted** stochastic process $\phi(t)$ such that

$$d(e^{-rt} P(t)) = \phi(t) dW(t) \quad (9)$$

whenever $\exp(-rt)P(t)$ is square-integrable with respect to $\mathbb{Q}$. By defining $a(t) \triangleq \phi(t)/\sigma \exp(-rt)S(t)$ and $b(t) \triangleq \exp(-rt)(P(t) - a(t)S(t))$, the portfolio $V(t)$ given by the investment strategy (a, b) is self-financing. Moreover, $V(T) = P(T) = X$, implying that it is a replicating strategy for the claim. Furthermore the market becomes **complete**, meaning that there exists a replicating strategy for all claims, X being square-integrable with respect to $\mathbb{Q}$.

Now, again appealing to the $\mathbb{Q}$ -martingale property of $e^{-rt} P(t)$, we find by definition that

$$P(t) = e^{-r(T-t)} \mathbb{E}_{\mathbb{Q}}[X | \mathcal{F}_t] \quad (10)$$

Thus, as a natural generalization of the binomial one-period model case, any claim has a price given as the expected present value, where the expectation is taken with respect to the risk-neutral probability. Note that S does not depend on its expected return μ under $\mathbb{Q}$, and therefore the price $P(t)$ is independent of this. The volatility σ is, however, a crucial parameter for the determination of price.

If we let X be the payoff of a call option written on S , one can calculate the conditional expectation in equation (10) to derive the famous **Black–Scholes formula**. Further, the process $\phi(t)$ is in this case explicitly known, and it turns out that the investment strategy $a(t)$ is the derivative of the price $P(t)$ with respect to $S(t)$. This derivative is known as the delta of the call option. Moreover, the strategy given by $a(t)$ is called *delta-hedging*.

Option Pricing in Incomplete Markets

Recall that we have assumed a frictionless market. In practice, transaction costs are normally incurred when buying and selling shares. Hence, since a delta-hedging strategy (a, b) involves incessant trading, it will become infinitely costly if implemented. In addition, there are practical limits to how big a short position we can take (e.g., due to credit limits and collaterals). Theoretically, there exists only *one* replicating strategy since the martingale representation theorem prescribes a unique integrand process $\phi(t)$. Introducing frictions such as transaction costs or short-selling limits in the market rules out the possibility to replicate claims in general, and the market is said to be **incomplete**. We remark that in an incomplete market, there still exists claims that can be replicated, and by the no-arbitrage principle, the price of these is characterized by the cost of replication, as we have argued above. However, a natural question arises: what can we say about pricing and hedging of claims where no replicating strategy exists?

One approach suggests to look at **super- and sub-replicating strategies**. A super(sub-)replicating strategy is a self-financing portfolio of stock and bank deposit, which at least(most) has the same value as the claim at maturity. Letting $P_{\max}$ ($P_{\min}$) be the infimum (supremum) over all prices of such super(sub-)replicating strategies, it follows that any price P in the interval $(P_{\min}, P_{\max})$ is arbitrage

free. Furthermore, any self-financing strategy that costs less than $P_{\max}$ will always have a positive probability of having a value lower than the claim at maturity, and thus full replication is impossible. This leaves the issuer of the claim with some unhedgable risk. An acceptable or fair price of the claim will reflect the compensation the issuer demands for taking on this risk.

A change in the stock price dynamics gives another source of incompleteness in the market. The GBM model is rather unnatural from an empirical point of view, since observed stock price returns on the marketplace are frequently far from being normally distributed nor are they independent. Stock price models including **stochastic volatility** and/or stochastic drivers other than Brownian motion have been proposed. For instance, on the basis of empirics, the returns may be modeled by a heavy-tailed distribution, which gives rise to a **Lévy process** in the geometric dynamics of the stock price. A consequence of such a seemingly innocent change in the structure is that there exists a continuum (in general) of **equivalent martingale measures** $\mathbb{Q}$ such that the discounted stock price is a martingale. The complicating implication of this is the absence of martingale representations when it becomes impossible to find an investment strategy replicating the claim. As for markets with frictions, we have no possibility of replication, but an interval of possible arbitrage-free prices. In addition, in this case, the issuer of the claim needs to accept a certain unhedgable risk.

To price claims in incomplete markets, one must resort to methods that take into account the risk posed on the issuer. Popular approaches include minimal-variance hedging, where the strategy minimizing the variance (that is, the risk) is sought for. The price of the claim is the cost of buying the minimal-variance strategy [8] plus a compensation for the unhedged risk. Another possibility that has gained a lot of attention in the option pricing literature is **indifference pricing** (see also the seminal work of Hodges and Neuberger [5]). Here, one considers an investor who has two opportunities. Either he/she can invest his/her funds in the market, or he/she can sell a claim and invest his/her funds along with claim price. In the latter case, he/she has more funds for investment, but on the other hand, he/she faces a claim at maturity. By optimizing his/her expected

utility from the two investment scenarios, the *indifference price* of the claim is defined as the price that makes one indifferent between the two opportunities. The choice of an exponential utility function leads to prices where the singular case of zero risk aversion coincides with the price defined by the **minimal entropy martingale measure** [4]. This price lends itself to the interpretation of being the price that is equally desirable for both the issuer and the buyer in the case when both parties have zero risk aversion. For all other risk aversions, the seller will charge higher prices, and the buyer will demand lower. The difference of the two optimal investment strategies obtained from utility maximization becomes the hedging strategy. This and other similar approaches have gained a lot of academic attention in the recent years.

Another path to pricing in incomplete markets is to try to complete the market by adding options. The required number of options to complete the market is closely linked to the number of sources of uncertainty and the number of assets. For example, considering a GBM with a stochastic volatility following the Heston model gives two random sources and one asset. Following the analysis in [2], one call option is sufficient to complete the market. In [2], the necessary and sufficient conditions to complete markets are given in the case when the filtration is spanned by more Brownian motions than there are traded assets.

References

- [1] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.
- [2] Davis, M. & Obloj, J. (2008). Market completion using options, in *Advances in Mathematics of Finance*, L. Stettner, ed., Banach Center Publications, pp. 49–60, Vol. 43.
- [3] Delbaen, F. & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing, *Mathematische Annalen* **300**, 463–520.
- [4] El Karoui, N. & Rouge, R. (2000). Pricing via utility maximization and entropy, *Mathematical Finance* **10**(2), 259–276.
- [5] Hodges, S. & Neuberger, A. (1989). Optimal replication of contingent claims under transaction costs, *Review of Futures Markets* **8**, 222–239.
- [6] Merton, R. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.

- [7] Samuelson, P.A. (1965). Proof that properly anticipating prices fluctuate randomly, *Industrial Management Reviews* **6**, 41–49.
- [8] Schweizer, M. (2001). A guided tour through quadratic hedging approaches, in *Option Pricing, Interest Rates, and Risk Management*, E. Jouini, J. Cvitanic & M. Musiela, eds, Cambridge University Press, pp. 538–574.

Related Articles

Binomial Tree; Black–Scholes Formula; Hedging; Option Pricing Theory: Historical Perspectives.

FRED E. BENTH

Forwards and Futures

Futures and forwards are financial contracts that make it possible to reduce the price risk that arises from the intention to buy or sell certain assets at a later date. A forward contract specifies in advance the price that will be paid at such a later date for the delivery of the asset. This obviously reduces the price risk for that transaction to zero for all parties involved. A futures contract, on the other hand, guarantees that changes in the asset's price that occur before the delivery date will be compensated for immediately when they arise. This compensation is achieved by offsetting payments into a bank account that is called the *margin account*. This significantly reduces the price risk associated with the futures transaction, since the only possible remaining source of uncertainty is now due to the interest rate used for the margin account.

The assets that are bought or sold at the delivery date can be storable commodities (such as gold, oil, and agricultural products), nonstorable commodities (such as electricity), or other financial assets (such as stocks, bonds, options, or currencies). Forward contracts are also used by parties to agree in advance on an interest rate that will be paid or charged during a later time period, in so-called forward rate agreements (FRAs). Similarly, one can buy and sell futures on the value of money deposited in a bank account. For such interest rate futures, which include the very popular eurodollar and euribor contracts, there is no actual delivery but the contract is fulfilled by cash settlement instead.

Here, we discuss only the general pricing principles for forwards and futures. We refer to other articles in the encyclopedia for detailed information concerning the delivery procedures and methods to quote prices for specific futures and forward contracts, such as eurodollar futures (*see Eurodollar Futures and Options*), forward rate agreements (*see LIBOR Rate*), electricity (*see Electricity Forward Contracts*), commodity (*see Commodity Forward Curve Modeling*), and foreign exchange forwards (*see Currency Forward Contracts*).

Using Futures and Forwards

Forward contracts are usually agreed upon by two parties who directly negotiate the terms of such contracts, which can therefore be very flexible. The two

parties need to agree on the specific asset (often called *the underlying asset*) and on the precise quantities that are bought or sold, on the exact date when the transactions take place (the *delivery date*), and the price that will be charged on that date (the *forward price*). Usually the forward price is chosen in such a way that both parties agree to sign the contract without any money changing hands before the delivery date. This implies that the forward contract starts with having zero market value, since both parties are willing to sign it without receiving or paying any money for it. Later on, the contract may have a positive or negative market value, since every change in the market price of the underlying asset will make the existing agreement as written in the contract more beneficial to one of the parties and less beneficial to the other one. The forward contract may, therefore, become a serious liability for one of the two parties involved, so there is the risk that this party is no longer willing or able to honor the terms of the contract on the delivery date. This counterparty risk problem can be avoided by the use of futures contracts.

Futures are standardized contracts that are traded on futures exchanges. When entering a futures contract, a margin account on the futures exchange is opened and a payment into that account is required, to make it possible for the exchange to withdraw money when appropriate. The exchange publishes a *futures price* for every contract, which is updated regularly to reflect price changes in the underlying. Whenever a new futures price is announced, an amount of cash that is equal to the difference between the new futures price and the previous one is paid into or withdrawn from the margin account, depending on whether one is *short* the contract or *long* the contract. Parties that intend to buy the underlying are long the contract, and they, therefore, receive money if the futures price goes up and pay when it goes down. Parties that intend to sell are short the contract, and they, therefore, pay money if the futures price goes down and receive money when it goes up. This procedure is known as *marking to market*. Since on the delivery date the futures price is always equal to the underlying asset price, a possible difference between the initial futures price and the current asset price has been compensated for by the intermediate payments into the margin account.

Parties with opposite positions in the futures market deal only with the exchange instead of with each other, which explains the need for standardized

contracts and the significant reduction in counterparty risk. Since no cash is needed to enter into a new (long or short) futures contract as long as there is enough money left in the margin account, it is easy to change a position in futures once such an account has been established. One can terminate existing long contracts by simply taking a position in offsetting short contracts or *vice versa*, and many parties close their position just before the delivery date if they are only interested in compensation for price changes and not in the actual delivery. This makes futures very convenient to use for hedging purposes (*see Hedging*) and for speculation on an underlying's price movements. Likewise, it is quite easy for the exchange to close the futures position of a party who refuses to put more money in their margin account when asked to do so in the so-called margin call.

These characteristics have made futures very popular financial instruments and the market for them is huge. In 2008, more than eight billion futures contracts were traded worldwide with underlying assets^a in equity indices (37%), individual equity (31%), interest rates (18%), agricultural goods (5%), energy (3%), currencies (3%), and metals (2%). The most popular are contracts on the S&P 500 and Dow Jones indices, followed by eurodollar and eurobund futures, and contracts on white sugar, soybeans, crude oil, aluminum, and gold. The notional amounts underlying futures on interest rates, equity indices, and currencies at the world's exchanges were estimated to be 27 trillion, 1.6 trillion, and 175 billion US dollars, respectively, in June 2008^b.

Pricing Methods for Forwards in Discrete Time

To analyze the futures and forward prices, we first look at discrete-time models, and then look at generalizations in continuous time.

Consider a discrete-time market model on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a filtration $(\mathcal{F}_n)_{n \in \mathcal{N}}$ where $\mathcal{N} = \{0, 1, \dots, N\}$ denotes our discrete-time set. We define assets S and B to model the underlying asset and a bank account, respectively, with associated stochastic price processes $(S_n)_{n \in \mathcal{N}}$ and $(B_n)_{n \in \mathcal{N}}$. We assume that S is adapted and that B is predictable with respect to this filtration, and that both B and $1/B$ are bounded. Associated with the asset S are cash flows $(D_n)_{n \in \mathcal{N}}$ where D_n denotes the sum of all cash flows caused by holding one unit of S at

time n . These cash flows can be positive (such as dividends when S is a stock, or interest when S is a currency) or negative (such as storage costs when S is a commodity). We will always assume perfect market liquidity (*see Liquidity*), so all assets can be bought and sold in all possible quantities for their current market prices and no transaction costs (*see Transaction Costs*) are charged.

The cash flows associated with a forward contract that is initiated at time $T_0 \in \mathcal{N}$ take place at the time of delivery $T_d \in \mathcal{N}$ that is specified in the contract, with $T_0 \leq T_d$. At time T_d , the asset is delivered while the forward price agreed upon at the initial time T_0 for delivery at time T_d , which we denote by $F(T_0, T_d)$, is paid in return. Since this forward price needs to be determined at time T_0 , it should be $\mathcal{F}_{T_0}$ -measurable. Moreover, the forward price is chosen in such a way that both parties agree to enter the contract without any cash changing hands at this initial time.

In complete and arbitrage-free markets (*see Arbitrage Pricing Theory*), it is often possible to find an explicit expression for the forward price $F(T_0, T_d)$, since the cash flows associated with the contract can then be replicated using other assets with known prices. Let us assume that there exists a unique martingale measure $\mathbb{Q}$, which is equivalent to $\mathbb{P}$, such that the discounted versions of tradable assets are martingales under this measure (*see Equivalent Martingale Measures*). This is almost equivalent to the assumption of a complete and arbitrage-free market; for the exact statement (*see Fundamental Theorem of Asset Pricing*). Contingent claims that pay a cash-flow stream of $\mathcal{F}_n$ -measurable amounts X_n at the times $n \in \mathcal{N}$ in such markets have a unique price p at time $k \in \mathcal{N}$ equal to

$$p_k = \sum_{n \in \mathcal{N}, n \geq k} B_k \mathbb{E}^{\mathbb{Q}}[X_n / B_n \mid \mathcal{F}_k] \quad (1)$$

A specific example is the zero-coupon bond price at time k for the delivery of one unit cash at time $T > k$, which is equal to $p(k, T) = B_k \mathbb{E}^{\mathbb{Q}}[1/B_T \mid \mathcal{F}_k]$.

Suppose that an investor enters into a forward contract at time T_0 , which obliges him/her to deliver the underlying asset S at time T_d , and that he/she buys the underlying asset at time T_0 to hold it until delivery. This will lead to a cashflow of $-S_{T_0}$ at time T_0 , to cash flows D_n at times $\{n \in \mathcal{N} : T_0 \leq n \leq T_d\}$, and a cash flow of $F(T_0, T_d)$ at time T_d when he/she delivers the asset. Since a forward contract is entered

into without any money changing hands and since the net position after delivery will be zero, the value of the cash-flow stream defined above must be zero if there is no arbitrage in the market. Using the previous equation, we thus find that

$$0 = -S_{T_0} + B_{T_0} \mathbb{E}^{\mathbb{Q}}[F(T_0, T_d)/B_{T_d} | \mathcal{F}_{T_0}] + \sum_{T_0 \leq n \leq T_d} B_{T_0} \mathbb{E}^{\mathbb{Q}}[D_n/B_n | \mathcal{F}_{T_0}] \quad (2)$$

Since $F(T_0, T_d)$ is $\mathcal{F}_{T_0}$ -measurable, this leads to the following expression for a forward price in a complete and arbitrage-free market:

$$\begin{aligned} F(T_0, T_d) &= \frac{S_{T_0}/B_{T_0} - \sum_{T_0 \leq n \leq T_d} \mathbb{E}^{\mathbb{Q}}[D_n/B_n | \mathcal{F}_{T_0}]}{\mathbb{E}^{\mathbb{Q}}[1/B_{T_d} | \mathcal{F}_{T_0}]} \\ &= \frac{S_{T_0} - B_{T_0} \sum_{T_0 \leq n \leq T_d} \mathbb{E}^{\mathbb{Q}}[D_n/B_n | \mathcal{F}_{T_0}]}{p(T_0, T_d)} \end{aligned} \quad (3)$$

In particular, when there are no dividends or storage costs, the forward price is simply equal to the current price of the underlying asset divided by the appropriate discount rate until delivery. For commodities, where the cash flows D_n are often negative since they represent storage costs, this formula (3) is known as the *cost-of-carry formula*. Conversely, when the actual possession of an underlying asset is more beneficial than just holding the forward contract, this can be modeled by introducing positive cash flows D_n . Such benefits are often expressed as a rate, the so-called convenience yield, which may fluctuate as a result of changing expectations concerning the availability of the underlying asset on the delivery date.

The initial price of a forward contract is zero, but when the underlying asset's price changes, so does the value of an existing contract. If we denote by $G(T_0, T_d, k)$ the value at time k of a forward contract entered at time $T_0 \leq k$ for delivery at time $T_d \geq k$, then a similar argument as before leads to

$$\begin{aligned} G(T_0, T_d, k) &= B_k \mathbb{E}^{\mathbb{Q}} \left[\frac{F(k, T_d) - F(T_0, T_d)}{B_{T_d}} | \mathcal{F}_k \right] \\ &= p(k, T_d) (F(k, T_d) - F(T_0, T_d)) \end{aligned} \quad (4)$$

Pricing Methods for Futures in Discrete Time

All cash flows associated with a futures contract take place *via* the margin account. Let $(M_n)_{n \in \mathcal{N}}$ be the process describing the value of the margin account associated with a long position in one future on the underlying asset S defined above. If $f(k, T_d)$ is the futures price at time k for delivery of one unit of the asset at time $T_d > k$ ($k, T_d \in \mathcal{N}$), then the margin account values will satisfy

$$M_{k+1} = \frac{B_{k+1}}{B_k} M_k + f(k+1, T_d) - f(k, T_d) \quad (5)$$

where we assume that the interest rate used for the margin account is the same as the one used for B .

Futures prices are determined by supply and demand on the futures exchanges, but if we assume a complete and arbitrage-free market for S and B , we can derive a theoretical formula for the futures price. We consider an investment strategy where at a certain time $k \in \mathcal{N}$, we open a new margin account, put an initial margin amount M_k into it, and take a long position in a futures contract for delivery at time T_d . One time step later, we go short one future contract for the same delivery date, which effectively closes our futures position, and we then empty our margin account. Since our net position is then zero again and since we do not pay or receive money to go long or short a futures contract, the total value of this cash-flow stream at time k should be equal to zero, so

$$\begin{aligned} 0 &= -M_k + B_k \mathbb{E}^{\mathbb{Q}} \left[\frac{M_{k+1}}{B_{k+1}} | \mathcal{F}_k \right] \\ &= B_k \mathbb{E}^{\mathbb{Q}} \left[\frac{f(k+1, T_d) - f(k, T_d)}{B_{k+1}} | \mathcal{F}_k \right] \end{aligned} \quad (6)$$

Since B was assumed to be predictable, that is, B_{k+1} is $\mathcal{F}_k$ -measurable for all $k \in \mathcal{N} \setminus \{N\}$, we may conclude from the above that the futures price process $f(\cdot, T_d)$ is a $\mathbb{Q}$ -martingale for any fixed delivery date $T_d \in \mathcal{N}$, and hence

$$f(k, T_d) = \mathbb{E}^{\mathbb{Q}}[S_{T_d} | \mathcal{F}_k] \quad (7)$$

since $f(T_d, T_d) = S_{T_d}$. Note that this formula no longer holds if B fails to be predictable or when the interest rates paid on the bank account B and the margin account M are different.

Continuous-time Models

The generalization to continuous-time models is rather straightforward for forward contracts, but more subtle for futures contracts.

Assume that the price process of the underlying asset is a stochastic process S on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a filtration $(\mathcal{F}_t)_{t \in [0, T]}$ that satisfies the usual conditions, that is, it is right continuous and $\mathcal{F}_0$ contains all $\mathbb{P}$ -null sets. We will assume that the process S is an adapted semimartingale and that the bank account process B is an adapted and predictable semimartingale, and B and $1/B$ are assumed to be bounded almost surely. We model the dividend and storage costs of the asset S using an adapted semimartingale D , with the interpretation that the total amount of dividends received minus the storage costs paid between two times t_1 and t_2 is equal to $D_{t_2} - D_{t_1-}$ where $0 \leq t_1 < t_2 \leq T$.

To determine the correct forward price $F(T_0, T_d)$ for a forward contract initiated at time T_0 for delivery at time T_d , we follow the same arguments as in the discrete-time case. If we borrow money to buy the underlying asset today and then hold on to it until we deliver it at the delivery date in return for a payment of the forward price, the total value of this cash-flow stream should be zero since we enter the forward contract without any cash payments. Therefore, p_t should be zero in the formula above if we substitute $t = T_0$ and the cash-flow stream

$$\begin{aligned} X_t &= 0, \quad (t < T_0, t > T_d), \quad X_{T_0} = -S_{T_0}, \\ X_t &= D_t \quad (t \in]T_0, T_d[), \quad X_{T_d} = D_{T_d} + F(T_0, T_d) \end{aligned} \quad (9)$$

Using the fact that the forward price $F(T_0, T_d)$ must be $\mathcal{F}_{T_0}$ -measurable then leads to

$$F(T_0, T_d) = \frac{1}{p(T_0, T_d)} \left(S_{T_0} - B_{T_0} \mathbb{E}^{\mathbb{Q}} \left[\int_{T_0}^{T_d} \left(\frac{dD_u}{B_{u-}} + d \left[D, \frac{1}{B} \right]_u \right) \middle| \mathcal{F}_{T_0} \right] \right) \quad (10)$$

As in the discrete-time case, we assume that we have a complete and arbitrage-free market and that there exists a unique measure $\mathbb{Q}$ that is equivalent to $\mathbb{P}$ such that discounted versions of tradable assets become martingales under this measure (*see Equivalent Martingale Measures*). We model contingent claims by a cumulative cash-flow stream $(X_t)_{t \in [0, T]}$, which is an adapted semimartingale. The total cash amount paid out by the contingent claim between two times t_1 and t_2 is given by $X_{t_2} - X_{t_1-}$, and $X_t - X_{t-}$ corresponds to a payment at the single time t (with $t, t_1, t_2 \in [0, T]$ and $t_2 \geq t_1$). Such contingent claims have a unique price p in a complete and arbitrage-free market, which at time t is equal to

$$p_t = B_t \mathbb{E}^{\mathbb{Q}} \left[\int_t^T \left(\frac{dX_u}{B_{u-}} + d \left[X, \frac{1}{B} \right]_u \right) \middle| \mathcal{F}_t \right] \quad (8)$$

The last term involving the brackets compensates for the fact that the cash flows X and the bank account may have nonzero covariation, so it disappears when B has finite variation and is continuous, or when B is deterministic. Compare this to the discrete-time case, where we assumed that $(B_n)_{n \in \mathbb{N}}$ is predictable.

The formula for the value of a forward contract at a later time after T_0 is the same as in the discrete-time case.

We now turn to the definition of a futures price process $(f(t, T_d))_{t \in [0, T_d]}$ in continuous time for delivery at a fixed time $T_d \in [0, T]$. Let $(\psi_t)_{t \in [0, T_d]}$ be a futures investment strategy: a bounded and predictable stochastic process such that ψ_t represents the number of futures contracts (positive or negative) we own at time t . The associated margin account process $(M_t)_{t \in [0, T]}$ is then defined on $[0, T]$ as

$$dM_t = M_t \frac{dB_t}{B_{t-}} + \psi_t df(t, T_d) \quad (11)$$

with $M_0 \in \mathbb{R}$, where we have again assumed that the margin account earns the same interest rate as the bank account B . As mentioned before, the futures price process should be equal to the underlying asset price at delivery, so $f(T_d, T_d) = S_{T_d}$.

In a complete and arbitrage-free market, we consider an investment strategy where at any time $t \in [0, T_d]$ we open a new margin account and put an initial margin amount M_t in, go long one future contract at time t , wait until a later date $s \in]t, T_d]$ and

close our futures position by going short one contract, and close our margin account. If there is no arbitrage, the discounted value of the cash flows from this strategy should be zero at time t since we start and end without any position, so

$$M_t = B_t \mathbb{E}^{\mathbb{Q}} \left[\frac{M_s}{B_s} \middle| \mathcal{F}_t \right] \quad (12)$$

This shows that M/B is a martingale under $\mathbb{Q}$, that is, the margin account should be a tradable asset. A bit of stochastic calculus shows that

$$d \left(\frac{M_t}{B_t} \right) = \frac{df(t, T_d)}{B_{t-}} + d \left[f(\cdot, T_d), \frac{1}{B} \right]_t \quad (13)$$

and we see that if B is continuous, of finite variation, bounded, and bounded away from zero, then the futures price process $f(\cdot, T_d)$ is itself a martingale under $\mathbb{Q}$ and hence

$$f(t, T_d) = \mathbb{E}^{\mathbb{Q}} [f(T_d, T_d) \mid \mathcal{F}_t] = \mathbb{E}^{\mathbb{Q}} [S_{T_d} \mid \mathcal{F}_t] \quad (14)$$

Note that in this case the difference between the forward and futures prices can be expressed as

$$\begin{aligned} F(t, T_d) - f(t, T_d) &= \frac{B_{T_0}}{p(T_0, T_d)} \left(\mathbb{E}^{\mathbb{Q}} \left[\frac{S_{T_d}}{B_{T_d}} \middle| \mathcal{F}_{T_0} \right] \right. \\ &\quad \left. - \mathbb{E}^{\mathbb{Q}} [S_{T_d} \mid \mathcal{F}_{T_0}] \mathbb{E}^{\mathbb{Q}} \left[\frac{1}{B_{T_d}} \middle| \mathcal{F}_{T_0} \right] \right) \end{aligned} \quad (15)$$

Since the expression in brackets is the $\mathcal{F}_{T_0}$ -conditional covariance between S_{T_d} and $1/B_{T_d}$, we immediately see that forward and futures prices coincide if and only if these two stochastic variables are uncorrelated when conditioned on $\mathcal{F}_{T_0}$, for example, when the bank account B is deterministic.

Extensions

For clarity of exposition, we have focused here on forward and future prices in complete and arbitrage-free markets without transaction costs (see **Transaction Costs**). Early papers on the theoretical pricing methods are by Black [3] for deterministic interest rates and Cox *et al.* [5] and Jarrow and Oldfield [9] for the general case. Continuous resettlement is

treated by Duffie and Stanton [7] and Karatzas and Shreve [10], see also [12]. See [2] for a very clear summary of the principles involved. For excellent introductions to the practical organization of futures and forward markets and for empirical results on prices, the books by Duffie [6], Hull [8], and Kolb [11] are recommended.

For incomplete markets, there is a theory of equilibrium in futures market under mean–variance preferences; see, for example, [14] and the consumption-based capital asset pricing model of Breeden [4] (see also **Capital Asset Pricing Model**). Many futures allow a certain flexibility regarding the exact product that must be delivered and regarding the time of delivery. The value of this last “timing option” is analyzed in a paper by Biagini and Björk [1].

When the bank account process B is not of finite variation and continuous, the futures price is no longer a martingale under $\mathbb{Q}$; however under some technical conditions, it can be shown to be a martingale under another equivalent measure that can be found using a multiplicative Doob–Meyer decomposition (see **Doob–Meyer Decomposition**) as shown in [15]. The assumption that B and $1/B$ are bounded is often too restrictive in practice; see [13] for weaker conditions.

End Notes

^aSector estimates based on the US data, by the Futures Industry Association.

^bQuarterly Review, December 2008, Bank for International Settlements.

References

- [1] Biagini, F. & Björk, T. (2007). On the timing option in a futures contract, *Mathematical Finance* **17**(2), 267–283.
- [2] Björk, T. (2004). *Arbitrage Theory in Continuous Time*, 2nd Edition, Oxford University Press.
- [3] Black, F. (1976). The pricing of commodity contracts, *Journal of Financial Economics* **3**(1–2), 167–179.
- [4] Breeden, D.T. (1980). Consumption risk in futures markets, *Journal of Finance* **35**(2), 503–520.
- [5] Cox, J.C., Ingersoll, J. Jr. & Ross, S.A. (1981). The relation between forward prices and futures prices, *Journal of Financial Economics* **9**(4), 321–346.
- [6] Darrall, D. (1989). *Futures Markets*, Prentice-Hall.
- [7] Duffie, D. & Stanton, R. (1992). Pricing continuously resettled contingent claims, *Journal of Economic Dynamics and Control* **16**(3–4), 561–573.

- [8] Hull, J. (2003). *Options, Futures and Other Derivatives*, 5th Edition, Prentice-Hall.
- [9] Jarrow, R.A. & Oldfield, G.S. (1981). Forward contracts and futures contracts, *Journal of Financial Economics* **9**(4), 373–382.
- [10] Karatzas, I. & Shreve, S. (1998). *Methods of Mathematical Finance*, Springer-Verlag.
- [11] Kolb, R. (2003). *Futures, Options, and Swaps*, 4th Edition, Blackwell Publishing.
- [12] Norberg, R. & Steffensen, M. (2005). What is the time value of a stream of investments? *Journal of Applied Probability* **42**, 861–866.
- [13] Pozdnyakov, V. & Steele, J.M. (2004). On the martingale framework for futures prices, *Stochastic Processes and Their Applications* **109**, 69–77.
- [14] Richard, S.F. & Sundaresan, M.S. (1981). A continuous time equilibrium model of forward prices and futures prices in a multigood economy, *Journal of Financial Economics* **9**(4), 347–371.
- [15] Vellekoop, M. & Nieuwenhuis, H. (2007). *Cash Dividends and Futures Prices on Discontinuous Filtrations*. Technical Report 1838, University of Twente.

Related Articles

Commodity Forward Curve Modeling; Currency Forward Contracts; Electricity Forward Contracts; Eurodollar Futures and Options; LIBOR Rate.

MICHEL VELLEKOOP

Black–Scholes Formula

“If options are correctly priced in the market, it should not be possible to make sure profits by creating portfolios of long and short positions in options and their underlying stocks. Using this principle, a theoretical valuation formula for options is derived.” These sentences, from the abstract of the great paper [2] by Fischer Black and Myron Scholes, encapsulate the basic idea that—with the asset price model they employ—insisting on absence of **arbitrage** is enough to obtain a unique value for a call option on the asset. The resulting formula, equation (6) below, is the most famous formula in financial economics, and, in fact, that whole subject splits decisively into the **pre-Black–Scholes** and post-Black–Scholes eras.

This article aims to give a self-contained derivation of the formula, some discussion of the hedge parameters, and some extensions of the formula, and to indicate why a formula based on a stylized mathematical model, which is known not to be a particularly accurate representation of real asset prices, has nevertheless proved so effective in the world of option trading. The section The Model and Formula formulates the model and states and proves the formula. As is well known, the formula can equally well be stated in the form of a partial differential equation (PDE); this is equation (9) below. The next section discusses the PDE aspects of Black–Scholes. The section Hedge Parameters summarizes information about the option ‘greeks’, while the sections The Black ‘Forward’ Option Formula and A Universal Black Formula introduce what is actually a more useful form of Black–Scholes, usually known as the *Black formula*. Finally, the section Implied Volatility and Market Trading discusses the applications of the formula in market trading. We define the *implied volatility* and demonstrate a “robustness” property of Black–Scholes, which implies that effective hedging can be achieved even if the “true” price process is substantially different from Black and Scholes’ stylized model.

The Model and Formula

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \in \mathbb{R}^+}, \mathbb{P})$ be a **probability space** with a given filtration $(\mathcal{F}_t)$ representing the flow of

information in the market. Traded asset prices are $\mathcal{F}_t$ -adapted stochastic processes on $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that the market is *frictionless*; assets may be held in arbitrary amount, positive and negative, the interest rate for borrowing and lending is the same, and there are no transaction costs (i.e., the bid–ask spread is 0). While there may be many traded assets in the market, we fix attention on two of them. First, there is a “risky” asset whose price process $(S_t, t \in \mathbb{R}^+)$ is assumed to satisfy the **stochastic differential equation** (SDE)

$$dS_t = \mu S_t dt + \sigma S_t dw_t \quad (1)$$

with given *drift* μ and *volatility* σ . Here $(w_t, t \in \mathbb{R}^+)$ is an $(\mathcal{F}_t)$ -Brownian motion. Equation (1) has a unique solution: if S_t satisfies equation (1), then by the **Itô formula**

$$d \log S_t = \left(\mu - \frac{1}{2}\sigma^2\right) dt + \sigma dw_t \quad (2)$$

so that S_t satisfies equation (1) if and only if

$$S_t = S_0 \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma w_t\right) \quad (3)$$

Asset S_t is assumed to have a constant *dividend yield* q , that is, the holder receives a dividend payment $qS_t dt$ in the time interval $[t, t + dt]$. Secondly, there is a riskless asset paying interest at a fixed continuously compounding rate r . The exact form of this asset is unimportant—it could be a money-market account in which \$1 deposited at time s grows to $\$e^{r(t-s)}$ at time t , or it could be a zero-coupon bond maturing with a value of \$1 at some time T , so that its value at $t \leq T$ is

$$B_t = \exp(-r(T - t)) \quad (4)$$

This grows, as required, at rate r :

$$dB_t = r B_t dt \quad (5)$$

Note that equation (5) does not depend on the final maturity T (the same growth rate is obtained from any zero-coupon bond) and the choice of T is a matter of convenience.

A *European call option* on S_t is a contract, entered at time 0 and specified by two parameters (K, T) , which gives the holder the right, but not the obligation, to purchase 1 unit of the risky asset at price K at time $T > 0$. (In the frictionless market setting, an option to buy N units of stock is equivalent

to N options on a single unit, so we do not need to include quantity as a parameter.) If $S_T \leq K$ the option is worthless and will not be exercised. If $S_T > K$ the holder can exercise his option, buying the asset at price K , and then immediately selling it at the prevailing market price S_T , realizing a profit of $S_T - K$. Thus, the exercise value of the option is $[S_T - K]^+ = \max(S_T - K, 0)$. Similarly, the exercise value of a *European put option*, conferring on the holder the right to *sell* at a fixed price K , is $[K - S_T]^+$. In either case, the exercise value is nonnegative and, in the above model, is strictly positive with positive probability, so the option buyer should pay the writer a premium to acquire it. Black and Scholes [2] showed that there is a unique arbitrage-free value for this premium.

Theorem 1

1. In the above model, the unique arbitrage-free value at time $t < T$ when $S_t = S$ of the call option maturing at time T with strike K is

$$C(t, S) = e^{-q(T-t)} SN(d_1) - e^{-r(T-t)} KN(d_2) \quad (6)$$

where $N(\cdot)$ denotes the cumulative standard normal distribution function

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}y^2} dy \quad (7)$$

and

$$\begin{aligned} d_1 &= \frac{\log(S/K) + (r + \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}} \\ d_2 &= d_1 - \sigma\sqrt{T - t} \end{aligned} \quad (8)$$

2. The function $C(t, S)$ may be characterized as the unique $C^{1,2}$ solution^a of the Black–Scholes PDE

$$\frac{\partial C}{\partial t} + rS \frac{\partial C}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} - rC = 0 \quad (9)$$

solved backward in time with the terminal boundary condition

$$C(T, S) = [S - K]^+ \quad (10)$$

3. The value of the put option with exercise time T and strike K is

$$\begin{aligned} P(t, S) &= e^{-r(T-t)} KN(-d_2) \\ &\quad - e^{-q(T-t)} SN(-d_1) \end{aligned} \quad (11)$$

To prove the theorem, we are going to show that the call option value can be replicated by a dynamic trading strategy investing in the asset S_t and in the zero-coupon bond $B_t = e^{-r(T-t)}$. A trading strategy is specified by an initial capital x and a pair of adapted processes α_t, β_t representing the number of units of S, B respectively held at time t ; the portfolio value at time t is then $X_t = \alpha_t S_t + \beta_t B_t$, and by definition $x = \alpha_0 S_0 + \beta_0 B_0$. The trading strategy (x, α, β) is *admissible* if

- (i) $\int_0^T \alpha_t^2 S_t^2 dt < \infty$ a.s.
 - (ii) $\int_0^T |\beta_t| dt < \infty$ a.s.
 - (iii) There exists a constant $L \geq 0$ such that $X_t \geq -L$ for all t , a.s.
- (12)

The *gain from trade* in $[s, t]$ is

$$\int_s^t \alpha_u dS_u + \int_s^t \beta_u dB_u + \int_s^t q\alpha_u S_u du$$

where the first integral is an Itô stochastic integral. This is the sum of the accumulated capital gains/losses in the two assets plus the total dividend received. The trading strategy is *self-financing* if

$$\begin{aligned} &\alpha_t S_t + \beta_t B_t - \alpha_s S_s - \beta_s B_s \\ &= \int_s^t \alpha_u dS_u + \int_s^t q\alpha_u S_u du + \int_s^t \beta_u dB_u \end{aligned} \quad (13)$$

implying that the change in value over any interval in portfolio value is entirely due to gains from trade (the accumulated increments in the value of the assets in the portfolio plus the total dividend received).

We can always create self-financing strategies by fixing α , the investment in the risky asset, and investing all residual wealth in the bond. Indeed, the value of the risky asset holding at time t is $\alpha_t S_t$, so if the total portfolio value is X_t we take

$\beta_t = (X_t - \alpha_t S_t)/B_t$. The portfolio value process is then defined implicitly as the solution of the SDE

$$\begin{aligned} dX_t &= \alpha_t dS_t + q\alpha_t S_t dt + \beta_t dB_t \\ &= \alpha_t dS_t + q\alpha_t S_t dt + (X_t - \alpha_t S_t)r dt \\ &= rX_t dt + \alpha_t S_t \sigma (\theta dt + dw_t) \end{aligned} \quad (14)$$

where $\theta = (\mu - r + q)/\sigma$. This strategy is always self-financing since X_t is, by definition, the gains from trade process, while the value is $\alpha S + \beta B = X$.

Proof of Theorem (1) The key step is to put the “wealth equation” (14) into a more convenient form by **change of measure**. Define a measure $\mathbb{Q}$, the so-called *risk-neutral measure* on $(\Omega, \mathcal{F}_T)$ by the Radon–Nikodým derivative

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\left(-\theta w_T - \frac{1}{2}\theta^2 T\right) \quad (15)$$

(The right-hand side has expectation 1, since $w_T \sim N(0, T)$.) Expectation with respect to $\mathbb{Q}$ will be denoted $\mathbb{E}_{\mathbb{Q}}$. By the *Girsanov theorem*, $\tilde{w} = w_t + \theta t$ is a $\mathbb{Q}$ -Brownian motion, so that from equation (1) the SDE satisfied by S_t under $\mathbb{Q}$ is

$$dS_t = (r - q)S_t dt + \sigma S_t d\tilde{w}_t \quad (16)$$

so that for $t < T$

$$\begin{aligned} S_T &= S_t \exp\left((r - q - \frac{1}{2}\sigma^2)(T - t) \right. \\ &\quad \left. + \sigma(\tilde{w}_T - \tilde{w}_t)\right) \end{aligned} \quad (17)$$

Applying the Itô formula and equation (14) we find that, with $\tilde{X}_t = e^{-rt} X_t$ and $\tilde{S}_t = e^{-rt} S_t$

$$d\tilde{X}_t = \alpha_t \tilde{S}_t \sigma d\tilde{w}_t \quad (18)$$

Thus $e^{-rt} X_t$ is a $\mathbb{Q}$ -local martingale under condition (12)(i). Let $h(S) = [S - K]^+$ and suppose there exists a replicating strategy, that is, a strategy (x, α, β) with value process X_t constructed as in equation (14) such that $X_T = h(S_T)$ a.s. Suppose also that α_t satisfies the stronger condition

$$\mathbb{E}_{\mathbb{Q}} \int_0^T \alpha_t^2 S_t^2 dt < \infty \quad (19)$$

Then $\tilde{X}_t$ is a $\mathbb{Q}$ -martingale, and hence for $t < T$

$$X_t = e^{-r(T-t)} \mathbb{E}_{\mathbb{Q}}[h(S_T) | \mathcal{F}_t] \quad (20)$$

and in particular

$$x = e^{-rT} \mathbb{E}_{\mathbb{Q}}[h(S_T)] \quad (21)$$

Now S_t is a **Markov process**, so the conditional expectation in equation (20) is a function of S_t , and indeed we see from equation (17) that S_T is a function of S_t and the increment $(\tilde{w}_T - \tilde{w}_t)$, which is independent of $\mathcal{F}_t$. Writing $(\tilde{w}_T - \tilde{w}_t) = Z\sqrt{T-t}$ where $Z \sim N(0, 1)$, the expectation is simply a one-dimensional integral with respect to the normal distribution. Hence, $X_t = C(t, S_t)$ where

$$\begin{aligned} C(t, S) &= \frac{e^{-r(T-t)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(S \exp((r - q - \sigma^2/2) \\ &\quad \times (T - t) - \sigma x \sqrt{T - t})) e^{-1/2x^2} dx \end{aligned} \quad (22)$$

Straightforward calculations show that this integral is equal to the closed-form expression in equation (6).

The argument so far shows that if there is a replicating strategy, the initial capital required must be $x = C(0, S_0)$ where C is defined by equation (22). It remains to identify the strategy (x, α, β) and to show that it is admissible. Let us temporarily take for granted the assertions of part (2) of the theorem; these will be proved in Theorem 3 below, where we also show that $(\partial C / \partial S)(t, S) = e^{-q(T-t)} N(d_1)$, so that in particular $0 < \partial C / \partial S < 1$.

The replicating strategy is $\mathcal{A} = (x, \alpha, \beta)$ defined by

$$\begin{aligned} x &= C(0, S_0), \quad \alpha_t = \frac{\partial C}{\partial S}(t, S_t) \\ \beta_t &= \frac{1}{rB_t} \left(\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S_t^2 \frac{\partial^2 C}{\partial S^2} - qS_t \frac{\partial C}{\partial S} \right) \end{aligned} \quad (23)$$

Indeed, using the PDE (9) we find that $X_t = \alpha_t S_t + \beta_t B_t = C(t, S_t)$, so that $\mathcal{A}$ is replicating and also $X_t \geq 0$, so that condition (12)(iii) is satisfied. From equation (17)

$$S_t^2 = S_0^2 \exp((2r - 2q - \sigma^2)t + 2\sigma \tilde{w}_t) \quad (24)$$

so that $\mathbb{E}_{\mathbb{Q}}[S_t^2] = \exp((2r - 2q + \sigma^2)t)$. Since $|e^{-r(T-t)} \partial C / \partial S| < 1$, this shows that $\mathbb{E}_{\mathbb{Q}} \int_0^T \alpha_t^2 S_t^2 dt < \infty$, that is, condition (19) is satisfied. Since β_t is, almost surely (a. s.), a continuous function of t it

satisfies equation (12)(ii). Thus $\mathcal{A}$ is admissible. The gain from trade in an interval $[s, t]$ is

$$\begin{aligned}
 & \int_s^t \alpha_u dS_u + \int_s^t q\alpha_u S_u du + \int_s^t \beta_u dB_u \\
 &= \int_s^t \frac{\partial C}{\partial S} dS + \int_s^t \left(\frac{\partial C}{\partial t} + \frac{1}{2} \sigma^2 S_t^2 \frac{\partial^2 C}{\partial S^2} \right) du \\
 &= \int_s^t dC \\
 &= C(t, S_t) - C(s, S_s)
 \end{aligned} \tag{25}$$

(We obtain the first equality from the definition of α, β , and it turns out to be just the Itô formula applied to the function C .) This confirms the self-financing property and completes the proof.

Finally, part (3) of the theorem follows from the model-free **put–call parity** relation $C - P = e^{-q(T-t)}S - e^{-r(T-t)}K$ and symmetry of the normal distribution: $N(-x) = 1 - N(x)$. $\square$

The replicating strategy derived above is known as **delta hedging**: the number of units of the risky asset held in the portfolio is equal to the Black–Scholes delta, $\Delta = \partial C / \partial S$.

So far, we have concentrated entirely on the hedging of call options. We conclude this section by showing that, with the class of trading strategies we have defined, there are no arbitrage opportunities in the Black–Scholes model.

Theorem 2 *There is no admissible trading strategy in a single asset and the zero-coupon bond that generates an arbitrage opportunity, in the Black–Scholes model.*

Proof Suppose X_t is the portfolio value process corresponding to an admissible trading strategy (x, α, β) . There is an arbitrage opportunity if $x = 0$ and, for some t , $X_t \geq 0$ a.s. and $\mathbb{P}[X_t > 0] > 0$, or equivalently $\mathbb{E}[X_t] > 0$. This is the $\mathbb{P}$ -expectation, but $\mathbb{E}[X_t] > 0 \Leftrightarrow \mathbb{E}_{\mathbb{Q}}[\tilde{X}_t] > 0$ since $\mathbb{P}$ and $\mathbb{Q}$ are equivalent measures and $e^{-rt} > 0$. From equation (18), $\tilde{X}_t$ is a $\mathbb{Q}$ -local martingale which, by the definition of admissibility, is bounded below by a constant $-L$. It follows that $\tilde{X}_t$ is a **supermartingale**, so if $x = 0$, then $\mathbb{E}_{\mathbb{Q}}[\tilde{X}_t] \leq 0$ for any t . So no arbitrage can arise from the strategy $(0, \alpha, \beta)$. $\square$

The Black–Scholes Partial Differential Equation

Theorem 3

1. The Black–Scholes PDE (9) with boundary condition (10) has a unique $C^{1,2}$ solution, given by equation (6).
2. The Black–Scholes “delta”, $\Delta(t, S)$, is given by

$$\Delta(t, S) = \frac{\partial}{\partial S} C(t, S) = e^{-q(T-t)} N(d_1) \tag{26}$$

Proof It can—with some pain—be directly checked that $C(t, S)$ defined by equation (6) does satisfy the Black–Scholes PDE (9), (10), and a further calculation (not quite as simple as it appears) gives the formula (26) for the Black–Scholes delta. It is, however, enlightening to take the original route of Black and Scholes and relate the equation (9) to a simpler equation, the *heat equation*. Note from the explicit expression (17) for the price process under the risk-neutral measure that, given the starting point S_t , there is a one-to-one relation between S_T and the Brownian increment $\tilde{w}_T - \tilde{w}_t$. We can therefore always express things interchangeably in “ S coordinates” or in “ $\tilde{w}$ coordinates”. In fact, we already made use of this in deriving the integral price expression (22). Here we proceed as follows.

For fixed parameters S_0, r, q, σ , define the functions $\phi : \mathbb{R}^+ \times \mathbb{R} \rightarrow \mathbb{R}^+$ and $u : [0, T] \times \mathbb{R} \rightarrow \mathbb{R}^+$ by

$$\phi(t, x) = S_0 \exp \left(\left(r - q - \frac{1}{2} \sigma^2 \right) t + \sigma x \right) \tag{27}$$

and

$$u(t, x) = C(t, \phi(t, x)) \tag{28}$$

Note that the inverse function $\psi(t, s) = \phi^{-1}(t, s)$ (i.e., the solution for x of the equation $s = \phi(t, x)$) is

$$\psi(t, s) = \frac{1}{\sigma} \left(\log \left(\frac{s}{S_0} \right) - \left(r - q - \frac{1}{2} \sigma^2 \right) t \right) \tag{29}$$

A direct calculation shows that C satisfies equation (9) if and only if u satisfies the heat equation

$$\frac{\partial u}{\partial t} + \frac{1}{2} \frac{\partial^2 u}{\partial x^2} - r u = 0 \tag{30}$$

If W_t is Brownian motion on some probability space and u is a $C^{1,2}$ function, then an application of the Itô formula shows that

$$d(e^{-rt}u(t, W_t)) = e^{-rt} \left(\frac{\partial u}{\partial t} + \frac{1}{2} \frac{\partial^2 u}{\partial x^2} - ru \right) dt + e^{-rt} \frac{\partial u}{\partial x} dW_t \quad (31)$$

If u satisfies equation (30) with boundary condition $u(T, x) = g(x)$ and

$$\mathbb{E} \int_0^T \left(\frac{\partial u}{\partial x}(t, W_t) \right)^2 dt < \infty \quad (32)$$

then the process $t \mapsto e^{-rt}u(t, W_t)$ is a martingale so that, with $\mathbb{E}_{t,x}$ denoting the conditional expectation given $W_t = x$,

$$\begin{aligned} e^{-rt}u(t, x) &= \mathbb{E}_{t,x}[e^{-rT}u(T, W_T)] \\ &= \mathbb{E}_{t,x}[e^{-rT}g(W_T)] \end{aligned} \quad (33)$$

Since $W_T \sim N(x, T - t)$, this shows that u is given by

$$u(t, x) = \frac{e^{-r(T-t)}}{\sqrt{2\pi(T-t)}} \int_{-\infty}^{\infty} g(y) e^{-\frac{(y-x)^2}{2(T-t)}} dy \quad (34)$$

A sufficient condition for equation (32) is

$$\frac{1}{\sqrt{2\pi T}} \int_{-\infty}^{\infty} g^2(y) e^{-y^2/2T} dy < \infty \quad (35)$$

In our case, the boundary condition is $g(x) = [\phi(t, x) - K]^+ < \phi(t, x)$ and this condition is easily checked. Hence, equation (30) with this boundary condition has unique $C^{1,2}$ solution (34), implying that the inverse function $C(t, S) = u(t, \psi(t, S))$ given by

equation (22) is the unique $C^{1,2}$, solution of equation (9) as claimed. $\square$

Hedge Parameters

Bringing in all the parameters, the Black–Scholes formula (6) is a six-parameter function $C(t, S) = C(\tau, S, K, r, q, \sigma)$, where $\tau = T - t$ is the time to maturity. For risk-management purposes, it is important to know the sensitivities of the option value to changes in the parameters. The conventional hedge parameters or “greeks” are given in Table 1. There are slight notational problems in that “vega” is not the name of a Greek letter (here we have used upper-case upsilon, but this is not necessarily a conventional choice) and upper-case rho coincides with Latin P, so this parameter is usually written ρ , risking confusion with correlation parameters. The expressions in the right-hand column are readily obtained from the sensitivity parameters (42) and (43) of the “universal” Black Formula introduced below.

Delta is, of course, the Black–Scholes hedge ratio. Gamma measures the convexity of C and is at its maximum when the option is close to being at the money. Since gamma is the rate of change of delta, frequent rebalancing of the hedge portfolio will be required in areas of high gamma. Theta is defined as $-\partial C/\partial \tau$ and is generally negative (as can be seen from the table, it is always negative for a call option on an asset with no dividends). It represents the “time decay” in the option value as the maturity time is reduced, that is, real time advances. As regards rho, it is not immediately obvious, without doing the calculation, what its sign will be: on one hand, increasing r increases the forward price, pushing a call option further into the money, while on the other hand increased r implies heavier discounting, reducing option value. As can be seen from the table,

Table 1 Black–Scholes risk parameters

Delta	Δ	$\frac{\partial C}{\partial S}$	$e^{-q\tau} N(d_1)$
Gamma	Γ	$\frac{\partial^2 C}{\partial S^2}$	$\frac{e^{-q\tau} N'(d_1)}{S\sigma\sqrt{\tau}}$
Theta	Θ	$-\frac{\partial C}{\partial \tau}$	$-\frac{e^{-q\tau} S N'(d_1) \sigma}{2\sqrt{\tau}} + q e^{-q\tau} S N(d_1) - r K e^{-r\tau} N(d_2)$
Rho	P	$\frac{\partial C}{\partial r}$	$K \tau e^{-r\tau} N(d_2)$
Vega	Υ	$\frac{\partial C}{\partial \sigma}$	$e^{-q\tau} S \sqrt{\tau} N'(d_1)$

the first effect wins: rho is always positive. Vega is in some ways the most important parameter, since a key risk in managing books of traded options is “vega risk”, and in Black–Scholes this is completely “outside the model”. Bringing it back inside the model is the subject of **stochastic volatility**.

An extensive discussion of the risk parameters and their uses can be found in Hull [6].

The Black “Forward” Option Formula

The six-parameter representation $\mathcal{C}(\tau, S, K, r, q, \sigma)$ is not the best parameterization of Black–Scholes. For the asset S_t with dividend yield q , the *forward price* at time t for delivery at time T is $F(t, T) = S_t e^{(r-q)(T-t)}$ (this is a model-free result, not related to the Black–Scholes model). We can trivially re-express the price formula (6) as

$$C(t, S_t) = B(t, T)(F(t, T)N(d_1) - KN(d_2)) \quad (36)$$

with

$$d_1 = \frac{\log(F(t, T)/K) + \frac{1}{2}\sigma^2(T-t)}{\sigma\sqrt{T-t}} \\ d_2 = d_1 - \sigma\sqrt{T-t} \quad (37)$$

where $B(t, T) = e^{-r(T-t)}$ is the zero-coupon bond value or “discount factor” from T to t . There is, however, far more to this than just a change of notation. First, the continuously compounding rate r is not market data. The market data at time t is the set of discount factors $B(t, t')$ for $t' > t$. We see from equation (36) that “ r ” plays two distinct roles in Black–Scholes: it appears in the computation of the forward price F and the discount factor B . But both of these are more fundamental than r itself and are, in fact, market data which, as equation (36) shows, can be used directly. A further advantage is that the exact mechanism of dividend payment is not important, as long as there is an unambiguously defined forward price.

Formula (36) is known as the *Black formula* and is the most useful version of Black–Scholes, being widely applied in connection with FX (**foreign exchange**) and **interest-rate options** as well as dividend-paying equities. Fundamentally, it relates to a price model in which the price is expressed in the

risk-neutral measure as $S_t = F(0, t)M_t$ where M_t is the exponential martingale

$$M_t = \exp\left(\sigma\check{w}_t - \frac{1}{2}\sigma^2 t\right) \quad (38)$$

which is equivalent to equation (17). This model accords with the general fact that, in a world of deterministic interest rates, the forward price is the expected price in the risk-neutral measure, that is, the ratio $S_t/F(0, t)$ is a positive martingale with expectation 1. The exponential martingale (38) is the simplest continuous-path process with these properties.

A Universal Black Formula

The parameterization of Black–Scholes can be further compressed as follows. First, note that σ and $\tau = (T - t)$ do not appear separately, but only in the combination $a = \sigma\sqrt{T - t}$, where a^2 is sometimes known as the *operational time*. Next, define the “moneyness” m as $m(t, T) = K/F(t, T)$, and define

$$d(a, m) = \frac{a}{2} - \frac{\log m}{a} \quad (39)$$

(so that $d_1 = d(\sigma\sqrt{T - t}, K/F(t, T))$). Then the Black formula (36) becomes

$$C = BF f(a, m) \quad (40)$$

where

$$f(a, m) = N(d(a, m)) - mN(d(a, m) - a) \quad (41)$$

Now BF is the price of a zero-strike call, or equivalently the price to be paid at time t for delivery of the asset at time T . Formula (40) says that the price of the K -strike call is the (model-free) price of the zero-strike call modified by a factor f that depends only on the moneyness and operational time. We call f the *universal Black–Scholes function*, and a graph of it is shown in Figure 1. With $N' = dN/dx$ and $d = d(a, m)$ we find that $mN'(d - a) = N'(d)$ and hence obtain the following very simple expressions for the first-order derivatives:

$$\frac{\partial f}{\partial a}(a, m) = N'(d) \quad (42)$$

$$\frac{\partial f}{\partial m}(a, m) = -N(d - a) \quad (43)$$

In particular, $\partial f/\partial a > 0$ and $\partial f/\partial m < 0$ for all a, m .

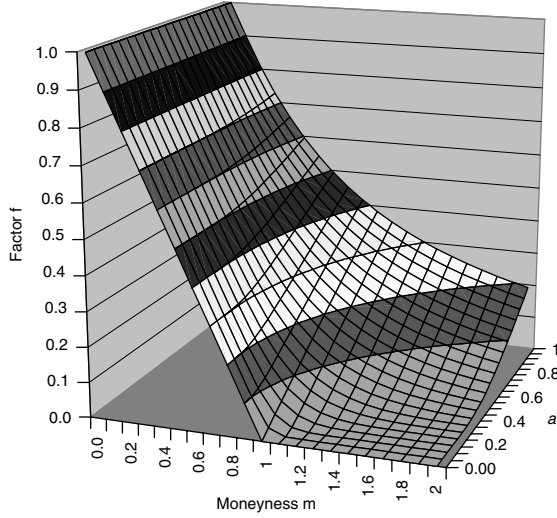


Figure 1 The universal Black–Scholes function

This minimal parameterization of Black–Scholes is used in studies of stochastic volatility; see, for example, **Gatheral** [5].

Implied Volatility and Market Trading

So far, our discussion has been entirely within the Black–Scholes model. What happens if we attempt to use Black–Scholes delta hedging in real market trading? This question has been considered by several authors, including El Karoui *et al.* [3] and Fouque *et al.* [4], though neither of these discusses the effect of jumps in the price process.

In the universal price formula (40), the parameters B, F, m are market data, so we can regard the formula as a mapping $a \mapsto p = BFf(a, m)$ from a to price $p \in [B[F - K]^+, BF)$. In a traded options market, p is market data (but must lie in the stated interval, else there is a static arbitrage opportunity). In view of equation (42), $f(a, m)$ is strictly increasing in a and hence there is a unique value $a = \hat{a}(p)$ such that $p = BFf(\hat{a}(p), m)$. The *implied volatility* is $\hat{\sigma}(p) = \hat{a}(p)/\sqrt{T - t}$. If the underlying price process S_t actually were geometric Brownian motion (1), then $\hat{\sigma}$ would be the same, and equal to the volatility σ , for call options of all strikes and maturities. Of course, this is never the case in practice—see [5] for

a discussion. Here, we restrict ourselves to examining what happens if we naïvely apply the Black–Scholes delta-hedge when in reality the underlying process is *not* geometric Brownian motion, taking $q = 0$ for simplicity. Specifically, we assume that the “true” price model, under measure $\mathbb{P}$, is

$$S_t = S_0 + \int_0^t \eta_t S_{t-} dt + \int_0^t \kappa_t S_{t-} dW_t + \int_{[0,t] \times \mathbb{R}} S_{t-} v_t(z) \mu(dt, dz) \quad (44)$$

where μ is a finite-activity *Poisson random measure*, so that there is a finite measure ν on $\mathbb{R}$ such that $\mu([0, t] \times A) - \nu(A)t \equiv (\mu - \pi)([0, t] \times A)$ is a martingale for each $A \in \mathcal{B}(\mathbb{R})$. η, κ, v are predictable processes. Assume that η, κ and v are such that the solution to the SDE (44) is well defined and, moreover, that $v_t(z) > -1$ so $S_t > 0$ almost surely. This is a very general model including path-dependent coefficients, stochastic volatility, and jumps. Readers unfamiliar with jump-diffusion models can set $\mu = \nu = \pi = 0$ below, and refer to the last paragraph of this section for comments on the effect of jumps.

Consider the scenario of selling at time 0 a European call option at implied volatility $\hat{\sigma}$, that is, for the price $p = C(T, S_0, K, r, \hat{\sigma})$ and then following a Black–Scholes delta-hedging trading strategy based on constant volatility $\hat{\sigma}$ until the option expires at time T . As usual, we shall denote $C(t, s) = C(T - t, s, K, r, \hat{\sigma})$, so that the hedge portfolio, with value process X_t , is constructed by holding $\alpha_t := \partial_S C(t, S_{t-})$ units of the risky asset S , and the remainder $\beta_t := \frac{1}{B_t}(X_{t-} - \alpha_t S_{t-})$ units in the riskless asset B (a unit notional zero-coupon bond). This portfolio, initially funded by the option sale (so $X_0 = p$), defines a self-financing trading strategy. Hence, the portfolio value process X satisfies the SDE

$$\begin{aligned} X_t = p &+ \int_0^t \partial_S C(u, S_{u-}) \eta_u S_{u-} du \\ &+ \int_0^t \partial_S C(u, S_{u-}) \kappa_u S_{u-} dW_u \\ &+ \int_{[0,t] \times \mathbb{R}} \partial_S C(u, S_{u-}) S_{u-} v_u(z) \mu(du, dz) \\ &+ \int_0^t (X_u - \partial_S C(u, S_{u-}) S_u) r du \end{aligned} \quad (45)$$

Now define $Y_t = C(t, S_t)$, so that, in particular, $Y_0 = p$. Applying the Itô formula (Lemma 4.4.6 of [1]) gives

$$\begin{aligned} Y_t &= p + \int_0^t \partial_t C(u, S_{u-}) du \\ &\quad + \int_0^t \partial_S C(u, S_{u-}) \eta_u S_{u-} du \\ &\quad + \int_0^t \partial_S C(u, S_{u-}) \kappa_u S_{u-} dW_u \\ &\quad + \frac{1}{2} \int_0^t \partial_{SS}^2 C(u, S_{u-}) \kappa_u^2 S_{u-}^2 du \\ &\quad + \int_{[0,t] \times \mathbb{R}} (C(u, S_{u-}(1 + v_u(z))) \\ &\quad - C(u, S_{u-})) \mu(dt, dz) \end{aligned} \quad (46)$$

Thus the ‘hedging error’ process defined by $Z_t := X_t - Y_t$ satisfies the SDE

$$\begin{aligned} Z_t &= \int_0^t r X_u du - \int_0^t (r S_{u-} \partial_S C(u, S_{u-}) \\ &\quad + \partial_t C(u, S_{u-}) + \frac{1}{2} \kappa_u^2 S_{u-}^2 \partial_{SS}^2 C(u, S_{u-})) du \\ &\quad - \int_{[0,t] \times \mathbb{R}} (C(u, S_{u-}(1 + v_u(z))) - C(u, S_{u-}) \\ &\quad - \partial_S C(u, S_{u-}) S_{u-} v_u(z)) \mu(du, dz) \\ &= \int_0^t r Z_u du + \frac{1}{2} \int_0^t \Gamma(u, S_{u-}) S_{u-}^2 (\hat{\sigma}^2 - \kappa_u^2) du \\ &\quad - \int_{[0,t] \times \mathbb{R}} (C(u, S_{u-}(1 + v_u(z))) \\ &\quad - C(u, S_{u-}) - \partial_S C(u, S_{u-}) S_{u-} v_u(z)) \mu(du, dz) \end{aligned} \quad (47)$$

where $\Gamma(t, S_t) = \partial_{SS}^2 C(t, S_t)$, and the last equality follows from the Black–Scholes PDE. Therefore, the final difference between the hedging strategy and the required option payout is given by

$$\begin{aligned} Z_T &= X_T - [S_T - K]^+ \\ &= \frac{1}{2} \int_0^T e^{r(T-t)} S_{t-}^2 \Gamma(t, S_{t-}) (\hat{\sigma}^2 - \kappa_t^2) dt \end{aligned}$$

$$\begin{aligned} &- \int_{[0,T] \times \mathbb{R}} e^{r(T-t)} \left(\int_0^1 \int_0^\epsilon \Gamma(t, S_{t-}(1 + \epsilon' v_t(z))) \right. \\ &\quad \times v_t^2(z) S_{u-}^2 d\epsilon' d\epsilon \Big) \pi(dt, dz) - M_T \end{aligned} \quad (48)$$

where M_T is the terminal value of the martingale

$$\begin{aligned} M_t &= \int_{[0,T] \times \mathbb{R}} e^{r(T-t)} \left(\int_0^1 \int_0^\epsilon \Gamma(t, S_{t-}(1 + \epsilon' v_t(z))) \right. \\ &\quad \times v_t^2(z) S_{u-}^2 d\epsilon' d\epsilon \Big) (\mu - \pi)(dt, dz) \end{aligned} \quad (49)$$

Equation (48) is a key formula, as it shows that successful hedging is quite possible even under significant model error. Without some ‘robustness’ property of this kind, it is hard to imagine that the derivatives industry could exist at all, since hedging under realistic conditions would be impossible.

Consider first the case $\mu \equiv 0$, where S_t has continuous sample paths and the last two terms in equation (48) vanish. Then, successful hedging depends entirely on the relationship between the implied volatility $\hat{\sigma}$ and the true ‘local volatility’ κ_t . Note from Table 1 that $\Gamma_t > 0$. If we, as option writers, are lucky and $\hat{\sigma}^2 \geq \kappa_t^2$ a.s. for all t , then the hedging strategy makes a profit *with probability 1* even though the true price model is substantially different from the assumed model as in equation (1). On the other hand, if we underestimate the volatility, we will consistently make a loss. The magnitude of the profit or loss depends on the option convexity Γ . If Γ is small, then hedging error is small even if the volatility has been grossly misestimated.

For the option writer, jumps in either direction are unambiguously bad news. Since C is convex, $\Delta C > (\partial C / \partial S) \Delta S$, so the last term in equation (47) is monotone decreasing: the hedge profit takes a hit every time there is a jump, either upward or downward, in the underlying price. However, there is some recourse: in equation (48), M_T has expectation 0 while the penultimate term is negative. By increasing $\hat{\sigma}$ we increase $\mathbb{E}[Z_T]$, so we could arrive at a situation where $\mathbb{E}[Z_T] > 0$, although in this case there is no possibility of *with probability 1* profit because of the martingale term. All of this reinforces the trader’s intuition that one can offset additional hedge costs by charging more upfront (i.e.,

increasing $\hat{\sigma}$) and hedging at the higher level of implied volatility.

End Notes

^aA two-parameter function is $C^{1,2}$ if it is once (twice) continuously differentiable in the first (second) argument.

References

- [1] Applebaum, D. (2004). *Lévy Processes and Stochastic Calculus*, Cambridge University Press.
- [2] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.
- [3] El Karoui, N., Jeanblanc-Picqué, M. & Shreve, S.E. (1998). Robustness of the Black and Scholes formula, *Mathematical Finance* **8**, 93–126.
- [4] Fouque, J.-P., Papanicolaou, G. & Sircar, K.R. (2000). *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press.
- [5] Gatheral, J. (2006). *The Volatility Surface*, Wiley.
- [6] Hull, J.C. 2005. *Options, Futures and Other Derivatives*, 6th Edition, Prentice Hall.

MARK H.A. DAVIS

Exchange Options

Definition and Examples

A European exchange option is a contract that gives the buyer the right to exchange two (possibly dividend-paying) assets A and B at a fixed expiration time T , say, to receive A and deliver (or pay) B ; thus, the option payoff is

$$(A_T - B_T)^+ := \max(A_T - B_T, 0) \quad (1)$$

(American and Bermudan exchange options are complicated by early optimal exercise and not discussed here.) An ordinary (European) call or put on an asset struck at K can be viewed, as in [9], as an option to exchange the asset with the T -maturity zero-coupon bond of principal K . More generally, a call or put on an s -maturity forward contract ($s \geq T$) on a zero-dividend asset is equivalent to an option to exchange the asset at time T with an s -maturity zero-coupon bond. Options to exchange two stocks or commodities provide good hypothetical examples but are not prevalent in the market place.

Exchange options are related to spread options with time- T payoffs of the form $(X - Y)^+$, given two prescribed time- T observables X and Y . A common structure is a CMS spread option, with X and Y , say, the 20-year and the 2-year spot swap rates at time T . A spread option can be viewed as an exchange option when there exist (or can be replicated) two zero-dividend assets A and B such that $A_T = X$ and $B_T = Y$. In the CMS case, A and B can be taken as the coupon cash flows of two CMS bonds or swaps. In practice, exchange options on dividend-paying assets are reduced to the zero-dividend case in a similar way.

Interest-rate swaptions, including caplets and floorlets as one-period special cases, can be viewed both as ordinary call or put options struck at par on coupon bonds and more directly as options to exchange the fixed and floating cash flow legs of a swap. The latter is the standard as it imposes the classical assumption of a lognormal ratio A_T/B_T on the forward swap rate (a swap-curve concept) rather than on the forward coupon bond price.

An exchange option is related to its reverse by parity: $(Y - X)^+ = (X - Y)^+ + Y - X$. (Hence, an

American option to exchange two fixed zero-dividend assets is not exercised early.)

Pricing and Hedging Approaches

The exchange option is a special case of a *path-independent* contingent claim with payoff being a *homogeneous function* of the underlying asset prices at expiration. It is governed by the same general theory (see **Option Pricing: General Principles**). One makes sure that the underlying assets are arbitrage free, which implies that there are no free lunches in a strong sense. If the payoff can be attained by a sufficiently regular self-financing trading strategy (SFTS) (e.g., a bounded number of shares or “deltas”), then the law of one price holds and the option price at each time is defined as the value of the self-financing portfolio. Otherwise, arbitrage-free pricing is not unique. We do not discuss this case, but only mention that one approach then chooses a linear pricing kernel (e.g., the minimal measure) among the many then available and another is nonlinear based on expected utility maximization.

Payoff replication by an SFTS is a question of predictable representation. As the payoff in this case is a path-independent function of the underliers, it seems natural that the option price as well as deltas be functions of time and the underliers at that time. This has been the traditional Markovian approach, beginning with Black and Scholes [1] and immediate extension by Merton [9] (see **Black–Scholes Formula**). Their simple choice of a geometric Brownian motion for the underlying asset in [1] and more generally of a deterministic-volatility forward price process in [9] meant that the underlying SDE and the associated PDE had constant coefficients (in log-state). Itô’s formula was applied to construct a riskless hedge, with the deltas (hedge ratios) simply given by partial derivatives of the unique solution to the PDE.

Black and Scholes constructed an SFTS for a call option struck at K by dynamically rebalancing long positions on the underlying asset A financed by shorting the riskless money market asset $B^* = (e^{rt})$, post an initial investment equal to the option price. Merton’s extension to stochastic interest rate r treated the call as an option C to exchange the asset A with the T -maturity zero-coupon bond B of principal K . The Black–Scholes model corresponded to a

deterministic bond price $B_t = e^{-r(T-t)}K$, but now, in general, B had infinite variation. The former's simplicity was nonetheless recaptured by exploiting the homogeneous symmetry of the option payoff to reduce dimensionality by 1—in effect, a projective transformation that hedged the forward option contract $F := C/B$ with trades in the forward asset $X := A/B$. The relevant volatility was accordingly the forward price volatility. An SFTS in the two assets and Itô's formula led to a PDE for the homogeneous option price function $C(t, A, B)$ and an equivalent PDE for the forward option price function $F(t, X)$.

Margrabe [8] extended the theory in [9] to an option to exchange any two correlated assets assuming constant volatilities (*see Margrabe Formula*). He observed akin to [9] that the self-financing equation with $\partial C/\partial A$ and $\partial C/\partial B$ as deltas is, by Itô's formula, equivalent to $C(t, A, B)$ satisfying a PDE with no first-order terms in A, B . Choosing C as the homogenized Black–Scholes function, it followed by Euler's formula for homogeneous functions that $\partial C/\partial A$ and $\partial C/\partial B$ in fact formed an SFTS. The result demonstrated that (in this case) the exchange option is replicated by dynamically going long in A and short in B , with no trades in any other asset. (This fails in general, e.g., a bond exchange option in a $k \geq 3$ factor non-Gaussian short-rate model.) “*Taking asset two as numéraire*,” Margrabe [8] also presented (acknowledging Stephen Ross) a key financial invariance argument as a heuristic alternative to the PDE algebraic proof of [9], reducing to a call on A/B struck at 1 in the Black–Scholes model with zero interest rate.

Martingale theory leads to a conceptual as well as computationally practical representation of solutions to the PDEs that describe option prices as a conditional expectation of terminal payoff. Harrison and Kreps [5] and Harrison and Pliska [6] developed, in related papers, an equivalent martingale measure framework that not only made this fruitful representation of the option price available but also laid a more general and probabilistic formulation of the notion of a dynamic hedge, or its mirror image, a replicating SFTS (*see Risk-neutral Pricing*). Their arbitrage-free semimartingale approach does permit path dependency, yet accommodates Markovian SDE/PDE models even better. They took the money market asset B^* as a tradable entering any hedge, giving it a general stochastic form

$B_t^* = e^{\int_0^t r_s ds}$ for discounting payoffs before expectation. In concert with Black–Scholes but in contrast to Merton and Margrabe, the finite variation asset B^* was their exclusive choice of numéraire.

With the advent of the forward measure sometime later (*see Forward and Swap Measures*), it was evident that Merton's choice of an infinite variation zero-coupon bond B as the financing hedge instrument fitted equivalent martingale measure theory perfectly well, and it led to quicker derivations of concrete pricing formulae than B^* , as discounting is conveniently performed outside the expectation [4, 7]. Another useful numéraire was the one by Neuberger [10] to price interest-rate swaptions. Viewed as an option to exchange the fixed and floating swap cash flows, the assets' ratio A/B represents the forward swap rate here. The assumption in [10] that the ratio has deterministic volatility yielded a model that has since served as industry standard to quote swaption-implied volatilities (*see Swap Market Models*). Here, it is noteworthy that the ratio A/B has deterministic volatility but A and B themselves decidedly do not. In time, El-Karoui *et al.* [4] showed that one can basically change numéraire to any asset B and associate with it an equivalent measure under which A/B is a martingale for every other asset A (*see Change of Numéraire*).

Today, option pricing and hedging theory has advanced farther and in many directions. Especially relevant to our discussion of exchange options is the principle of numéraire invariance and arbitrage-free modeling. For in-depth studies of these and related topics, we refer the reader to [3] and [2], among other excellent books. Our approach is to concentrate on the modeling in “projective coordinate” $X := A/B$, and impose for the most part conditions that are invariant under the transformation $X \mapsto 1/X$.

The Deterministic-volatility and Exponential-Poisson Models

The option to exchange two assets with a *deterministic volatility* $\sigma(t)$ of the asset price ratio $X = A/B$ is celebrated as the simplest nontrivial example in option pricing theory. Its classical Black–Scholes/Merton option price function and explicit representation of the “deltas” (“hedge ratios”) illustrate the principles that underline options in many assets with arbitrary homogeneous payoffs and more

general dynamics. There is another concrete albeit less known example with simple jumps in X involving the Poisson rather than the normal distribution. The pattern is similar, with the main difference being that the deltas are the partial differences rather than the partial derivatives of the option price function.

We fix, throughout, a stochastic basis $(\Omega, (\mathcal{F}_t), \mathcal{F}, \mathbb{P})$ with time horizon $t \in [0, T]$, $T > 0$. In this section, we fix two *zero-dividend* assets with price processes $A = (A_t)$ and $B = (B_t)$.

The Exchange Option Price Process

When A and B are semimartingales, we call a pair (δ^A, δ^B) of (locally) bounded predictable processes a (locally) *bounded SFTS* (see, more generally, the section Self-financing Trading Strategies) if $C = C_0 + \int \delta^A dA + \int \delta^B dB$, where

$$C = \delta^A A + \delta^B B \quad (2)$$

Clearly, C is then a semimartingale, $\Delta C = \delta^A \Delta A + \delta^B \Delta B$, and hence $C_- = \delta^A A_- + \delta^B B_-$.

The differential form of the self-financing equation is often handy:

$$dC = \delta^A dA + \delta^B dB \quad (3)$$

SFTSs form a linear space. If there exists a unique *bounded SFTS* (δ^A, δ^B) such that

$$C_T = (A_T - B_T)^+ \quad (4)$$

then it is justified to call C the *exchange option price process* and δ^A and δ^B the *deltas*.

Assume now that the semimartingales A and B are positive and have positive left limits.

The numéraire invariance principle (see the section Numéraire Invariance and more comprehensively the section The Invariance Principle) states that if (δ^A, δ^B) is a locally bounded SFTS, then $C = \delta^A A + \delta^B B$ satisfies $d\left(\frac{C}{B}\right) = \delta^A d\left(\frac{A}{B}\right)$ (similarly by symmetry with A as numéraire). This is useful for uniqueness. Numéraire invariance also states the converse: if C is a semimartingale and δ^A a locally bounded predictable process such that $d\left(\frac{C}{B}\right) = \delta^A d\left(\frac{A}{B}\right)$, then (δ^A, δ^B) is an SFTS and equations (2) and (3) hold, where $\delta^B = \frac{C_-}{B_-} - \delta^A \frac{A_-}{B_-}$.

This reduces existence to finding an F_0 and δ^A such that

$$\left(\frac{A_T}{B_T} - 1\right)^+ = F_0 + \int_0^T \delta_t^A d\left(\frac{A_t}{B_t}\right) \quad (5)$$

The exchange option price process is then the semimartingale $C = B \left(F_0 + \int \delta^A d\left(\frac{A}{B}\right) \right)$.

Numéraire invariance in effect reduces general option pricing and hedging to a market where one of the asset price processes equals 1 identically. The remaining task is to find the above “projective” predictable representation of the ratio payoff against the ratio process.

Deterministic-volatility Exchange Option Model

Let $\sigma(t) > 0$ be a continuous positive function. Define the *Black–Scholes/Merton projective option price function*

$$f(t, x) := x\delta_A(t, x) + \delta_B(t, x) \quad (6)$$

for $t \leq T$, $x > 0$, where $\delta_A(T, x) := 1_{x>1}$, $\delta_B(T, x) := -1_{x>1}$, and for $t < T$,

$$\begin{aligned} \delta_A(t, x) &:= N\left(\frac{\log x}{\sqrt{v_t}} + \frac{\sqrt{v_t}}{2}\right), \\ \delta_B(t, x) &:= -N\left(\frac{\log x}{\sqrt{v_t}} - \frac{\sqrt{v_t}}{2}\right) \end{aligned} \quad (7)$$

where $v_t := \int_t^T \sigma^2(s) ds$ and $N(\cdot)$ is the normal distribution function. The function $f(t, x)$ is continuous, and on $t < T$ is C^1 in t and analytic in x . In addition, $-1 \leq \delta_B \leq 0 \leq \delta_A \leq 1$, and

$$f(T, x) = (x - 1)^+, \quad \frac{\partial f}{\partial x}(t, x) = \delta_A(t, x) \quad (8)$$

As is well known and seen in the sections Deterministic-volatility Model Uniqueness and Projective Continuous SDE SFTS, the function $f(t, x)$ is the unique $C^{1,2}$ (on $t < T$) solution with bounded partial derivative $\frac{\partial f}{\partial x}(t, x)$ subject to $f(T, x) = (x - 1)^+$ of the PDE

$$\frac{\partial f}{\partial t}(t, x) + \frac{1}{2}\sigma^2(t)x^2 \frac{\partial^2 f}{\partial x^2}(t, x) = 0 \quad (9)$$

4 Exchange Options

Assume now $A = BX$ for some positive continuous semimartingale $X > 0$ satisfying

$$d[\log X]_t = \sigma^2(t)dt, \quad (A = BX) \quad (10)$$

Under this assumption, one traditionally defines the exchange option price process C by

$$C := BF, \quad F = (F_t), \quad F_t := f(t, X_t) \quad (11)$$

Clearly, $C_T = (A_T - B_T)^+$. The definition is justified using the continuous semimartingales

$$\begin{aligned} \delta_t^A &:= \delta_A(t, X_t) = \frac{\partial f}{\partial x}(t, X_t), \\ \delta_t^B &:= \delta_B(t, X_t) = F_t - \delta_t^A X_t \end{aligned} \quad (12)$$

Clearly, $C = \delta^A A + \delta^B B$, and the deltas are bounded: $0 \leq \delta^A \leq 1$ and $-1 \leq \delta^B \leq 0$. Since $f(t, x)$ satisfies the PDE (9) (as directly verified) and $\frac{\partial f}{\partial x}(t, X_t) = \delta_t^A$, by Itô's formula, the continuous semimartingale $F := (f(t, X_t))$ satisfies the predictable representation

$$dF = \delta^A dX \quad (13)$$

If, at this stage, we assume that B is a semimartingale, then A and C are semimartingales too, and by the invariance principle discussed next, $dC = \delta^A dA + \delta^B dB$ and (δ^A, δ^B) is a bounded SFTS.

Numéraire Invariance

Let X and F be two semimartingales and δ^A be a locally bounded predictable process such that $dF = \delta^A dX$. Set $\delta^B = F - \delta^A X$. Clearly $\delta^B = F_- - \delta^A X_-$ since $\Delta F = \delta^A \Delta X$. Let B be any semimartingale. Set $A = BX$, $C = BF$. Clearly $C = \delta^A A + \delta^B B$. We claim $dC = \delta^A dA + \delta^B dB$, so (δ^A, δ^B) is an SFTS.

Indeed, this follows by applying Itô's product rule to BF , then substituting $dF = \delta^A dX$ and $F_- = \delta^B + \delta^A X_-$, followed by Itô's product rule on BX :

$$\begin{aligned} dC &= d(BF) = B_- dF + F_- dB + d[B, F] \\ &= B_- \delta^A dX + (\delta^B + \delta^A X_-)dB + \delta^A d[B, X] \\ &= \delta^A d(BX) + \delta^B dB = \delta^A dA + \delta^B dB \end{aligned} \quad (14)$$

Conversely, if A and B are semimartingales with $B, B_- > 0$ and (δ^A, δ^B) is an SFTS, then

$d\left(\frac{C}{B}\right) = \delta^A d\left(\frac{A}{B}\right)$, where $C = \delta^A A + \delta^B B$. (See the section The Invariance Principle for a more lucid treatment.)

Exponential-Poisson Exchange Option Model

Assume that the two zero-dividend asset price processes A and B satisfy $A = BX$, where X is a semimartingale satisfying

$$X_t = X_0 e^{\beta P_t - (\beta - 1)\lambda t} \quad (15)$$

for some constants $\beta \neq 0$, $\lambda > 0$ and semimartingale P such that $[P] = P$ and $P_0 = 0$ (Thus, $P_t = \sum_{s \leq t} 1_{\Delta P_s \neq 0}$). Define the projective option price function $f(t, x)$, $x > 0$ by

$$\begin{aligned} f(t, x) &:= \sum_{n=0}^{\infty} (x e^{\beta n - (\beta - 1)\lambda(T-t)} - 1)^+ \\ &\quad \times \frac{\lambda^n}{n!} (T - t)^n e^{-\lambda(T-t)} \end{aligned} \quad (16)$$

and exchange option price process by

$$C := BF, \quad F = (F_t), \quad F_t := f(t, X_t) \quad (17)$$

Clearly $f(T, x) = (x - 1)^+$ and $C_T = (A_T - B_T)^+$. One has the predictable representation

$$dF = \delta^A dX \quad (18)$$

as shown shortly, where

$$\delta_t^A := \delta_A(t, X_{t-}), \quad \delta_A(t, x) := \frac{f(t, e^\beta x) - f(t, x)}{(e^\beta - 1)x} \quad (19)$$

Thus by numéraire invariance, (δ^A, δ^B) is an SFTS if A and B are semimartingales, where

$$\delta^B := F - \delta^A X = F_- - \delta^A X_- \quad (20)$$

Moreover, it is bounded. Indeed, since $|(e^\beta y - 1)^+ - (y - 1)^+| \leq |e^\beta - 1|y$ for any $y > 0$,

$$\begin{aligned} 0 \leq \delta_A(t, x) &\leq \sum_{n=0}^{\infty} e^{\beta n - (\beta - 1)\lambda(T-t)} \\ &\quad \times \frac{\lambda^n}{n!} (T - t)^n e^{-\lambda(T-t)} = 1 \end{aligned} \quad (21)$$

Hence, $0 \leq \delta^A \leq 1$. Similarly, $-1 \leq \delta^B \leq 0$.

We note that $f(t, x)$ is *not* C^1 in x (though convex, absolutely continuous, and piecewise analytic in x). We also caution that this model is arbitrage free only when $\mathbb{P}\{P_t = n\} > 0$ for all $t > 0$ and $n \in \mathbb{N}$, for example, when P is a Poisson process under an equivalent measure.

Derivation of the Predictable Representation

To show $dF = \delta^A dX$ (equation (18)), we first note that $[P]^c = 0$ since $[P] = P$; hence, $(\Delta P)^2 = \Delta P$ and $P_t = [P]_t = \sum_{s \leq t} \Delta P_s$. If $v(p)$, $p \in \mathbb{R}$, is any function, then clearly $V = (v(P_t))$ is a semimartingale and we have

$$\begin{aligned} \Delta V_t &= v(P_t) - v(P_{t-}) = (v(P_t) - v(P_{t-}))\Delta P_t \\ &= (v(P_{t-} + 1) - v(P_{t-}))\Delta P_t \end{aligned} \quad (22)$$

Hence, as V is clearly the sum of its jumps,

$$\begin{aligned} V_t - v(0) &= \sum_{s \leq t} \Delta V_s \\ &= \sum_{s \leq t} (v(P_{s-} + 1) - v(P_{s-}))\Delta P_s \\ &= \int_0^t (v(P_{s-} + 1) - v(P_{s-}))dP_s \end{aligned} \quad (23)$$

Likewise, $(u(t, P_t))$ is a semimartingale for any C^1 in t function $u(t, p)$, $p \in \mathbb{R}$, and one has

$$\begin{aligned} du(t, P_t) &= \frac{\partial u}{\partial t}(t, P_{t-})dt + (u(t, P_{t-} + 1) \\ &\quad - u(t, P_{t-}))dP_t \end{aligned} \quad (24)$$

Now, define the function

$$x(t, p) := X_0 e^{\beta p - (e^\beta - 1)\lambda t} \quad (p \in \mathbb{R}) \quad (25)$$

Clearly $X_t = x(t, P_t)$. Applying equation (24) to the function $x(t, p)$ and using that

$$\begin{aligned} \frac{\partial x}{\partial t}(t, p) &= -x(t, p)(e^\beta - 1)\lambda, \\ x(t, p + 1) - x(t, p) &= x(t, p)(e^\beta - 1) \end{aligned} \quad (26)$$

(or alternatively applying Itô's formula to $x(t, P_t)$ and simplifying) yields

$$dX_t = X_{t-}(e^\beta - 1)d(P_t - \lambda t) \quad (27)$$

Next, define the function of $t \leq T$ and $p \in \mathbb{R}$,

$$\begin{aligned} u(t, p) &:= f(t, x(t, p)) \\ &= \sum_{n=0}^{\infty} (X_0 e^{\beta(p+n) - (e^\beta - 1)\lambda T} - 1)^+ \\ &\quad \times \frac{\lambda^n}{n!} (T - t)^n e^{-\lambda(T-t)} \end{aligned} \quad (28)$$

Clearly, $u(t, P_t) = F_t$. One readily verifies that $u(t, p)$ satisfies the equation

$$\frac{\partial u}{\partial t}(t, p) + \lambda(u(t, p + 1) - u(t, p)) = 0 \quad (29)$$

Hence by equation (24) we have,

$$dF_t = (u(t, P_{t-} + 1) - u(t, P_{t-}))d(P_t - \lambda t) \quad (30)$$

Combining this with equation (27) and the fact that clearly

$$\begin{aligned} u(t, p + 1) - u(t, p) &= f(t, e^\beta x(t, p)) \\ &\quad - f(t, x(t, p)) \end{aligned} \quad (31)$$

we conclude that, as desired,

$$dF_t = \frac{f(t, e^\beta X_{t-}) - f(t, X_{t-})}{(e^\beta - 1)X_{t-}} dX_t \quad (32)$$

The Homogeneous Option Price Function

There is an alternative derivation of the self-financing equation $dC = \delta^A dA + \delta^B dB$ much along that in [9] and [8] that does not employ numéraire invariance. It is related to a family of two-dimensional PDEs satisfied by the *Merton/Margrabe homogeneous option price function* $c(t, a, b)$ below.

Let $f(t, x)$, $x > 0$, be any $C^{1,2}$ function, for example, as in equation (6). Define the homogenized function

$$c(t, a, b) := bf\left(t, \frac{a}{b}\right) \quad (a, b > 0) \quad (33)$$

Then $c(t, a, b)$ is homogeneous of degree 1 in (a, b) , and hence by Euler's formula

$$c(t, a, b) = \frac{\partial c}{\partial a}(t, a, b)a + \frac{\partial c}{\partial b}(t, a, b)b \quad (34)$$

A laborious repeated application of the chain rule on equation (33) gives

$$\begin{aligned} a^2 \frac{\partial^2 c}{\partial a^2}(t, a, b) &= b^2 \frac{\partial^2 c}{\partial b^2}(t, a, b) \\ &= -ab \frac{\partial^2 c}{\partial a \partial b}(t, a, b) \\ &= b x^2 \frac{\partial^2 f}{\partial x^2}(t, x), \quad x := \frac{a}{b} \end{aligned} \quad (35)$$

Let $\sigma(t)$, $\sigma_A(t, a, b)$, $\sigma_B(t, a, b)$, $\sigma_{AB}(t, a, b)$ be any functions ($a, b > 0$) such that

$$\sigma^2(t) = \sigma_A^2(t, a, b) + \sigma_B^2(t, a, b) - 2\sigma_{AB}(t, a, b) \quad (36)$$

Using equations (35), (36), and $\frac{\partial c}{\partial t}(t, a, b) = b \frac{\partial f}{\partial t}\left(t, \frac{a}{b}\right)$, we see that $c(t, a, b)$ satisfies the PDE

$$\begin{aligned} \frac{\partial c}{\partial t} + \frac{1}{2} \sigma_A^2(t, a, b) a^2 \frac{\partial^2 c}{\partial a^2} + \frac{1}{2} \sigma_B^2(t, a, b) b^2 \frac{\partial^2 c}{\partial b^2} \\ + \sigma_{AB}(t, a, b) ab \frac{\partial^2 c}{\partial a \partial b} = 0 \end{aligned} \quad (37)$$

if and only if $f(t, x)$ satisfies the PDE (9): $\frac{\partial f}{\partial t} + \frac{1}{2} \sigma^2(t) x^2 \frac{\partial^2 f}{\partial x^2} = 0$.

The PDE (9) was utilized in [1] and [9] (but not in [8]), and Merton [9] stated its equivalence to the PDE (37) (assuming σ_A , etc., depend only on t). As noted in [9] and expounded in [8], if $d[\log A]_t = \sigma_A^2(t)dt$, $d[\log B]_t = \sigma_B^2(t)dt$ and $d[\log A, \log B]_t = \sigma_{AB}(t)dt$, then Itô's formula and equation (37) imply at once $dc(t, A_t, B_t) = \delta_t^A dA_t + \delta_t^B dB_t$, with δ^A and δ^B as in equation (40), and thus (δ^A, δ^B) is an SFTS with price process $c(t, A, B)$ by Euler's formula (34).

Let us expand on this (see also the sections Self-financing Trading Strategies and Homogeneous Continuous Markovian SFTS). Let $\sigma(t) > 0$ be a continuous function, and $f(t, x)$ be the Black-Scholes/Merton function (6). Set $c(t, a, b) := bf(t, a/b)$. Clearly,

$$\frac{\partial c}{\partial a}(t, a, b) = \frac{\partial f}{\partial x}\left(t, \frac{a}{b}\right) = \delta_A\left(t, \frac{a}{b}\right) \quad (38)$$

This combined with Euler's formula (34) and the definition (6) $f := \delta_A x + \delta_B$ give

$$\frac{\partial c}{\partial b}(t, a, b) = \delta_B\left(t, \frac{a}{b}\right) \quad (39)$$

Assume that A and B are positive semimartingales with positive left limits and $X := A/B$ has deterministic volatility $\sigma(t)$: $d[X]_t = X_t^2 \sigma^2(t)dt$. Using equation (12), the deltas are conveniently the sensitivities of the homogeneous Merton/Margrabe function:

$$\delta_t^A = \frac{\partial c}{\partial a}(t, A_t, B_t), \quad \delta_t^B = \frac{\partial c}{\partial b}(t, A_t, B_t) \quad (40)$$

Since X is continuous, we also have $\delta_t^A = \frac{\partial c}{\partial a}(t, A_{t-}, B_{t-})$ and similarly δ_t^B . The section Deterministic-Volatility Exchange Option Model yields $dC = \delta^A dA + \delta^B dB$ with $C_t = B_t f(t, X_t) = c(t, A_t, B_t)$. Therefore, by equation (40) and Itô's formula,

$$\begin{aligned} \frac{\partial c}{\partial t} dt + \frac{1}{2} \frac{\partial^2 c}{\partial a^2} d[A]_t^c + \frac{1}{2} \frac{\partial^2 c}{\partial b^2} d[B]_t^c \\ + \frac{\partial^2 c}{\partial a \partial b} d[A, B]_t^c = 0 \end{aligned} \quad (41)$$

where the partial derivatives are evaluated at (t, A_{t-}, B_{t-}) and $[\cdot]^c$ is the bracket continuous part. (The jump term in Itô's formula vanishes as it equals $\sum_{s \leq t} (\Delta C_s - \delta_s^A \Delta A_s - \delta_s^B \Delta B_s) = 0$.)

Returning to the approach of Merton [9], assume now that $d[\log A]_t = \sigma_A^2(t, A_t, B_t)dt$ for some function σ_A and similarly $d[\log B] = \sigma_B^2 dt$ and $d[\log A, \log B] = \sigma_{AB} dt$. Then equation (36) holds using $\log X = \log A - \log B$. Since $f(t, x)$ satisfies the PDE (9), the PDE (37) follows as before by the chain rule. However, equation (37) implies equation (41), which by Itô's formula in turn implies the self-financing equation $dC = \delta^A dA + \delta^B dB$ with δ^A and δ^B given by equation (40).

Change of Numéraire

The solution $c(t, a, b)$ to the PDE (37) subject to $c(T, a, b) = (a - b)^+$ can be expressed in a form $\mathbb{E}(X - Y)^+$ for some random variables X and $Y > 0$ with means a and b . Expectations of this form often become more tractable by a change of measure as in

[4]. Define the equivalent probability measure $\mathbb{Q}$ by $\frac{d\mathbb{Q}}{d\mathbb{P}} := \frac{Y}{\mathbb{E}Y}$. Clearly,

$$\mathbb{E}^{\mathbb{Q}}\left(\frac{X}{Y}\right) = \frac{\mathbb{E}(X)}{\mathbb{E}(Y)} \quad \left(\frac{d\mathbb{Q}}{d\mathbb{P}} := \frac{Y}{\mathbb{E}(Y)}\right) \quad (42)$$

Replacing X by $(X - Y)^+$ in equation (42) and using the homogeneity to factor out Y , we get

$$\mathbb{E}(X - Y)^+ = \mathbb{E}(Y)\mathbb{E}^{\mathbb{Q}}\left(\frac{X}{Y} - 1\right)^+ \quad (43)$$

If X/Y is $\mathbb{Q}$ -lognormally distributed then equation (43) together with equation (42) readily yields

$$\begin{aligned} \mathbb{E}(X - Y)^+ &= \mathbb{E}(X)N\left(\frac{\log(\mathbb{E}X/\mathbb{E}Y)}{\sqrt{\nu^{\mathbb{Q}}}} + \frac{\sqrt{\nu^{\mathbb{Q}}}}{2}\right) \\ &\quad - \mathbb{E}(Y)N\left(\frac{\log(\mathbb{E}X/\mathbb{E}Y)}{\sqrt{\nu^{\mathbb{Q}}}} - \frac{\sqrt{\nu^{\mathbb{Q}}}}{2}\right) \end{aligned} \quad (44)$$

where $\nu^{\mathbb{Q}} := \text{var}^{\mathbb{Q}}[\log(X/Y)]$. When X and Y are bivariate lognormally distributed, it is not difficult to show that X/Y is lognormally distributed in both $\mathbb{P}$ and $\mathbb{Q}$ with the same log-variance $\nu^{\mathbb{Q}} = \nu := \text{var}[\log(X/Y)]$. Then $\nu^{\mathbb{Q}}$ can be replaced with ν in equation (44). This occurs when the functions σ_A , σ_B and σ_{AB} in equation (37) are independent of a and b , as in [8, 9].

Uniqueness

Assume that A and B are positive semimartingales with positive left limits such that $X := A/B$ is square-integrable martingale under an equivalent probability measure $\mathbb{Q}$ and $d\langle X \rangle_t^{\mathbb{Q}} = X_t^2 \sigma_t^2 dt$ for some nowhere zero process σ , where $\langle X \rangle^{\mathbb{Q}}$ is the $\mathbb{Q}$ -compensator of $[X]$. (Of course, $\langle X \rangle^{\mathbb{Q}} = [X]$ if X is continuous.) Let (δ^A, δ^B) be an SFTS and set $C := \delta^A A + \delta^B B$. We claim that $\delta^A = \delta^B = 0$ if $C_T = 0$ and δ^A is bounded.

Indeed, set $F := C/B$. By numéraire invariance, $dF = \delta^A dX$. Hence, F is a $\mathbb{Q}$ -square-integrable martingale since X is and δ^A is bounded. Thus, $F = 0$ since $F_T = C_T/B_T = 0$. Hence, $0 = d\langle F \rangle_t^{\mathbb{Q}} = (\delta^A)^2 X_t^2 \sigma_t^2 dt$. However, $X_t^2 \sigma_t^2 > 0$. Thus, $\delta^A = 0$ and $\delta^B = F - \delta^A X = 0$.

In general, since $F := C/B$ is a $\mathbb{Q}$ -martingale, we have the following pricing formula:

$$C_t = B_t \mathbb{E}^{\mathbb{Q}}[C_T/B_T | \mathcal{F}_t] \quad (45)$$

Deterministic-volatility Model Uniqueness

Assume that A and B are positive semimartingales with positive left limits and $X := A/B$ is an Itô process following

$$\frac{dX_t}{X_t} = \mu_t dt + \sigma_t dZ_t, \quad \left(X := \frac{A}{B}\right) \quad (46)$$

where Z is a Brownian motion and μ and $\sigma > 0$ are predictable processes with σ bounded and $\mathbb{E}[e^{1/2 \int_0^T (\mu_t/\sigma_t)^2 dt}] < \infty$. Let (δ^A, δ^B) be an SFTS with δ^A bounded. Set $C := \delta^A A + \delta^B B$. We claim that $\delta^A = \delta^B = 0$ if $C_T = 0$. Indeed, the process

$$M := \mathcal{E}\left(-\int \frac{\mu}{\sigma} dZ\right) = e^{-\int \frac{\mu}{\sigma} dZ - \frac{1}{2} \int \left(\frac{\mu}{\sigma}\right)^2 dt} \quad (47)$$

is then a positive martingale with $M_0 = 1$. Define the equivalent probability measure $\mathbb{Q}$ by $d\mathbb{Q} = M_T d\mathbb{P}$. The process $W := Z + \int \frac{\mu}{\sigma} dt$ is a $\mathbb{Q}$ -Brownian motion because $[W]_t = t$ and W is $\mathbb{Q}$ -local martingale as MW is a local martingale using Itô's product rule:

$$\begin{aligned} d(MW) - WdM &= MdW + d[W, M] \\ &= M\left(dZ + \frac{\mu}{\sigma} dt\right) - M\frac{\mu}{\sigma} d[Z] \\ &= MdZ \end{aligned} \quad (48)$$

Moreover, $dX = X\sigma dW$ by equation (46). Therefore, X is a $\mathbb{Q}$ -square integrable martingale since σ is bounded. The claim, thus, follows by the section Uniqueness.

Assume now that σ_t is *deterministic*. The results of the section Deterministic-Volatility Exchange Option Model hold since $d[\log X] = \sigma_t^2 dt$. However, we can now derive them more conceptually. Indeed, both conditioned on $\mathcal{F}_t$ and unconditionally, X_T/X_t is $\mathbb{Q}$ -lognormally distributed with mean 1 and log-variance

$\int_t^T \sigma_s^2 ds$ since $X_T = X_t e^{\int_t^T \sigma_s dW_s - 1/2 \int_t^T \sigma_s^2 ds}$. Hence, by equation (45),

$$f(t, X_t) = \mathbb{E}^{\mathbb{Q}}[(X_T - 1)^+ | \mathcal{F}_t] \quad \text{where}$$

$$f(t, x) := \mathbb{E}^{\mathbb{Q}}\left(x \frac{X_T}{X_t} - 1\right)^+ \quad (49)$$

which function readily equals the Black–Scholes/Merton option price function (6). Thus, $F := (f(t, X_t))$ is a $\mathbb{Q}$ -martingale. Therefore, Itô's formula implies that $f(t, x)$ satisfies the PDE (9) and $dF = \delta^A dX$ where $\delta^A := \frac{\partial f}{\partial x}(t, X_t)$. Numéraire invariance now yields that the pair $(\delta^A, \delta^B := F - \delta^A X)$ is an SFTS. Clearly, $C_T = (A_T - B_T)^+$ where $C := \delta^A A + \delta^B B = BF$.

Exponential-Poisson Model Uniqueness

Let $\beta \neq 0$ be a constant and κ and λ be positive continuous adapted processes such that λ is bounded and $\mathbb{E} \int_0^T \left(\frac{\lambda_t}{\kappa_t} - 1\right)^2 \kappa_t dt < \infty$. Let P be semimartingale satisfying $[P] = P$ with $P_0 = 0$ and compensator $\int \kappa dt$. Assume that A and B are positive semimartingales with positive left limits and $X := \frac{A}{B}$ satisfies

$$dX_t = X_{t-}(e^\beta - 1)(dP_t - \lambda_t dt) \quad (50)$$

Using $de^{\beta P} = (e^\beta - 1)e^{\beta P} dP$ or as in the section Derivation of the Predictable Representation, this is equivalent to the integrated form

$$X_t = X_0 e^{\beta P_t - (e^\beta - 1) \int_0^t \lambda_s ds} \quad (51)$$

Let (δ^A, δ^B) be an SFTS with δ^A bounded. Set $C := \delta^A A + \delta^B B$. We claim that $\delta^A = \delta^B = 0$ if $C_T = 0$. Indeed, $\mathbb{E} \left\langle \int \left(\frac{\lambda}{\kappa} - 1\right) (dP - \kappa dt) \right\rangle_T = \mathbb{E} \int_0^T \left(\frac{\lambda_t}{\kappa_t} - 1\right)^2 \kappa_t dt < \infty$, so the positive local martingale

$$M := \mathcal{E} \left(\int \left(\frac{\lambda}{\kappa} - 1 \right) (dP - \kappa dt) \right)$$

$$= e^{-\int (\lambda - \kappa) dt} \prod_{s \leq \cdot} \left(1 + \left(\frac{\lambda_s}{\kappa_s} - 1 \right) \Delta P_s \right) \quad (52)$$

is a martingale. Define the equivalent probability measure $\mathbb{Q}$ by $d\mathbb{Q} = M_T d\mathbb{P}$. Then $N := P - \int \lambda dt$ is a $\mathbb{Q}$ -local martingale as MN is a local martingale by Itô's product rule:

$$d(MN) - N_- dM = M_- dN + d[M, N]$$

$$= M_- (dP - \lambda dt)$$

$$+ M_- \left(\frac{\lambda}{\kappa} - 1 \right) dP$$

$$= M_- \frac{\lambda}{\kappa} (dP - \kappa dt) \quad (53)$$

Therefore, by equation (50), X is a $\mathbb{Q}$ -square-integrable martingale (in fact, in $\mathcal{H}^p(\mathbb{Q})$ for all $p > 0$) since λ is bounded. Thus, by the section Uniqueness, $\delta^A = \delta^B = 0$ if $C_T = 0$, as claimed.

Assume now that λ is a positive constant. By equation (51) we have a special case of the exponential-Poisson model. Further, P is a $\mathbb{Q}$ -Poisson process with intensity λ since $[P] = P$. We now have uniqueness, but additionally, the previous results follow more conceptually as follows.

Conditioned on $\mathcal{F}_t$, $P_T - P_t$ is $\mathbb{Q}$ -Poisson distributed with mean $\lambda(T - t)$. Its unconditional $\mathbb{Q}$ -distribution is identical. Thus, the $\mathcal{F}_t$ -conditional and the unconditional $\mathbb{Q}$ -distribution of X_T/X_t are identical and are exponentially Poisson distributed with mean 1. Hence, by equation (45),

$$f(t, X_t) = \mathbb{E}^{\mathbb{Q}}[(X_T - 1)^+ | \mathcal{F}_t] \quad \text{where}$$

$$f(t, x) := \mathbb{E}^{\mathbb{Q}}\left(x \frac{X_T}{X_t} - 1\right)^+ \quad (54)$$

which function readily equals that defined in equation (16). Thus, $F := (f(t, X_t))$ is a $\mathbb{Q}$ -martingale. Using this and equation (24), one shows that F satisfies equation (32) and with it that the pair (δ^A, δ^B) as defined in equation (19), equation (20) is a bounded SFTS for the exchange option.

Extension to Dividends

Consider two assets with positive price processes $\hat{A}$ and $\hat{B}$ and continuous dividend yields y_t^A and y_t^B . When there exist traded or replicable zero-dividend assets A and B such that $A_T = \hat{A}_T$ and $B_T = \hat{B}_T$ (if not, there is little hope of replication), it is natural to

define the price process of the option to exchange $\hat{A}$ and $\hat{B}$ to be that of the option to exchange A and B . If y^A and y^B are deterministic, then consistent with the treatment of dividends in [9], A (and similarly B) is simply given by

$$\begin{aligned} A_t &:= a\tilde{A}_t = e^{-\int_t^T y_s^A ds} \hat{A}_t, \\ \tilde{A}_t &:= e^{\int_0^t y_s^A ds} \hat{A}_t, \quad a := e^{-\int_0^T y_t^A dt} \end{aligned} \quad (55)$$

Note A/B is a semimartingale if and only if $\hat{A}/\hat{B}$ is, in which case $[\log A/B] = [\log \hat{A}/\hat{B}]$.

In general, $\tilde{A}_t$ is the price of the zero-dividend asset that initially buys one share of $\hat{A}$ and thereon continually reinvests all dividends in $\hat{A}$ itself. What is required is that the four zero-dividend assets A , $\tilde{A}$, B , and $\tilde{B}$ be arbitrage free in relation to one another (see the section Arbitrage-free Semimartingales and Uniqueness).

For instance, say $\hat{A}$ and $\hat{B}$ are the yen/dollar and yen/Euro exchange rates viewed as yen-denominated dividend assets. Then A is the yen-value of the US T -maturity zero-coupon bond and $\tilde{A}$ is the yen-value of the US money market asset. This exchange option is equivalent to a Euro-denominated call struck at 1 on the Euro/dollar exchange rate $\hat{A}/\hat{B}$. The ratio A/B is the *forward* Euro/dollar exchange rate. If it has deterministic volatility, we are as in a setting of [7], which yields the same pricing formula as that from the section Deterministic-volatility Exchange Option Model.

Pricing and Hedging Options with Homogeneous Payoffs

We took some shortcuts to quickly present the main results for two of the simplest and among the most interesting examples. A better understanding of the principles at work requires generalization to contingent claims C on many assets with price processes $A = (A^1, \dots, A^m) > 0$ and a path-independent payoff $C_T = h(A_T)$ given as a homogeneous function $h(a)$, $a \in \mathbb{R}_+^m$, of the asset prices A_T at expiration time T . Combined with an underlying SDE and the resulting PDE, such a Markovian setting utilizes the invariance principle and equivalent martingale measures to derive unique pricing and construct an SFTS that

replicates the given payoff $h(A_T)$ in general. The construction is explicit in the multivariate extensions of the deterministic-volatility and exponential-Poisson models.

The homogeneity of the payoff function $h(a)$ implies $h(A_T) = A_T^m g(X_T)$ where $g(x) := h(x, 1)$, $x \in \mathbb{R}_+^n$, $n := m - 1$, and $X := \left(\frac{A^1}{A^m}, \dots, \frac{A^n}{A^m}\right)$. Once a predictable representation $F = F_0 + \delta' \cdot X$, $F_T = g(X_T)$ is found, then by numéraire invariance $\delta := (\delta', \delta^m)$ will be an SFTS with payoff $h(A_T)$, where $\delta^m := F_- - \sum_{i=1}^n \delta^i X_- = F - \sum_{i=1}^n \delta^i X$. Uniqueness of pricing requires boundedness of partial derivatives (or differences) of $h(a)$ (or $g(x)$) and that A be arbitrage free, meaning X is a martingale under an equivalent measure. Arbitrage freedom holds “generically” when the matrix $(\langle X^i, X^j \rangle)$ is nonsingular, basically a “no-redundant-asset” condition. Then the SFTS is also unique.

Libor and swap derivatives are among contingent claims with homogeneous payoffs.

Self-financing Trading Strategies

By an SFTS we mean a pair (δ, A) of an m -dimensional semimartingale $A = (A^1, \dots, A^m)$ and an A -integrable predictable vector process $\delta = (\delta^1, \dots, \delta^m)$ such that (with $\delta \cdot A$ denoting the m -dimensional stochastic integral)

$$\sum_{i=1}^m \delta^i A^i = \sum_{i=1}^m \delta_0^i A_0^i + \delta \cdot A \quad (56)$$

We then say δ is an SFTS for A . This is equivalent to saying that the SFTS price process

$$C := \sum_{i=1}^m \delta^i A^i \quad (57)$$

satisfies $C = C_0 + \delta \cdot A$. Clearly, C is then a semimartingale, $\Delta C = \sum_i \delta^i \Delta A_i$, and hence

$$C_- = \sum_{i=1}^m \delta^i A_-^i \quad (58)$$

If δ^i are bounded (say by b) and A^i are martingales, then the SFTS price process C is a martingale

because C is then a local martingale that is dominated by a martingale M :

$$\begin{aligned} |C_t| &\leq b \sum_i |A_t^i| = b \sum_i |\mathbb{E}[A_T^i | \mathcal{F}_t]| \\ &\leq b \sum_i \mathbb{E}[|A_T^i| | \mathcal{F}_t] =: M_t \end{aligned} \quad (59)$$

As suggested by the case of a locally bounded δ , we often use the differential form

$$dC = \sum_{i=1}^m \delta^i dA^i \quad (60)$$

of the equation $C = C_0 + \delta \cdot A$ as a convenient symbolic equivalent in calculations. One interprets A^i as prices of m zero-dividend assets and δ_t^i as the number of shares invested in them at time t . Then C_t indicates the resultant self-financing portfolio price by equation (57), and equation (60) is the self-financing equation, implying that the change dC in the portfolio price is only due to the changes dA^i in the asset prices with no financing from outside.

Assume for the remainder of this subsection as a way of motivation that A is continuous and $C_t = c(t, A_t)$ for some $C^{1,2}$ function $c(t, a)$.^a Then by equation (60) and Itô's formula, we have

$$\begin{aligned} \frac{\partial c}{\partial t}(t, A_t)dt + \frac{1}{2} \sum_{i,j=1}^m \frac{\partial^2 c}{\partial a_i \partial a_j}(t, A_t) d[A^i, A^j]_t \\ = \sum_{i=1}^m \left(\delta_t^i - \frac{\partial c}{\partial a_i}(t, A_t) \right) dA_t^i \end{aligned} \quad (61)$$

In particular, if $\delta_t^i = \frac{\partial c}{\partial a_i}(t, A_t)$ for all i then $c(t, A_t) = \sum_i \frac{\partial c}{\partial a_i}(t, A_t) A_t^i$ by equation (57) and

$$\frac{\partial c}{\partial t}(t, A_t)dt + \frac{1}{2} \sum_{i,j=1}^m \frac{\partial^2 c}{\partial a_i \partial a_j}(t, A_t) d[A^i, A^j]_t = 0 \quad (62)$$

In general, $\sum_{i,j} \left(\delta^i - \frac{\partial c}{\partial a_i} \right) \left(\delta^j - \frac{\partial c}{\partial a_j} \right) d[A^i, A^j] = 0$ since the (left) right-hand side of equation

(61) has finite variation. Thus, if $[A^i]$ are absolutely continuous and the $m \times m$ matrix $(d/dt[A^i, A^j])$ is nonsingular, then $\delta_t^i = \partial c / \partial a_i(t, A_t)$, so equation (62) holds and $c(t, A_t) = \sum_i \partial c / \partial a_i(t, A_t) A_t^i$. If further the support of A_t is a cone, it follows $c(t, a)$ is *homogeneous of degree 1* in a on that cone.

Assume that $M^i := e^{-\int r dt} A^i$ are local martingales under an equivalent measure for some locally bounded predictable process r . Then $dA^i = r A^i dt + e^{\int r dt} dM^i$; thus, by equations (61) and (57)

$$\begin{aligned} \frac{\partial c}{\partial t}(t, A_t)dt + \frac{1}{2} \sum_{i,j=1}^m \frac{\partial^2 c}{\partial a_i \partial a_j}(t, A_t) d[A^i, A^j]_t \\ = r_t \left(C_t - \sum_{i=1}^m \frac{\partial c}{\partial a_i}(t, A_t) A_t^i \right) dt \end{aligned} \quad (63)$$

Hence, if $c(t, a)$ is homogeneous (in a), then by Euler's formula equation (62) holds (yet δ_t^i may differ from $\partial c / \partial a_i(t, A_t)$ if there are redundancies, for then a regular replicating SFTS is not unique).

Given a homogeneous payoff function $h(a)$, the section Homogeneous Continuous Markovian SFTS constructs under suitable assumptions a homogeneous solution $c(t, a)$ to equation (62) with $c(T, a) = h(a)$. Clearly then, by Euler and Itô formulae, $(\partial c / \partial a_i(t, A_t))$ is an SFTS for A (as observed in [9] and highlighted in [8], see the section The Homogeneous Option Price Function). To this end, we first factor out the homogeneous symmetry of $h(a)$ next.

The Invariance Principle

Let (δ, A) be an SFTS and S a (scalar) semimartingale such that δ is $SA := (SA^1, \dots, SA^m)$ -integrable. Then (δ, SA) is an SFTS. Consequently,

$$d(SC) = \sum_{i=1}^m \delta^i d(SA^i) \quad (64)$$

where $C := \sum_i \delta^i A^i = C_0 + \delta \cdot A$, that is, $SC = S_0 C_0 + \delta \cdot (SA)$. Indeed, by Itô's product rule, then substituting for dC and C_- and regrouping, followed

by Itô's product rule again,

$$\begin{aligned}
d(SC) &= S_-dC + C_-dS + d[S, C] \\
&= S_- \sum_{i=1}^m \delta^i dA^i + \sum_{i=1}^m \delta^i A_-^i dS + \sum_{i=1}^m \delta^i d[S, A^i] \\
&= \sum_{i=1}^m \delta^i (S_-dA^i + A_-^i dS + d[S, A^i]) \\
&= \sum_{i=1}^m \delta^i d(SA^i) \tag{65}
\end{aligned}$$

Interpreting S as an exchange rate, this result [3, 4, 8], called *numéraire invariance*, means that the self-financing property is independent of the base currency. (To the best of our knowledge, the term was coined in the 1992 edition of [3], where a similar proof is given.)

If $S, S_- > 0$, then applied to the semimartingale $1/S$ we see that δ is an SFTS for A if and only if it is one for SA . Thus, if equation (57) holds, then equations (60) and (64) are equivalent.

Assume now that $A^m, A_-^m > 0$ and $m \geq 2$. Define the $n := m - 1$ dimensional semimartingale

$$X := \left(\frac{A^1}{A^m}, \dots, \frac{A^n}{A^m} \right), \quad n := m - 1 \tag{66}$$

Taking $S = 1/A^m$, it follows that δ is an SFTS for A if and only if it is an SFTS for $A/A^m = (X, 1)$, that is, if and only if $F := C/A^m$ satisfies $F = F_0 + \delta' \cdot X$ where $\delta' := (\delta^1, \dots, \delta^n)$. Clearly in this case, $F = \sum_{i=1}^n \delta^i X^i + \delta^m$ and $F_- = \sum_{i=1}^n \delta^i X_-^i + \delta^m$ as $\Delta F = \delta' \cdot \Delta X$. Thus,

$$\delta^m = F - \sum_{i=1}^n \delta^i X^i = F_- - \sum_{i=1}^n \delta^i X_-^i, \quad \left(F := \frac{C}{A^m} \right) \tag{67}$$

(When $m = 1$, a similar argument shows that δ must be a constant, as intuitively obvious.)

Conversely, suppose that δ' is an X -integrable process and F is a process such that $F = F_0 + \delta' \cdot X$. Define δ^m by either of the above formulas—the other then holds as before. Obviously then $\delta = (\delta', \delta^m)$ is an SFTS for $(X, 1)$ with price process F . Hence by

numéraire invariance, δ is an SFTS for A with price process $C = A^m F$, provided δ is A -integrable.

Thus, numéraire invariance shows that *in order to find an SFTS with a given time- T payoff C_T it is sufficient to find processes δ' and F such that $F = F_0 + \delta' \cdot X$ and $F_T = C_T/A_T^m$.*

Since $\delta^m = F - \sum_{i=1}^n \delta^i X^i$, the m th delta δ^m is like F determined by δ' and F_0 . As such, one interprets the m -th asset as the “numéraire asset” chosen to finance an otherwise arbitrary trading strategy δ' in the other assets, post an initial investment of $C_0 = A_0^m F_0$.

We often use the differential form $dF = \sum_{i=1}^n \delta^i dX^i$ of the equation $F = F_0 + \delta' \cdot X$.

Arbitrage-free Semimartingales and Uniqueness

We call a semimartingale $A = (A^1, \dots, A^m)$, $m \geq 2$, *arbitrage free* if there exists a positive semimartingale S with $S_- > 0$ such that SA^i are martingales for all i . Such a process S is called a *state price density* or *deflator* for A . The *law of one price* (with bounded deltas) justifies the terminology:

If A is arbitrage free and δ is a bounded SFTS for A , then SC is a martingale where $C := \sum_{i=1}^m \delta^i A^i$; consequently, $C = 0$ if $C_T = 0$.

Indeed, by numéraire invariance δ is then an SFTS for SA with price process SC . Hence by the section Self-financing Trading Strategies, SC is a martingale, implying $SC = 0$ if $C_T = 0$, and with it $C = 0$, as claimed.

A simple and well-known argument yields that *if $A^m, A_-^m > 0$, then A is arbitrage free if and only if there exists an equivalent probability measure $\mathbb{Q}$ such that X is a $\mathbb{Q}$ -martingale, where $X := \left(\frac{A^1}{A^m}, \dots, \frac{A^n}{A^m} \right)$, $n := m - 1$.*^b Numéraire invariance then implies that C/A^m is a $\mathbb{Q}$ -martingale for the price process $C := \sum_i \delta^i A^i$ of any bounded SFTS δ , and hence

$$C_t = A_t^m \mathbb{E}^{\mathbb{Q}} \left[\frac{C_T}{A_T^m} \mid \mathcal{F}_t \right] \tag{68}$$

Indeed, by numéraire invariance, δ is an SFTS for A/A^m with price process C/A^m . Hence, C/A^m is a $\mathbb{Q}$ -martingale by the section Self-Financing Trading Strategies since A/A^m is a $\mathbb{Q}$ -martingale and δ is bounded.

Suppose that X is a $\mathbb{Q}$ -square-integrable martingale and δ^i are bounded for $i \leq n$. Then $F := C/A^m$ is a $\mathbb{Q}$ -square-integrable martingale

since $dF = \sum_{i=1}^n \delta^i dX^i$ by numéraire invariance. Moreover, $d\langle F \rangle^{\mathbb{Q}} = \sum_{i,j=1}^n \delta^i \delta^j d\langle X^i, X^j \rangle^{\mathbb{Q}}$. Thus, if $\langle X^i \rangle^{\mathbb{Q}}$ are absolutely continuous and the $n \times n$ matrix $(d/dt \langle X^i, X^j \rangle^{\mathbb{Q}})$ is nonsingular, then given any random variable R , there exists at most one SFTS δ for A such that $\sum_{i=1}^m \delta_T^i A_T^i = R$ and δ^i are bounded for $i \leq n$.

Projective Continuous Markovian SFTS

Let $X = (X^1, \dots, X^n)$ be a continuous vector martingale. In this, subsection $x \in \mathbb{R}_+^n$ if $X > 0$ (the main case of interest); otherwise, $x \in \mathbb{R}^n$. Let $g(x)$ be a Borel function of linear growth (so $\mathbb{E}|g(X_T)| < \infty$), and $f(t, x)$ be a continuous function, $C^{1,2}$ on $t < T$. Set $m := n + 1$ and define the C^1 functions

$$\begin{aligned} \delta_i(t, x) &:= \frac{\partial f}{\partial x_i}(t, x), \quad i \leq n, \\ \delta_m(t, x) &:= f(t, x) - \sum_{i=1}^n \delta_i(t, x) x_i \end{aligned} \quad (69)$$

and the continuous vector process

$$\delta = (\delta^1, \dots, \delta^m), \quad \delta_t^i := \delta_i(t, X_t) \quad (70)$$

First suppose that

$$f(t, X_t) = \mathbb{E}[g(X_T) | \mathcal{F}_t] \quad (71)$$

Then the process $F := (f(t, X_t))$ is a martingale, and since X^i are also martingales, Itô's formula yields

$$dF_t = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(t, X_t) dX_t^i, \quad (72)$$

and

$$\frac{\partial f}{\partial t}(t, X_t) dt + \frac{1}{2} \sum_{i,j=1}^n \frac{\partial^2 f}{\partial x_i \partial x_j}(t, X_t) d[X^i, X^j]_t = 0 \quad (73)$$

Clearly, $F_T = g(X_T)$ and equation (72) imply δ is an SFTS for $(X, 1)$ with price process F .

Conversely, suppose that $f(t, x)$ satisfies equation (73) or equivalently, by Itô's formula, equation (72). By equation (72), δ is an SFTS for $(X, 1)$ with price process $F := f(t, X_t)$. Thus by the section Self-financing Trading Strategies, if $\delta_i(t, x)$ are bounded then F is a martingale and if further $f(T, x) = g(x)$ then equation (71) holds. Moreover, as in the section Arbitrage-free Semimartingales and Uniqueness, δ given by equation (70) is then the

unique bounded SFTS for $(X, 1)$ with payoff $g(X_T)$, provided $d[X^i, X^j] = X^i X^j \sigma^{ij} dt$ for some nonsingular matrix process (σ_t^{ij}) .

Example: Projective Deterministic Volatility

Let $X = (X^1, \dots, X^n) > 0$ be a continuous n -dimensional martingale such that

$$d[X^i, X^j]_t = X_t^i X_t^j \sigma_{ij}(t) dt \quad (74)$$

for some n^2 deterministic continuous functions $\sigma_{ij}(t)$. So, $d[\log X^i, \log X^j]_t = \sigma_{ij}(t) dt$. Conditioned on $\mathcal{F}_t$ and unconditionally, X_T/X_t is then *multivariately lognormally distributed*, with mean $(1, \dots, 1)$ and log-covariance matrix $(\int_t^T \sigma_{ij}(s) ds)$. Let $P(t, T, z)$ denote its distribution function. Let $g(x)$ be a Borel function of linear growth. Define the function

$$f(t, x) := \mathbb{E} \left[g \left(x_1 \frac{X_T^1}{X_t^1}, \dots, x_n \frac{X_T^n}{X_t^n} \right) \right] \quad (75)$$

Obviously, $f(T, x) = g(x)$. Clearly, $f(t, x)$ can also be represented in two other ways as

$$\begin{aligned} f(t, x) &= \int_{\mathbb{R}_+^n} g(x_1 z_1, \dots, x_n z_n) P(t, T, dz) \\ &= \mathbb{E} \left[g \left(x_1 \frac{X_T^1}{X_t^1}, \dots, x_n \frac{X_T^n}{X_t^n} \right) \middle| \mathcal{F}_t \right] \end{aligned} \quad (76)$$

Equation (71) holds by the second equality, and $f(t, x)$ is C^1 in t and smooth (even analytic) in x on $t < T$ as seen by changing variable in the integral to $y^i = x^i z^i$ and differentiating under the integral sign in the first equality. Therefore by equation (73), $f(t, x)$ satisfies the PDE

$$\frac{\partial f}{\partial t} + \frac{1}{2} \sum_{i,j=1}^n \sigma_{ij}(t) x_i x_j \frac{\partial^2 f}{\partial x_i \partial x_j} = 0 \quad (77)$$

on the support of X , equation (72) holds, and δ is an SFTS for $(X, 1)$ with price process $F := (f(t, X_t))$, a martingale by equation (71). If $g(x)$ is dx -absolutely continuous with bounded partial derivatives $\frac{\partial g}{\partial x_i}$ (as L^1_{loc} functions), then $g(x)$ has linear growth, $\mathbb{E}|g(X_T)|^p < \infty$ for $p > 0$, and

$$\frac{\partial f}{\partial x_i}(t, x) = \mathbb{E} \left[\frac{X_T^i}{X_t^i} \frac{\partial g}{\partial x_i} \left(x_1 \frac{X_T^1}{X_t^1}, \dots, x_n \frac{X_T^n}{X_t^n} \right) \right] \quad (78)$$

Thus, $\delta_i(t, x) = \frac{\partial f}{\partial x_i}(t, x)$ are bounded. If $g(x) - \sum \frac{\partial g}{\partial x_i} x_i$ is bounded, then so is $\delta_m(t, x)$ as

$$\delta_m(t, x) = \mathbb{E} \left[g \left(x \frac{X_T}{X_t} \right) - \sum_{i=1}^n \frac{\partial g}{\partial x_i} \left(x \frac{X_T}{X_t} \right) \frac{X_T^i}{X_t^i} \right] \quad (79)$$

It further follows that if $f(t, x)$ is any $C^{1,2}$ function with bounded partials $\frac{\partial f}{\partial x_i}(t, x)$ satisfying $f(T, x) = 0$ for all x and the PDE (77), then $F := (f(t, X_t)) = 0$. Indeed, equation (72) then holds by PDE (77) and Itô's formula, implying F is a square-integrable martingale. Thus $F = 0$ since $F_T = 0$. As such, $f(t, x) = 0$ identically if the support of X_t equals $\mathbb{R}_+^n$ for every t . This is so if the matrix $(\sigma_{ij}(t))$ is nonsingular at least near 0, and it is "generically" so even when the matrix has rank 1 but is time dependent.

Projective Continuous SDE SFTS

Continuous Markovian positive martingales $X = (X^1, \dots, X^n)$ often arise as solutions to an SDE system of the form

$$dX_t^i = X_t^i \sum_{j=1}^k \varphi_{ij}(t, X_t) dW_t^j \quad (80)$$

where $W^1, \dots, W^k$ are independent Brownian motions and $\varphi_{ij}(t, x)$, $x \in \mathbb{R}_+^n$, are continuous bounded functions. As is well known, for each $s \leq T$ and $x \in \mathbb{R}_+^n$, there is a unique continuous semimartingale $X^{s,x} = (X_t^{s,x})$ on $[s, T]$ with $X_s^{s,x} = x$ satisfying this SDE; moreover, $X^{s,x}$ is a positive square-integrable martingale (in fact in all $\mathcal{H}^p$) since $\varphi_{ij}(t, x)$ are bounded. Fixing an $X_0 \in \mathbb{R}_+^n$, the solution on $[0, T]$ starting at X_0 at time 0 is denoted as $X = X^{0,X_0}$. The Markov property holds: for any Borel function $g(x)$ of linear growth,

$$\mathbb{E}[g(X_T) | \mathcal{F}_t] = f(t, X_t) \quad \text{where} \quad f(t, x) := \mathbb{E} g(X_T^{t,x}) \quad (81)$$

Clearly $f(T, x) = g(x)$. (Intuitively, $f(t, x) = \mathbb{E}[g(X_T) | X_t = x]$.)

Thus if we assume that $\varphi_{ij}(t, x)$ are sufficiently regular so that $f(t, x)$ is $C^{1,2}$ on $t < T$ for every

bounded (hence of linear growth) Borel function $g(x)$, then the assumptions of the section Projective Continuous Markovian SFTS are satisfied and the conclusions hold. In particular, equation (72) then holds, and since

$$d[X^i, X^j] = X^i X^j \sigma_{ij}(t, X) dt \quad \text{where} \quad \sigma_{ij}(t, x) := \sum_{l=1}^k \varphi_{il}(t, x) \varphi_{jl}(t, x) \quad (82)$$

it follows from equation (73) that, at least on the support of X , $f(t, x)$ satisfies the PDE

$$\frac{\partial f}{\partial t}(t, x) + \frac{1}{2} \sum_{i,j=1}^n x_i x_j \sigma_{ij}(t, x) \frac{\partial^2 f}{\partial x_i \partial x_j}(t, x) = 0 \quad (83)$$

In the deterministic-volatility case, the functions φ_{ij} and hence σ_{ij} are independent of x and simply $X_T^{t,x} = x X_T / X_t$, explaining why in this special case $f(t, x)$ is also given by equation (73).

In general, if $g(x)$ is absolutely continuous with bounded derivatives and the probability transition function of X is sufficiently regular, one shows, as in the deterministic volatility case, that the x -partial derivatives of f (the deltas) are bounded and thereby concludes uniqueness.

If $\sigma_{ij}(t, x)$ are homogeneous of degree 0 in x , then (assumed) uniqueness and symmetry of PDE (73) under dilation in x imply that $f(t, x)$ is homogeneous of degree 1 in x if $g(x)$ is so. By Euler's formula then $\delta_m(t, x) = 0$ in equation (69), implying $(\delta^1, \dots, \delta^n)$ is an SFTS for X .

Homogeneous Continuous Markovian SFTS

Let $A = (A^1, \dots, A^m)$ be a semimartingale with $A, A_- > 0$ such that $X^i := A^i / A^m$ are Itô processes following

$$dX_t^i = X_t^i \sum_{j=1}^k \varphi_t^{ij} (dZ_t^j + \phi_t^j dt) \quad (i = 1, \dots, n := m - 1) \quad (84)$$

where Z^j are independent Brownian motions and ϕ^j , φ^{ij} are locally bounded predictable processes with

φ^{ij} bounded and $\mathbb{E} e^{1/2 \sum_j \int_0^T (\phi_j^i)^2 dt} < \infty$. Define the martingale

$$\begin{aligned} M &:= \mathcal{E} \left(- \sum_{j=1}^k \int \phi^j dZ^j \right) \\ &= e^{-\sum_{j=1}^k \left(\int \phi^j dZ^j + \frac{1}{2} \int (\phi^j)^2 dt \right)} \end{aligned} \quad (85)$$

and the measure $\mathbb{Q}$ by $d\mathbb{Q} = M_T d\mathbb{P}$. Then $W^i := Z^i + \int \phi^i dt$ are $\mathbb{Q}$ -Brownian motions and are $\mathbb{Q}$ -independent since $[W^k, W^l] = 0$ for $k \neq l$. Hence, X^i are $\mathbb{Q}$ -square-integrable martingales as $dX^i = X^i \sum_{j=1}^k \varphi^{ij} dW^j$ and φ^{ij} are bounded. Thus, A is arbitrage free.

Now let $h(a)$, $a \in \mathbb{R}_+^m > 0$, be a homogeneous function of linear growth. Define $g(x) := h(x, 1)$, $x \in \mathbb{R}_+^n$. Assume further that $\varphi_i^{ij} = \varphi_{ij}(t, X_t)$ for some continuous bounded functions $\varphi_{ij}(t, x)$. Then equation (80) holds, and hence the section Projective Continuous SDE SFTS applied under measure $\mathbb{Q}$ shows that X is $\mathbb{Q}$ -Markovian in that $\mathbb{E}^{\mathbb{Q}}[g(X_T) | \mathcal{F}_t] := f(t, X_t)$ where $f(t, x) = \mathbb{E}^{\mathbb{Q}}[g(X_T^{t,x})]$, as in equation (81). Thus, by the section Projective Continuous SDE SFTS, equations (72) and (73) hold and δ as defined in equation (70) is an SFTS for $(X, 1)$. Therefore by numéraire invariance, δ is an SFTS for A with price process $C = A^m F$. The homogeneity of $h(a)$ further implies $C_T = A_T^m g(X_T) = h(A_T)$.

We have thus constructed an SFTS with the given payoff $h(A_T)$. As in the section Example: Projective Deterministic Volatility or Projective Continuous SDE SFTS, we ensure its boundedness by requiring the x -partial derivatives of $g(x)$ or equivalently a -partial derivatives of $h(a)$ (as L_{loc}^1 functions) be bounded and thereby get unique pricing. For (very) low dimensions n , the PDE (83) is suitable for numerical valuation in the absence of a closed-form solution.

Although the option price process and the deltas are already found, let us also consider the homogeneous option price function referred to in the section Self-financing Trading Strategies, and now naturally defined by

$$c(t, a) := a_m f \left(t, \frac{a^1}{a^m}, \dots, \frac{a^n}{a^m} \right) \quad (86)$$

Then $C_t = c(t, A_t)$. Agreeably, $\delta_t^i = \frac{\partial c}{\partial a_i}(t, A_t)$ by equation (69). (For $i = m$ use Euler's formula for $c(t, a)$.) By the continuity of X and equation (69), $\delta_t^i = \frac{\partial c}{\partial a_i}(t, A_{t-})$ too. Therefore by Itô's formula,

$$\frac{\partial c}{\partial t}(t, A_{t-})dt + \frac{1}{2} \sum_{i,j=1}^m \frac{\partial^2 c}{\partial a_i \partial a_j}(t, A_{t-})d[A^i, A^j]_t^c = 0 \quad (87)$$

(The term for the sum of jumps in Itô's formula vanishes since $\Delta C = \sum \delta^i \Delta A^i$.) This yields the PDE $\frac{\partial c}{\partial t} + \frac{1}{2} \sum_{i,j} a_i a_j \sigma_{ij}^A(t, a) \frac{\partial^2 c}{\partial a_i \partial a_j} = 0$ for the special case $d[A^i, A^j]_t = A_t^i A_t^j \sigma_{ij}^A(t, A_t)dt$ for some functions $\sigma_{ij}^A(t, a)$. The quotient-space PDE (83) is more fundamental for it holds in general (even when A is discontinuous) and has one lower dimension. Change of variable $L^i = \frac{X^i}{X^{i+1}} - 1$ ($i < n$), $L^n = X^n - 1$, transforms equation (83) to the Libor market model PDE.

Multivariate Poisson Predictable Representation

Let $P = (P^1, \dots, P^k)$ be a vector of independent Poisson processes P^i with intensities $\lambda_i > 0$. For any C^1 in t function $u(t, p)$, $p \in \mathbb{R}^k$, the process $u(t, P) = (u(t, P_i))$ is a finite activity semimartingale, and using $[P^i, P^j] = 0$, one has $\Delta u(t, P) = \sum_i \Delta_i u(t, P_-) \Delta P^i$, where

$$\Delta_i u(t, p) := u(t, p_1, \dots, p_i + 1, \dots, p_n) - u(t, p) \quad (88)$$

denotes the i th forward partial difference of $u(t, p)$ in p . This in turn readily implies

$$du(t, P) = \frac{\partial u}{\partial t}(t, P_-)dt + \sum_{i=1}^k \Delta_i u(t, P_-)dP^i \quad (89)$$

Let $v(p)$, $p \in \mathbb{R}^k$ be a function of exponential linear growth. Define the function

$$\begin{aligned} u(t, p) &:= \sum_{q_1, \dots, q_k=0}^{\infty} v(p + q) \\ &\times \prod_{i=1}^k \frac{\lambda_i^{q_i}}{q_i!} (T - t)^{q_i} e^{-\lambda_i(T-t)} \quad (p \in \mathbb{R}^k) \end{aligned} \quad (90)$$

Clearly, $u(T, p) = v(p)$. Since the unconditional distribution of P_{T-} is Poisson and is the same as the distribution of $P_T - P_t$ conditioned on $\mathcal{F}_t$, we have

$$\begin{aligned} u(t, p) &= \mathbb{E}[v(p + P_T - P_t)] \\ &= \mathbb{E}[v(p + P_T - P_t) | \mathcal{F}_t] \end{aligned} \quad (91)$$

Hence, $u(t, P_t) = \mathbb{E}[v(P_T) | \mathcal{F}_t]$. (Intuitively, $u(t, p) = \mathbb{E}[v(P_T) | P_t = p]$.) Thus, the process

$$F = (F_t), \quad F_t := u(t, P_t) = \mathbb{E}[v(P_T) | \mathcal{F}_t] \quad (92)$$

is a martingale. But so are $P^j - \lambda_j t$. Therefore in view of equation (89), it follows that

$$dF = \sum_{i=1}^k \Delta_i u(t, P_-) d(P^i - \lambda_i t) \quad (93)$$

and $u(t, p)$ satisfies the equation

$$\frac{\partial u}{\partial t}(t, P_{t-}) + \sum_{i=1}^k \lambda_i \Delta_i u(t, P_{t-}) = 0 \quad (94)$$

Since $F_T = v(P_T)$ and $F_0 = u(0, 0)$, combining equations (90) and (93) yields the following representation:

$$\begin{aligned} v(P_T) &= \sum_{q_1, \dots, q_k=0}^{\infty} v(q_1, \dots, q_k) \prod_{i=1}^k \frac{\lambda_i^{q_i}}{q_i!} T^{q_i} e^{-\lambda_i T} \\ &\quad + \sum_{i=1}^k \int_0^T \Delta_i u(t, P_{t-}) d(P_t^i - \lambda_i t) \end{aligned} \quad (95)$$

Projective Exponential-Poisson SFTS

Let $P = (P^1, \dots, P^k)$ be a vector of independent Poisson processes P^j with intensities $\lambda_j > 0$. Let $X_0 \in \mathbb{R}_+^n$, $n \geq k$, and $\beta = (\beta_{ij})$ be an $n \times k$ matrix such that the $n \times k$ matrix $(e^{\beta_{ij}} - 1)$ has full rank. Then the processes $X^i := (x_i(t, P_t))$, $i = 1, \dots, n$, are square-integrable martingales (in fact in all $\mathcal{H}^p$), where

$$\begin{aligned} x_i(t, p) &:= X_0^i \exp \left(\sum_{j=1}^k (\beta_{ij} p_j - (e^{\beta_{ij}} - 1) \lambda_j t) \right) \\ (p \in \mathbb{R}^k) \end{aligned} \quad (96)$$

Since

$$\begin{aligned} \frac{\partial x_i}{\partial t}(t, p) &= -x_i(t, p) \sum_{j=1}^k (e^{\beta_{ij}} - 1) \lambda_j, \\ \Delta_j x_i(t, p) &= x_i(t, p) (e^{\beta_{ij}} - 1) \end{aligned} \quad (97)$$

it follows from equation (89) (or easily also from Itô's formula) that

$$\begin{aligned} dX^i &= X_-^i \sum_{j=1}^k (e^{\beta_{ij}} - 1) \\ &\quad \times d(P^j - \lambda_j t) \quad (X_t^i := x_i(t, P_t)) \end{aligned} \quad (98)$$

Let $\alpha = (\alpha_{ij})$ be any $n \times k$ matrix such that $\sum_i (e^{\beta_{il}} - 1) \alpha_{ij} = \delta_{jl}$, all $1 \leq j, l \leq k$. Then

$$d(P^j - \lambda_j t) = \sum_{i=1}^n \alpha_{ij} \frac{dX^i}{X_-^i} \quad (99)$$

Now let $g(x)$, $x \in \mathbb{R}_+^n$, be a function of linear growth; define the function

$$v(p) := g(x_1(T, p), \dots, x_n(T, p)), \quad (p \in \mathbb{R}^n) \quad (100)$$

and the function $u(t, p)$ by equation (90). By the section Multivariate Poisson Predictable Representation, $F := (u(t, P_t))$ is a martingale with $F_T = v(P_T) = g(X_T)$ and is represented as equation (93). Substituting equation (99) into equation (93) yields

$$dF = \sum_{i=1}^n \delta^i dX^i \quad (101)$$

where

$$\delta_t^i := \frac{1}{X_{t-}^i} \sum_{j=1}^k \alpha_{ij} \Delta_j u(t, P_{t-}) \quad (102)$$

Thus, $\delta = (\delta^1, \dots, \delta^m)$ is an SFTS for $(X, 1)$ where $m := n + 1$ and $\delta^m := F - \sum_{i=1}^n \delta^i X^i$.

It is more desirable to express δ in terms of X . One has $u(t, p) = f(t, x(t, p))$, where

$$\begin{aligned}
f(t, x) &:= \mathbb{E} \left[g \left(x \frac{X_T}{X_t} \right) \right] = \mathbb{E} \left[g \left(x \frac{X_T}{X_t} \right) \mid \mathcal{F}_t \right] \\
&= \sum_{q_1, \dots, q_n=0}^{\infty} g \left(x_1 e^{\sum_{j=1}^n (\beta_{1j} q_j - (e^{\beta_{1j}} - 1) \lambda_j (T-t))}, \dots, x_n e^{\sum_{j=1}^n (\beta_{nj} q_j - (e^{\beta_{nj}} - 1) \lambda_j (T-t))} \right) \prod_{i=1}^n \frac{\lambda_i^{q_i}}{q_i!} (T-t)^{q_i} e^{-\lambda_i (T-t)}
\end{aligned} \tag{103}$$

The equalities follow from the definition of $v(p)$ above and of $u(t, p)$ in equation (90) together with the two formulae following it.^c We clearly have $f(T, x) = g(x)$ and

$$F_t := u(t, P_t) = f(t, X_t) = \mathbb{E}[g(X_T) \mid \mathcal{F}_t] \tag{104}$$

Since $u(t, p) = f(t, x(t, p))$, the deltas in equation (102) are given by partial differences of $f(t, x)$ as

$$\begin{aligned}
\delta_t^i &= \delta_i(t, X_{t-}) \quad \text{where} \\
\delta_i(t, x) &:= \frac{1}{x_i} \sum_{j=1}^k \alpha_{ij} (f(t, e^{\beta_{1j}} x_1, \dots, e^{\beta_{nj}} x_n) \\
&\quad - f(t, x))
\end{aligned} \tag{105}$$

We have unique pricing since $(X, 1)$ is arbitrage free (as X^i are martingales). Specifically, if $\hat{\delta}$ is another SFTS for $(X, 1)$ with payoff $\hat{F}_T = g(X_T)$, then $\hat{F} := \sum_{i=1}^n \hat{\delta}^i X^i + \hat{\delta}^m = F$ provided that either all $\hat{\delta}^i$, $i \leq n$ are bounded or all $\hat{\delta}^i - \delta^i$, $i \leq n$ are bounded.

Indeed, then $\hat{F} = \hat{F}_0 + \hat{\delta}' \cdot X$ is a martingale, since X is square integrable (in the second case, also use that F is a martingale). Hence, $\hat{F} = F$ as $\hat{F}_T = F_T$.

Moreover, if $k = n$ we have unique hedging, that is, $\hat{\delta} = \delta$ for any bounded SFTS $\hat{\delta}$ for $(X, 1)$ with payoff $\hat{F}_T = g(X_T)$. Indeed, $\hat{F} = F$, as before; thus, setting $\theta^i := \hat{\delta}^i - \delta^i$ gives

$$\begin{aligned}
0 &= d(\hat{F} - F) = \sum_{i,j=1}^n \theta^i \theta^j d\langle X^i, X^j \rangle \\
&= \sum_{i,j=1}^n \theta^i \theta^j X_-^i X_-^j \sum_{l=1}^n (e^{\beta_{il}} - 1)(e^{\beta_{jl}} - 1) \lambda_l dt
\end{aligned} \tag{106}$$

the last equality following from equation (98). However, the $n \times n$ matrix $(\sum_{l=1}^n (e^{\beta_{il}} - 1)(e^{\beta_{jl}} - 1) \lambda_l)_{i,j=1}^n$ is nonsingular. Therefore, $\theta^i = 0$, that is, $\hat{\delta}^i = \delta^i$ for $i \leq n$, implying $\hat{\delta}^m = \delta^m$ too as $\hat{F} = F$.

One shows, as in the section Exponential-Poisson Exchange Option Model, that the processes δ^i are bounded if $\gamma_i(x)$ are bounded, where

$$\begin{aligned}
\gamma_i(x) &:= \frac{1}{x_i} \sum_{j=1}^k \alpha_{ij} (g(e^{\beta_{1j}} x_1, \dots, e^{\beta_{nj}} x_n) - g(x)), \\
\gamma_m(x) &:= g(x) - \sum_{i=1}^n \gamma_i(x) x_i
\end{aligned} \tag{107}$$

Homogeneous Exponential-Poisson SFTS

Let $A > 0$ be an m -dimensional semimartingale with $A_- > 0$ and set $X := (A^i / A^m)_{i=1}^n$, $n := m - 1$, as before. Assume that

$$dX_t^i = X_{t-}^i \sum_{j=1}^k (e^{\beta_{ij}} - 1) (dP_t^j - \lambda_t^j dt) \tag{108}$$

where $1 \leq k \leq n$, β_{ij} are constants with the $n \times k$ matrix $(e^{\beta_{ij}} - 1)$ of full rank, $\lambda^j > 0$ are bounded predictable processes, and P^j are semimartingales with $[P^j, P^l] = 0$ for $j \neq l$ such that $[P^j] = P^j$, $P_0^j = 0$, and $P^j - \int \kappa^j dt$ are local martingales for some locally bounded predictable processes $\kappa^j > 0$. Assume further that $\mathbb{E} \exp \left(\sum_{j=1}^k \int_0^T \left(\frac{\lambda_t^j}{\kappa_t^j} - 1 \right)^2 \kappa_t^j dt \right) < \infty$.

Owing to the above growth condition, the positive local martingale

$$\begin{aligned} M &:= \mathcal{E} \left(\sum_{j=1}^k \int \left(\frac{\lambda^j}{\kappa^j} - 1 \right) (dP^j - \kappa^j dt) \right) \\ &= e^{-\sum_{j=1}^k \int (\lambda^j - \kappa^j) dt} \prod_{s \leq \cdot} \left(1 + \sum_{j=1}^n \left(\frac{\lambda_s^j}{\kappa_s^j} - 1 \right) \Delta P_s^j \right) \end{aligned} \quad (109)$$

is a martingale. Define the measure $\mathbb{Q}$ by $d\mathbb{Q} = M_T d\mathbb{P}$. As in the section Exponential-Poisson Model Uniqueness, $\int \lambda^j dt$ are the $\mathbb{Q}$ -compensator of P^j . This, equation (108), and boundedness of λ^j imply that X^i are $\mathbb{Q}$ -square integrable martingales. Thus, A is arbitrage free. As before, the SDE (108) integrates to

$$X_t^i = X_0^i e^{\sum_{j=1}^k \beta_{ij} P_t^j - (e^{\beta_{ij}} - 1) \int_0^t \lambda_s^j ds} \quad (110)$$

Now assume λ^j are constant. Then P^j are $\mathbb{Q}$ -Poisson processes with intensities λ^j and are independent since $[P^j, P^l] = 0$, $j \neq l$. Let $h(a)$, $a \in \mathbb{R}_+^m$, be a homogeneous function of linear growth. Define $g(x) := h(x, 1)$, $x \in \mathbb{R}_+^n$. The section Projective Exponential-Poisson SFTS applied under $\mathbb{Q}$ implies that δ given by equation (105) (with $\delta^m = F - \sum_{i=1}^n \delta^i X^i$) is an SFTS for $(X, 1)$ with price process $F = (f(t, X_t))$ satisfying $F_T = g(X_T)$, where $f(t, x)$ is defined explicitly by equation (103), or equivalently, $f(t, x) = \mathbb{E}^{\mathbb{Q}} g(x X_T / X_t)$. Therefore, by numéraire invariance, δ is an SFTS for A with price process $C := A^m F$ satisfying $C_T = A^m g(X_T) = h(A_T)$ by homogeneity.

Assume finally that the payoff function $h(a)$ is such that the functions $\gamma_i(x)$ defined in equation (107) are bounded (e.g., $h(a) = \max(a^1, \dots, a^m)$). By the section Projective Exponential-Poisson SFTS, if $k = n$, then δ is the unique bounded SFTS for A with payoff $C_T = h(A_T)$. In general, since A is arbitrage free, $\hat{C} = C$ for any other bounded SFTS $\hat{\delta}$ for A with payoff $\hat{C}_T = h(A_T)$, where $\hat{C} := \sum_i \hat{\delta}^i A^i$.

End Notes

^aClearly, then the restriction of (any such) $c(t, a)$ to the support of A is unique, and if $\hat{c}(t, a)$ is any function that equals $c(t, a)$ on the support of A , then $C_t = \hat{c}(t, A_t)$ too.

If the support of A_t is a proper surface, for example, if $m = 2$ and A^2 is deterministic as in the Black–Scholes model or $A_t^2 = a_2(t, A_t^1)$ as in Markovian short-rate models, then obviously there exist infinitely many nonhomogeneous functions $\hat{c}(t, a)$ such that $C_t = \hat{c}(t, A_t)$. (Such a homogeneous function also exists under some assumptions as in the section Homogeneous Continuous Markovian SFTS.)

^bIndeed, first assume that A is arbitrage free and let S be a state price density. The martingale $M := \frac{SA^m}{\mathbb{E}[S_0 A_0^m]}$ clearly satisfies $\mathbb{E} M_T = 1$. Hence, the equivalent measure $\mathbb{Q}$ defined by $d\mathbb{Q} = M_T d\mathbb{P}$ is a probability measure.

Since $MX^i = \frac{SA^i}{\mathbb{E}[S_0 A_0^m]}$ is a martingale, X^i is a $\mathbb{Q}$ -martingale by Bayes' rule. Conversely, assume that X^i are $\mathbb{Q}$ -martingales for some $\mathbb{Q}$. Define $M_t := \mathbb{E} \left[\frac{d\mathbb{Q}}{d\mathbb{P}} \mid \mathcal{F}_t \right] > 0$. Then (the right continuous version of) $M = (M_t)$ is a martingale (so $M_- > 0$). By Bayes' rule MX^i are martingales since X^i are $\mathbb{Q}$ -martingales. Set $S := M/A^m$. Then $S, S_- > 0$ and $SA^i = MX^i$. Thus S is a deflator, as desired. Further, since SC is a martingale for any bounded SFTS δ , by the Bayes' rule $SC/M = C/A^m$ is a $\mathbb{Q}$ -martingale.

^cThe projective option price function $f(t, x) := \mathbb{E} \left[g \left(x \frac{X_T}{X_t} \right) \right]$, also encountered for the log-Gaussian case in equation (75), satisfies $f(t, X_t) = \mathbb{E}[g(X_T) \mid \mathcal{F}_t]$ in general when X is the exponential of any n -dimensional process of independent increments (inhomogeneous Lévy process), but we no longer have hedging in general.

References

- [1] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economics* **81**, 637–659.
- [2] Delbaen, F. & Schachermayer, W. (2006). *The Mathematics of Arbitrage*, Springer.
- [3] Duffie, D. (2001). *Dynamic Asset Pricing Theory*, 3rd Edition, Princeton University Press.
- [4] El-Karoui, N., Geman, H. & Rochet, J.C. (1995). Change of numéraire, change of probability measure, and option pricing, *Journal of Applied Probability* **32**, 443–458.
- [5] Harrison, M.J. & Kreps, D.M. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- [6] Harrison, M.J. & Pliska, S. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and their Applications* **11**, 215–260.
- [7] Jamshidian, F. (1993). Options and futures evaluation with deterministic volatilities, *Mathematical Finance* **3**(2), 149–159.
- [8] Margrabe, W. (1978). The value of an option to exchange one asset for another, *Journal of Finance* **33**, 177–186.

- [9] Merton, R. (1973). Theory of rational option pricing, *Bell Journal of Economics* **4**(1), 141–183.
- [10] Neuberger, A. (1990). *Pricing Swap Options Using the Forward Swap Market*, IFA Preprint.

eign Exchange Options; Forward–Backward Stochastic Differential Equations (SDEs); Hedging; Itô’s Formula; Markov Processes; Martingales; Poisson Process.

Related Articles

FARSHID JAMSHIDIAN

Arbitrage Strategy; Caps and Floors; CMS Spread Products; Equivalent Martingale Measures; For-

Binomial Tree

This model, introduced by Cox *et al.* [1] in 1979, has played a decisive role in the development of the derivatives industry. Its simple structure and easy implementation gave analysts the ability to price a huge range of financial derivatives in an almost routine way. Nowadays its value is largely pedagogical, in that the whole theory of arbitrage pricing in complete markets can be explained in a couple of pages in the context of the binomial model. The model is covered in every mathematical finance textbook, but we mention in particular [4], which is entirely devoted to the binomial model, and [2] for a careful treatment of American options.

The One-period Model

Suppose we have an asset whose price is S today and whose price tomorrow can only be one of two known values S_0, S_1 (we take $S_0 > S_1$); see Figure 1. This apparently highly artificial situation is the kernel of the binomial model. We also suppose there is a bank account paying a daily rate of interest r_1 , so that \$1 today is worth $\$R = \$(1 + r_1)$ tomorrow. We assume that borrowing is possible from the bank at the same rate of interest r_1 , and that the risky asset can also be borrowed (sold short, in the usual financial terminology). The only other assumption is that

$$S_1 < RS < S_0 \quad (1)$$

If $RS \leq S_1$, we could borrow $\$B$ from the bank and buy B/S shares of the risky asset. Tomorrow these will be worth at least $S_1 B/S$, while only RB has to be repaid to the bank, leaving a profit of either $B(S_1 - RS)/S$ or $B(S_0 - RS)/S$. Both of these are nonnegative and at least one is strictly positive. This is an *arbitrage opportunity*: no initial investment, no loss and the chance of a positive profit at the end

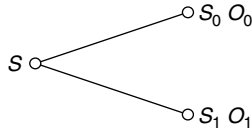


Figure 1 One-period binomial tree

(see **Arbitrage Strategy**). There is also an arbitrage opportunity if $RS > S_0$, realized by short-selling the risky asset.

A *derivative security*, *contingent claim*, or *option* is a contract that pays tomorrow an amount that depends only on tomorrow's asset price. Thus any such claim can only have values, say O_0 and O_1 corresponding to "underlying" prices S_0, S_1 , as shown in Figure 1.

Suppose we form a portfolio today consisting of N shares of the risky asset and $\$B$ in the bank (either of both of N, B could be negative). The value today of this portfolio is $p = B + NS$ and its value tomorrow will be $RB + NS_0$ or $RB + NS_1$. Now choose B, N such that

$$\begin{aligned} RB + NS_0 &= O_0 \\ RB + NS_1 &= O_1 \end{aligned} \quad (2)$$

There is a unique solution as long as $S_1 \neq S_0$, given by

$$N^* = \frac{O_0 - O_1}{S_0 - S_1}, \quad B^* = \frac{1}{R}(O_0 - NS_0) \quad (3)$$

With these choices, the portfolio value tomorrow exactly coincides with the derivative security payoff, whichever way the price moves. If the derivative security is offered today for any price other than $p = RB^* + N^*S$ there is an arbitrage opportunity (realized by "borrowing the portfolio" and buying the option or conversely). Thus "arbitrage pricing" reduces to the solution of a pair of simultaneous linear equations.

It is easily checked that $p = (q_0 O_0 + q_1 O_1)/R$ where

$$q_0 = \frac{RS - S_1}{S_0 - S_1}, \quad q_1 = \frac{S_0 - RS}{S_0 - S_1} \quad (4)$$

We see that q_0, q_1 depend only on the underlying market parameters, not on O_0 or O_1 , that $q_0 + q_1 = 1$ and that $q_0, q_1 > 0$ if and only if the no-arbitrage condition (1) holds. Thus under this condition q_0, q_1 define a probability measure Q and we can write the price of the derivative as

$$p = E_Q \left(\frac{1}{R} O \right) \quad (5)$$

Note that Q , the so-called *risk-neutral* measure, emerges from the “no-arbitrage” argument. We said nothing in formulating the model about the probability of an upward or downward move and the above argument does not imply that this probability has to be given by Q . A further feature of Q is that if we compute the expected price tomorrow under Q we find that

$$S = \frac{1}{R}(q_0 S_0 + q_1 S_1) \quad (6)$$

showing that the discounted price process is a Q -martingale. This is summarized as follows:

- Under condition (1) there is a unique arbitrage-free price for the contingent claim.
- Condition (1) is equivalent to the existence of a unique probability measure Q under which the discounted asset price is a martingale.
- The contingent claim value is obtained by computing the discounted expectation of its exercise value with respect to a certain probability measure Q .

Much of the classic theory of mathematical finance (see **Fundamental Theorem of Asset Pricing; Risk-neutral Pricing**) is concerned with identifying conditions under which these three statements hold for the more general price models. They hold in particular for the multiperiod models discussed below.

The Multiperiod Model

More realistic models can be obtained by generalizing the binomial model to n periods. We consider a discrete-time price process $S(i)$, $i = 0, \dots, n$ such that, at each time i , $S(i)$ takes one of $i + 1$ values $S_{i0} > S_{i1} > \dots > S_{ii}$. While we could consider general values for these constants, the most useful case is that in which the price moves “up” by a factor u or “down” by a factor $d = 1/u$, giving a recombining tree with $S_{ij} = Su^{i-2j}$ where $S = S(0)$; see Figure 2 for the two-period case. We can define a probability measure Q by specifying that $P[S(i+1) = uS(i)|S(i)] = q_0$ and $P[S(i+1) = dS(i)|S(i)] = q_1$ where q_0 and q_1 are given by equation (4) above; in this case, $q_0 = (Ru - 1)/(u^2 - 1)$, $q_1 = 1 - q_0$. Thus $S(i)$ is a discrete-time Markov process under Q with homogeneous transition probabilities. Specifically, it is a *multiplicative random walk* in that each

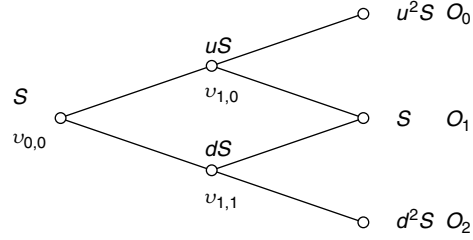


Figure 2 Two-period binomial tree

successive value is obtained from the previous one by multiplication by an independent positive random factor.

Consider the two-period case of Figure 2 and a contingent claim with exercise value O at time 2 where $O = O_0, O_1, O_2$ in the three states as shown. By the one-period argument, the no-arbitrage price for the claim at time 1 is $v_{1,0} = (q_0 O_0 + q_1 O_1)/R$ if the price is uS and $v_{1,1} = (q_0 O_1 + q_1 O_2)/R$ if the price is dS . However, now our contingent claim is equivalent to a one-period claim with payoff $v_{1,0}, v_{1,1}$, so its value at time 0 is just $(q_0 v_{1,0} + q_1 v_{1,1})/R$, which is equal to

$$v_{0,0} = E_Q \left[\frac{1}{R^2} O \right] \quad (7)$$

Generalizing to n periods and a claim that pays amounts $O_0, \dots, O_n$ at time n , the value at time 0 is

$$v_{0,0} = E_Q \left[\frac{1}{R^n} O \right] = \frac{1}{R^n} \sum_{j=0}^n C_j^n q_0^{n-j} q_1^j O_j \quad (8)$$

where C_j^n is the binomial coefficient $C_j^n = n!/(j!(n-j)!)$. From equation (3) the initial hedge ratio (the number N of shares is the hedging portfolio at time 0) is

$$\begin{aligned} N &= \frac{v_{1,0} - v_{1,1}}{uS - dS} \\ &= \frac{1}{SR^{n-1}(u-d)} \sum_{j=0}^{n-1} C_j^{n-1} q_0^{n-1-j} q_1^j (O_j - O_{j+1}) \end{aligned} \quad (9)$$

For example, suppose $S = 100$, $R = 1.001$, $u = 1.04$, $n = 25$, and O is a call option with strike $K = 100$, so that $O_j = [Su^{n-2j} - K]^+$. The option value

is $v_{0,0} = 9.086$ and $N = 0.588$. The initial holding in the bank is therefore $v_{0,0} - NS = -49.72$. This is the typical situation: hedging involves leverage (borrowing from the bank to invest in shares).

Scaling the Binomial Model

Now let us consider scaling the binomial model to a continuous limit. Take a fixed time horizon T and think of the price $S(i)$ above, now written $S_n(i)$, as the price at time $iT/n = i\Delta t$. Suppose the continuously compounding rate of interest is r , so that $R = e^{r\Delta t}$. Finally, define $h = \log u$ and $X(i) = \log(S(i)/S(0))$; then $X(i)$ is a random walk on the lattice $\{\dots -2h, -h, 0, h, \dots\}$ with right and left probabilities q_0, q_1 as defined earlier and $X(0) = 0$. If we now take $h = \sigma\sqrt{\Delta t}$ for some constant σ , we find that

$$q_0, q_1 = \frac{1}{2} \pm \frac{h}{2\sigma^2} \left(r - \frac{1}{2}\sigma^2 \right) + O(h^2) \quad (10)$$

Thus $Z(i) := X(i) - X(i-1)$ are independent random variables with

$$\begin{aligned} EZ(i) &= \frac{h^2}{\sigma^2} \left(r - \frac{1}{2}\sigma^2 \right) + O(h^3) \\ &= \left(r - \frac{1}{2}\sigma^2 \right) \Delta t + O(n^{-3/2}) \end{aligned} \quad (11)$$

and

$$\text{var}(Z(i)) = \sigma^2 \Delta t + O(n^{-2}) \quad (12)$$

Hence $X_n(T) := X(n) = \sum_{i=1}^n Z(i)$ has mean μ_n and variance V_n such that $\mu_n \rightarrow (r - \sigma^2/2)T$ and $V_n \rightarrow \sigma^2 T$ as $n \rightarrow \infty$. By the central limit theorem, the distribution of $X_n(T)$ converges weakly to the normal distribution with the limiting mean and variance. If the contingent claim payoff is a continuous bounded function $O = h(S_n(n))$, then the option value converges to a normal expectation that can be written as

$$\begin{aligned} V_0(S) &= \frac{e^{-rT}}{\sqrt{2\pi}} \int_{-1}^1 h \left(S \exp \left(r - \frac{1}{2}\sigma^2 \right) T \right. \\ &\quad \left. + \sigma\sqrt{T}x \right) e^{-\frac{1}{2}x^2} dx \end{aligned} \quad (13)$$

This is the Black–Scholes formula. It can be given in more explicit terms when, for example, $h(S) = [S - K]^+$, the standard call option (see **Black–Scholes Formula**).

American Options

In the multiperiod binomial model, the basic computational step is the backward recursion

$$v_{i-1,j} = \frac{1}{R} (q_0 v_{i,j} + q_1 v_{i,j+1}) \quad (14)$$

defining the values at time step $i-1$ from those at time i by discounted conditional expectation, starting with the exercise values $v_{n,j} = O_j$ at the final time n . In an American option, we have the right to exercise at any time, the exercise value at time i being some given function $h(i, S_i)$, for example, $h(i, S_i) = [K - S_i]^+$ for an American put. The exercise value at node (i, j) in the binomial tree is therefore $\tilde{h}(i, j) = h(i, Su^{i-2j})$. In this case, it is natural to replace equation (14) by

$$v_{i-1,j} = \max\{v_{i-1,j}^c, \tilde{h}(i-1, j)\} \quad (15)$$

where $v_{i-1,j}^c$ is given by the right-hand side of equation (14). At each node $(i-1, j)$, we compare the “continuation value” $v_{i-1,j}^c$ with the “immediate exercise” value $\tilde{h}(i-1, j)$ and take the larger value. This intuition is correct, and the value $v_{0,0}$ obtained by applying equation (15) for $i = n, n-1, \dots, 1$ with starting condition $v_{n,j} = \tilde{h}(n, j)$ is the unique arbitrage-free value of the American option at time 0.

The reader should refer to **American Options** for a complete treatment, but, in outline, the argument establishing the above claim is as follows. The algorithm divides the set of nodes into two, the *stopping set* $\mathcal{S} = \{(i, j) : v_{i,j} = \tilde{h}(i, j)\}$ and the complementary *continuation set* $\mathcal{C}$. By definition, $(n, j) \in \mathcal{S}$ for $j = 0, \dots, n$. Let τ^* be the stopping time $\tau^* = \min\{i : S_i \in \mathcal{S}\}$. Then τ^* is the optimal time at which the holder of the option should exercise. The process $V_i = v_{i,S_i}/R^i$ is a supermartingale, while the stopped process $V_{i \wedge \tau^*}$ is a martingale with the property that $V_{i \wedge \tau^*} \geq h(i \wedge \tau^*, S_{i \wedge \tau^*})/R^i$. These facts follow from the general theory of optimal stopping, but are not hard to establish directly in the present case. The value $V_{i \wedge \tau^*}$ can be replicated by trading in the underlying asset (using the basic hedging strategy (3))

derived for the one-period model). It follows that this strategy (call it $\mathcal{SR}$) is the *cheapest superreplicating strategy*, that is, $x = v_{0,0}$ is the minimum capital required to construct a trading strategy with value X_i at time i with the property that $X_i \geq h(i, S_i)$ for all i almost surely. If the seller of the option is paid more than $v_{0,0}$, then he or she can put the excess in the bank and employ the trading strategy $\mathcal{SR}$, which is guaranteed to cover his or her obligation to the buyer whenever he or she chooses to exercise. Conversely, if the seller will accept $p < v_{0,0}$ for the option then the buyer should short $\mathcal{SR}$, obtaining an initial value $v_{0,0}$ of which p is paid to the seller and $v_{0,0} - p$ placed in (for clarity) a second bank account. The short strategy has value $-X_i$ and the buyer exercises at τ^* , receiving from the seller the exercise value $h(\tau^*, S_{\tau^*}) = X_{\tau^*}$, which is equal and opposite to the value of the short hedge at τ^* . Thus, there is an arbitrage opportunity for one party or the other unless the price is $v_{0,0}$.

The impact of the binomial model as introduced by Cox *et al.* [1] is largely due to the fact that the European option pricer can be turned into an American option pricer by a trivial one-line modification of the code. Pricing American options in (essentially) the Black–Scholes model was recognized as a free-boundary problem in partial differential equations (PDE) by McKean [3] in 1965, but the only computational techniques were PDE methods (see **Finite Difference Methods for Early Exercise Options**) generally designed for much more complicated problems.

Computations in the Binomial Model

Nowadays the binomial model is rarely, if ever, used for practical problems, largely because it is comprehensively outperformed by the trinomial tree (see **Tree Methods**).

First, the form of the tree given above is probably not the best if we want to regard the tree as an approximation to the Black–Scholes model. We see from equation (10) that the risk-neutral probabilities q_0, q_1 depend on r , so if we want to calibrate the model to the market yield curve we will need time-varying q_0, q_1 . This can be avoided if we write the Black–Scholes model as

$$S_t = F_{0,t} M_t \quad (16)$$

where $F_{0,t}$ is the forward price quoted at time 0 for exchange at time t and

$$M_t = \exp\left(\sigma W_t - \frac{1}{2}\sigma^2 t\right) \quad (17)$$

is the exponential martingale with Brownian motion W_t . See **Black–Scholes Formula** for this representation. $F_{0,t}$ only depends on the spot price S_0 and the yield curve (and the dividend yield, if any), so the only stochastic modeling required relates to the Brownian motion σW_t . Here we can use a standard “symmetric random walk” approximation: divide the time interval $[0, T]$ into n intervals of length $\delta = T/n$ and take a space step of length $h = \sigma\sqrt{\delta}$. At each discrete time point, the random walk (denoted X_i) takes a step of $\pm h$ with probability 1/2 each—this is just a binomial tree with equal up and down probabilities. For a single step $Z = X_i - X_{i-1} = \pm h$ we have $E[e^Z] = \cosh h$, so if we define $\alpha = \log(\cosh h)$ then $M_i^{(n)} = \exp(X_i - \alpha i)$ is a positive discrete-time martingale with $E[M_i^{(n)}] = 1$. It is a standard result that the sequence $M^{(n)}$ (suitably interpreted) converges weakly to M given by equation (17) as $n \rightarrow \infty$. This gives us a discrete-time model

$$S_i^{(n)} = F_{0,i\delta} M_i^{(n)} \quad (18)$$

such that $E[S_i^{(n)}] = F_{0,i\delta}$ holds *exactly* at each i . At node (i, j) in the tree the corresponding price is $F_{0,i\delta} \exp((n-j)h - i\alpha)$. Essentially, we have replaced the original multiplicative random walk representing the price $S(t)$ by an additive random walk representing the return process $\log S(t)$. The advantages of this are (i) all the yield curve aspects are bundled up in the model-free function F , and (ii) the stochastic model is “universal” (and very simple).

The decisive drawback of any binomial model is the absolute inflexibility with respect to volatility: it is impossible to maintain a recombining tree while allowing time-varying volatility. This means that the model cannot be calibrated to more than a single option price, making it useless for real pricing applications. The trinomial tree gets around this: we can adjust the local volatility by changing the transition probabilities while maintaining the tree geometry (i.e., the constant spatial step h).

References

- [1] Cox, J., Ross, S. & Rubinstein, M. (1979). Option pricing, a simplified approach, *Journal of Financial Economics* **7**, 229–263.
- [2] Elliott, R.J. & Kopp, P.E. (2005). *Mathematics of Financial Markets*, 2nd Edition, Springer.
- [3] McKean, H.P. (1965). Appendix to P.A. Samuelson, rational theory of warrant pricing, *Industrial Management Review* **6**, 13–31.
- [4] Shreve, S.E. (2005). *Stochastic Calculus for Finance, Vol 1: The Binomial Asset Pricing Model*, Springer.

Related Articles

Black–Scholes Formula; Quantization Methods; Tree Methods.

MARK H.A. DAVIS

American Options

An American option is a contract between the seller and the buyer. It is characterized by a nonnegative random function of time Z and a maturity. The option can be exercised at any time t between the initial date and the maturity. If the buyer exercises the option at time t , he/she receives the amount of money $Z(t)$ at time t . The buyer may exercise the option only once before the maturity. The price of an American option is always greater than or equal to the price of the corresponding European option (see **Black–Scholes Formula**). Indeed, the buyer of an American option gets more rights than the one who holds a European option, as he/she may exercise the option at any time before the maturity. This is the right of early exercise, and the difference between the American and European options prices is called the *early exercise premium*. The basic American options are American call and put options (see **Call Options**): they allow the buyer to sell or buy a financial asset at a price K (the strike price) and before a date (maturity) agreed before. The function Z associated to the call (respectively put) option is then $Z(t) = (S_t - K)^+$ (respectively $Z(t) = (K - S_t)^+$), where S_t is the value at time t of the underlying financial asset.

The study of American options began in 1965 with McKean [41] who considered the pricing problem as an optimal stopping problem and reduced it to a free boundary problem. The option value is then computable if one knows the free boundary called the *optimal exercise boundary*. In 1976, Van Moerbeke [48] exhibited some properties of this boundary. The formalization of the American option pricing problem as an optimal stopping problem was done in the two pioneering works of Benssoussan and Karatzas [5, 32].

They have proved that, under no arbitrage and completeness assumptions (see **Complete Markets**), the value process of an American option is the Snell envelope of the pay-off process, that is, the smallest supermartingale greater than the pay-off process. From previous works on these processes [23], we can derive some properties of the value process. Especially, we obtain characterization of optimal exercise times. In the section American Option and Snell Envelope, we present the main results and some

numerical methods based on this characterization of the option value process.

We can adopt a complementary point of view to study an American option. If we specify the evolution model for the underlying assets of the option, we could characterize the option value as the solution of a variational inequality. This method, introduced by Benssoussan and Lions [6], was applied to American options by Jaillet *et al.* [31]. We present this variational approach in the section Analytic Properties of American Options. We conclude this survey by giving results on exercise regions. In particular, we recall a formula linking the European and the American option prices known as the early exercise premium formula (see the section Exercise Region).

American Option and Snell Envelope

To price and hedge an American option, we have to choose a model for the financial market. We consider a filtered probability space $(\Omega, \mathbb{F} = (\mathcal{F}_t)_{0 \leq t \leq T}, \mathbb{P})$, where T is the maturity of our investment, $\mathcal{F}_t$ the information available at time t , and $\mathbb{P}$ the historical probability. We assume that the market is composed of $d + 1$ assets: $S^0, S^1, \dots, S^d$. S^0 is a deterministic process representing the time value of money. The others are risky assets such that S_t^i is the value of asset i at time t . In this section, we assume that the market does not offer arbitrage opportunities and is complete (see **Complete Markets**). Harisson and Pliska [28] observed that the no arbitrage assumption is equivalent to the existence of a probability measure equivalent to the historical one under which the discounted asset price processes are martingales. In a complete market, such a probability measure is unique and called the *risk-neutral probability measure* (see **Risk-neutral Pricing**). We will denote it by $\mathbb{P}^*$.

American Option Pricing

We present the problems linked to the American option study. The first one is the option pricing. An American option is characterized by an adapted and nonnegative process $(Z_t)_{t \geq 0}$, which represents the option pay-off if its owner exercises it at time t . We generally define Z as a function of one or several underlying assets. For instance, for a call option with strike price K , we have $Z_t = (S_t - K)^+$

or for a put option on the minimum of two assets, we have $Z_t = (K - \min(S_t^1, S_t^2))^+$. There also exist options, called *Amerasian options*, where the pay-off depends on the whole path of the assets, for instance, $Z_t = \left(K - \frac{1}{t} \int_0^t S_u du\right)^+$.

Using arbitrage arguments, Benssoussan and Karatzas [5, 32] have shown that the discounted American option value at time t is the Snell envelope of the discounted pay-off process [19, 43]. For the definition and general properties on the Snell envelope, we refer to [23] for continuous time and to [44] for discrete time. We can then assert that the price at time t of an American option with pay-off process Z and maturity T is

$$P_t = \text{esssup}_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}_{\mathbb{P}^*} \left[\frac{S_t^0}{S_\tau^0} Z_\tau \mid \mathcal{F}_t \right] \quad (1)$$

where $\mathcal{T}_{t,T}$ is the set of $\mathbb{F}$ -stopping times with values in $[t, T]$.

The second problem appearing in the option theory consists in determining a hedging strategy for the option seller (*see Hedging*). The solution follows directly from the Snell envelope properties. Indeed, if X is a process, we will denote the discounted process by $\tilde{X} = \frac{X}{S^0}$ and we have the following result ([35, Corollary 10.2.4]).

Proposition 1 *The process $(\tilde{P}_t)_{0 \leq t \leq T}$ is the smallest right-continuous super martingale that dominates $(\tilde{Z}_t)_{0 \leq t \leq T}$.*

As $(\tilde{P}_t)_{0 \leq t \leq T}$ is a super martingale, it admits a Doob decomposition (*see Doob–Meyer Decomposition*). There exist a unique right-continuous martingale $(M_t)_{0 \leq t \leq T}$ and a unique nondecreasing, continuous, adapted process $(A_t)_{0 \leq t \leq T}$ such that $A_0 = 0$ and $\tilde{P}_t = M_t - A_t$ for all $t \in [0, T]$. This decomposition of P is very useful to determine a surreplication strategy for an American option (*see Superhedging*). A strategy is defined as a predictable process $(\phi_t)_{0 \leq t \leq T}$ such that the value, at time t , of the portfolio associated with this strategy is $V_t(\phi) = \sum_{i=0}^d \phi_t^i S_t^i$. In a complete market, each contingent claim is replicable; then there exists a self-financing strategy ϕ such that $V_T(\phi) = S_T^0 M_T$. As $\tilde{V}(\phi)$ is a martingale under the risk-neutral probability, we get $\tilde{V}_t(\phi) = M_t$ for all $t \in [0, T]$. In conclusion, we have constructed a self-financing strategy such that

$$\forall t \in [0, T], \quad V_t(\phi) = P_t + S_t^0 A_t \geq P_t \quad (2)$$

This is a surreplication strategy for American options. Moreover, for this strategy, the initial wealth for hedging the option is minimum because we have $V_0(\phi) = P_0$.

The third problem arising in the American option theory is linked to early exercise opportunity. Contrary to European options, for the American option holder, knowing the arbitrage price of his/her option is not enough. He/she has to know when it is optimal for him/her to exercise the option. The tool to study this problem is the optimal stopping theory.

Optimal Exercise

We recall some useful results of the optimal stopping theory and apply them to the American put option in the famous Black–Scholes model. These results are proved in [23] in a larger setting and their financial applications have been developed in [35]. An optimal stopping time for an American option holder is a stopping time that maximizes his/her gain. Consequently, a stopping time ν is optimal if we have

$$\mathbb{E}_{\mathbb{P}^*}[\tilde{Z}_\nu] = \text{esssup}_{\tau \in \mathcal{T}_{0,T}} \mathbb{E}_{\mathbb{P}^*}[\tilde{Z}_\tau \mid \mathcal{F}_t] \quad (3)$$

We have a characterization of optimal stopping times, thanks to the following theorem.

Theorem 1 *Let $\tau^* \in \mathcal{T}_{0,T}$. τ^* is an optimal stopping time if and only if $P_{\tau^*} = Z_{\tau^*}$ and the process $(\tilde{P}_{t \wedge \tau^*})_{0 \leq t \leq T}$ is a martingale.*

It follows from this result that the stopping time

$$\tau^* = \inf\{t \geq 0 : P_t = Z_t\} \wedge T \quad (4)$$

is an optimal stopping time and, obviously, it is the smallest one. We can easily determine the largest optimal stopping time by using the Doob decomposition of super martingale. We introduce the following stopping time:

$$\nu^* = \inf\{t \geq 0 : A_t > 0\} \wedge T \quad (5)$$

and it is easy to see that ν^* is the largest optimal stopping time.

We then apply these results to an American put option in Black–Scholes framework (*see Black–Scholes Formula*). We assume that the

underlying asset S of the option is solution, under the risk-neutral probability, to the following equation:

$$dS_t = S_t (r dt + \sigma dW_t) \quad (6)$$

with $r, \sigma > 0$ and W a standard Brownian motion.

From the Markov property of S , we can deduce that the option price at time t is $P(t, S_t)$, where

$$P(t, x) = \sup_{\tau \in T_{0, T-t}} \mathbb{E}_{\mathbb{P}^*} [e^{-r\tau} (K - S_\tau)^+ | S_0 = x] \quad (7)$$

It is easy to see that $t \rightarrow P(t, x)$ is nonincreasing for all $x \in [0, +\infty)$. Moreover, for $t \in [0, T]$, the function $x \rightarrow P(t, x)$ is convex [24, 29, 30]. From the convexity of P , we deduce that there exists a unique optimal stopping time: $\tau^* = \inf\{t \geq 0 : P(t, S_t) = (K - S_t)^+\} \wedge T$. We introduce the so-called critical price or free boundary $s(t) = \inf\{x \in [0, +\infty) : P(t, x) > (K - x)^+\}$ and can write that

$$\begin{aligned} \tau^* &= \inf\{t \geq 0 : S_t \leq s(t)\} \wedge T \\ &= \inf\{t \geq 0 : W_t \leq \alpha(t)\} \wedge T \\ \text{with } \alpha(t) &= \frac{1}{\sigma} \left(\ln \left(\frac{s(t)}{S_0} \right) - \left(r - \frac{\sigma^2}{2} \right) t \right) \end{aligned} \quad (8)$$

Hence, τ^* is the reaching time of α by a Brownian motion. If α was known, we could compute τ^* and then P . However, the only way to get the law of τ^* explicitly is to reduce the dimension by considering options with infinite maturity (also known as *perpetual options*). In this case, we have the following result ([37, Proposition 4.5]).

Proposition 2 *The value function of an American perpetual put option is*

$$\begin{aligned} P^\infty(x) &= \sup_{\tau \in T_{0, +\infty}} \mathbb{E}_{\mathbb{P}^*} [e^{-r\tau} (K - S_\tau)^+ \\ &\quad \times \zeta_{\tau < +\infty} | S_0 = x] \end{aligned} \quad (9)$$

and is given by

$$\begin{aligned} P^\infty(x) &= \begin{cases} K - x & \text{if } x \leq s^* \\ (K - s^*) \left(\frac{s^*}{x} \right)^\gamma & \text{if } x > s^* \end{cases} \\ \text{where } \gamma &= \frac{2r}{\sigma^2} \text{ and } s^* = \frac{K\gamma}{1 + \gamma} \end{aligned} \quad (10)$$

is the critical price.

Another technique to reduce the dimension of the problem is the randomization of the maturity applied in [9, 13], but only approximations of the option price can be obtained in this way. In the following section, we present methods to approximate P based on the discretization of the problem.

Approximation of the American Option Value

To approximate P_t , it is natural to restrict the set of exercise dates to a finite one. We then introduce a subdivision $\mathcal{S} = \{t_1, \dots, t_n\}$ of the interval $[0, T]$ and assume that the option owner can exercise only at a date in $\mathcal{S}$. Such options are called *Bermuda options* and their price at time t is given by

$$P_t^n = \text{esssup}_{\tau \in T_{t, T}^n} \mathbb{E}[S_t^0 \tilde{Z}_\tau | \mathcal{F}_t] \quad (11)$$

where $T_{t, T}^n$ is the set of $\mathbb{F}$ -stopping times with values in $\mathcal{S} \cap [t, T]$. We obviously have $\lim_{n \rightarrow +\infty} P^n = P$ and some estimates of the error have been given in [1, 15]. For perpetual put options, Dupuis and Wang [21] have obtained a first-order expansion of the error on the value function and on the critical prices. In the case of finite maturity, this problem is still open; we just know that the error is proportional to $\frac{1}{n}$ for the value function and to $\frac{1}{\sqrt{n}}$ for the critical prices [18].

We have to determine $P_{t_i}^n$ for all $i \in \{1, \dots, n\}$. For this, we use the so-called dynamic programming equation:

$$\begin{cases} P_T^n = Z_T \\ P_{t_i}^n = \max \left(Z_{t_i}, \mathbb{E}_{\mathbb{P}^*} \left[\frac{S_{t_i}^0}{S_{t_{i+1}}^0} P_{t_{i+1}}^n | \mathcal{F}_{t_i} \right] \right) \end{cases} \quad (12)$$

This equation is easy to understand with financial arguments. At maturity of the option, it is obvious that the option price P_T^n is equal to the pay-off Z_T . At time $t_i < T$, the option holder has two choices: he/she exercises and then earns Z_{t_i} ; else he/she keeps the option and then would have the option value at time $n+1$, $P_{t_{i+1}}^n$. Hence, using the no arbitrage assumption, one can prove that at time t_i the option seller should receive

$$\max \left(Z_{t_i}, \mathbb{E}_{\mathbb{P}^*} \left[\frac{S_{t_i}^0}{S_{t_{i+1}}^0} P_{t_{i+1}}^n | \mathcal{F}_{t_i} \right] \right) \quad (13)$$

Computing the Bermuda option price consists now in calculating the expectations in the dynamic

programming equation. On the one hand, Monte Carlo techniques have been applied to solve this problem (see **Monte Carlo Simulation for Stochastic Differential Equations; Bermudan Options** and [11]). More precisely, we can quote some regression methods based on projections on Hilbert space base [40, 47], quantization algorithms proposed in [1, 2], and some Monte Carlo methods based on Malliavin calculus [3, 8]. On the other hand, we can use a discrete approximation of the underlying assets process. A widely used model is the Cox, Ross, and Rubinstein model (see **Binomial Tree**). We introduce a family of independent and identically distributed Bernoulli variables $(U_n)_{1 \leq n \leq N}$ with values in $\{b, h\}$, where $-1 < b < h$. We then consider only two assets S^0 and S whose respective initial values are 1 and S_0 such that

$$S_n^0 = (1+r)^n \quad \text{and} \quad S_n = S_{n-1}(1+U_n) \\ \forall n \in \{1, \dots, N\} \quad (14)$$

where $r > 0$ is the constant interest rate of the market. From the no arbitrage assumption, it follows that $b < r < h$ and that, under the risk-neutral probability $\mathbb{P}^*$, we have $p := \mathbb{P}^*(U_1 = h) = \frac{r-b}{h-b}$. Hence, using the Markov property of S , we can price an American option on S . For instance, for a call option with exercise price K , we get $P_n = F(n, S_n)$, where F is the solution to the following equation:

$$\left\{ \begin{array}{l} F(N, x) = (K - x)^+ \\ F(n, x) = \max \left\{ K - x, \frac{1}{1+r} \right. \\ \quad \times (pF(n+1, x(1+h)) \\ \quad \left. + (1-p)F(n+1, x(1+b))) \right\} \end{array} \right. \quad (15)$$

The convergence of binomial approximations was first studied in a general setting in [34]. The rate of convergence is difficult to get, but some estimates are given in [36, 38].

In conclusion, for some simple models, one can numerically solve the option pricing problem. However, only the time variable is discretized. Analytical methods have been developed and provide a better understanding of the links between time and space variables. In particular, we can characterize the option value as a solution to a variational inequality and get an approximation of its solution, thanks to finite difference methods. This characterization has many

consequences from the theoretical point of view and on practical aspects.

Analytic Properties of American Options

In this section, we assume that the assets prices process follows a model called *local volatility model* (see **Local Volatility Model**). This model is complete and takes into account the smile of volatility observed when one calibrates the Black–Scholes model (see **Model Calibration; Implied Volatility Surface** and [20]). We suppose that the assets prices process is solution to the following stochastic differential equation:

$$dS_t^i = S_t^i \left(b_i(t, S_t) dt + \sum_{j=1}^d \sigma_{i,j}(t, S_t) dW_t^j \right) \quad (16)$$

where W is a standard Brownian motion on $\mathbb{R}^d$, b a function mapping $[0, T] \times [0, +\infty)^d$ into $\mathbb{R}^d$, and σ a function mapping $[0, T] \times [0, +\infty)^d$ into $\mathbb{R}^{d \times d}$. Moreover, we assume that b is bounded and Lipschitz continuous, that σ is Lipschitz continuous in the space variable, and that there exists $\alpha \geq 1/2$ and σ_H such that $\forall x \in [0, +\infty)^d$, $(t, s) \in [0, T]^2$, $|\sigma(t, x) - \sigma(s, x)| \leq \sigma_H |t - s|^\alpha$. Moreover, to ensure the completeness of the market and the nondegeneracy of the partial differential equation satisfied by European option price functions, we assume that there exist $m > 0$ and $M > 0$ such that

$$\forall (t, x, \xi) \in [0, T] \times [0, +\infty)^d \times \mathbb{R}^d, \\ m^2 \|\xi\|^2 \leq \xi^* \sigma^* \sigma(t, x) \xi \leq M^2 \|\xi\|^2 \quad (17)$$

From the Markov property of the process S , at time t , the price of an American option with maturity T and pay-off process $(f(S_t))_{0 \leq t \leq T}$ is $P(t, S_t)$, where

$$P(t, x) = \sup_{\tau \in T_{t,T}} \mathbb{E} \left[e^{-r(\tau-t)} f(S_\tau) \mid S_t = x \right] \quad (18)$$

The Value Function

To compute the option price, we now have to study the option value function P . From its definition, we can derive immediate properties:

- $\forall x \in [0, +\infty)^d$, $P(T, x) = f(x)$

- $\forall(t, x) \in [0, T] \times [0, +\infty)^d$, $P(t, x) \geq f(x)$
- If the coefficients σ and b do not depend on time, we can write

$$P(t, x) = \sup_{\tau \in \mathcal{T}_0, T-t} \mathbb{E} [e^{-r\tau} f(S_\tau) \mid S_0 = x] \quad (19)$$

then the function $t \rightarrow P(t, x)$ is nonincreasing on $[0, T]$.

Up to imposing some assumptions on the regularity of the pay-off function, we can derive some important continuity properties of P . In this section, we assume that f is nonnegative and continuous on $[0, +\infty)$ such that

$$\begin{aligned} \exists(M, n) \in [0, +\infty) \times \mathbb{N}, \quad \forall x \in [0, +\infty)^d, \\ |f(x)| + \sum_{i=1}^d \left| \frac{\partial f}{\partial x_i}(x) \right| \leq M(1 + |x|^n) \end{aligned} \quad (20)$$

These assumptions are generally satisfied by the pay-off functions appearing in finance, especially by the pay-off functions of put and call options. In this setting, we have the following result [31].

Proposition 3 *There exists a constant $C > 0$ such that*

$$\begin{aligned} \forall t \in [0, T], \quad \forall(x, y) \in [0, +\infty)^{2d}, \\ |P(t, x) - P(t, y)| \leq C |x - y| \\ \forall x \in [0, +\infty)^d, \quad \forall(t, s) \in [0, T]^2, \\ |P(t, x) - P(s, y)| \leq C \left| (T - t)^{\frac{1}{2}} - (T - s)^{\frac{1}{2}} \right| \end{aligned} \quad (21)$$

(22)

As a consequence of this result, we can assert that the first-order derivatives of P in the sense of distributions are locally bounded on the open set $(0, T) \times (0, +\infty)^d$. This plays a crucial role in the characterization of P as a solution to a variational inequality.

Variational Inequality

In a more general setting, Benssoussan and Lions [6] have studied existence and uniqueness of solutions of variational inequalities and linked these solutions to those of optimal stopping problems. Applying

this method to the American option problem, Jaillet *et al.* have proved that the value function P can be characterized as the unique solution, in the sense of distribution, of the following variational inequality [31]:

$$\begin{cases} \mathcal{D}P \leq 0, & f \leq P, & (P - f)\mathcal{D}P = 0 & \text{a.e.} \\ P(x, T) = f(x) & \text{on } [0, +\infty) \end{cases} \quad (23)$$

where we set

$$\begin{aligned} \mathcal{D}h(t, x) = \frac{\partial h}{\partial t} + \frac{1}{2} \sum_{i,j=1}^d (\sigma \sigma^*)_{i,j}(t, x) x_i x_j \frac{\partial^2 h}{\partial x_i \partial x_j} \\ + \sum_{i=1}^d b_i(t, x) x_i \frac{\partial h}{\partial x_i} - rh \end{aligned} \quad (24)$$

This inequality directly derives from the properties of the Snell envelope. Indeed, the condition $\mathcal{D}P \leq 0$ is the analytic translation of the super martingale property of $\tilde{P}$, $f \leq P$ corresponds to $Z \leq P$, and the fact that one of this two inequalities has to be an equality follows from the martingale property of $(P_{t \wedge \tau^*})_{0 \leq t \leq T}$.

From the variational inequality, we can use numerical methods, such as finite difference methods, to compute the option price (see **Finite Difference Methods for Early Exercise Options** and [31]). From a theoretical point of view, we can deduce some analytic properties of P . If we add the condition that second-order derivatives of the pay-off function are bounded from below, we have the following result.

Proposition 4 *Regularity of P*

1. *Smooth fit property: For $t \in [0, T)$, the function $x \rightarrow P(t, x)$ is continuously differentiable and its first derivatives are uniformly bounded on $[0, T] \times [0, +\infty)^d$.*
2. *There exists a constant $C > 0$ such that for all $(t, x) \in [0, T) \times [0, +\infty)^d$, we have*

$$\left| \frac{\partial P}{\partial t}(t, x) \right| + |D^2 P(t, x)| \leq \frac{C}{(T - t)^{\frac{1}{2}}} \quad (25)$$

where $D^2 P$ is the Hessian matrix of P .

The *smooth fit property* has equally been established with probabilistic arguments, using the early

exercise premium formula presented in the section Exercise Region [30, 43]. In connection with free boundary problems, some analytic methods have been developed in [26] from which we can deduce the continuity of $\frac{\partial P}{\partial t}$ on $[0, T) \times [0, +\infty)^d$.

Thanks to the variational inequality, we can establish the so-called robustness of Black–Scholes formula [24]. The two main results obtained are the following.

Proposition 5 *We assume that $d = 1$. If the pay-off function is convex, then the value function P is equally convex. Moreover, if there exist $\sigma_1, \sigma_2 > 0$ such that $\sigma_1 \leq \sigma \leq \sigma_2$, then we have*

$$P^{\sigma_1} \leq P \leq P^{\sigma_2} \quad (26)$$

where P^{σ_i} is the value function of the American option on an underlying asset with volatility σ_i .

The propagation of convexity has been proved with probabilistic arguments in [29] and can be extended to the case $d > 1$. The robustness of Black–Scholes formula is equally useful from a practical point of view because it allows to construct surreplication and subreplication strategies using a constant volatility.

When there is only one risky asset modeled as a geometric Brownian motion, the analytic properties presented in this section can be used to transform, thanks to Green’s theorem, the variational inequality in an integral equation (see **Integral Equation Methods for Free Boundaries**). This point of view has been adopted to provide new numerical methods [16] to get theoretical results such as the convexity of the critical price for the put option [22] or its behavior near maturity [16, 25].

Integro-differential Equation

The integro-differential approach can be extended to the American option on jump diffusions (see **Partial Integro-differential Equations (PIDEs)**). In 1976, Merton (see **Merton, Robert C.** and [42]) introduced a model including some discontinuities in the assets value process. He considered a risky asset whose value process is solution to the following equation:

$$dS_t = S_{t-} \left(\mu dt + \sigma dW_t + d \left(\sum_{i=1}^{N_t} U_i \right) \right) \quad (27)$$

where $\mu \in \mathbb{R}$, $\sigma > 0$, W is a standard Brownian motion, N is a Poisson process with intensity $\lambda > 0$, and the U_i are independent and identically distributed variables with values in $(-1, +\infty)$ such that $\mathbb{E}[U_i^2] < +\infty$.

This model is not complete but up to a change of probability measure, we can suppose that $\mu = r - \lambda \mathbb{E}[U_1]$, where $r > 0$ is the constant interest rate of the market. Hence, $\tilde{S}$ is a martingale with respect to the filtration generated by W , N , and $(U_i \mathbb{1}_{\{i \leq N_t\}})_{0 \leq t \leq T}$. The option price is then determined as the initial wealth of a replication portfolio, which minimizes the quadratic risk. Merton obtained closed formulas to calculate the European options price. In this model, Zhang [50] extended the variational inequality approach to evaluate the American options price and he got a characterization of the value function as solution to the following integro-differential equation:

$$\begin{cases} \mathcal{D}P + \mathcal{I}P \leq 0, & f \leq P, \\ (\mathcal{D}P + \mathcal{I}P)(P - f) = 0 & \text{a.e.} \\ P(x, T) = f(x) & \text{on } [0, +\infty) \end{cases} \quad (28)$$

with

$$\begin{aligned} \mathcal{D}h(t, x) &= \frac{\partial h}{\partial t} + \frac{\sigma^2 x^2}{2} \frac{\partial^2 h}{\partial x^2} + \mu x \frac{\partial h}{\partial x} - rh \\ \mathcal{I}h(t, x) &= \lambda \int (h(t, x + z) - h(t, x)) \nu(dz) \end{aligned} \quad (29)$$

where ν is the law of $\ln(1 + U_1)$. Zhang used this equation to derive numerical schemes for approximating P . However, he could not obtain a description of the optimal exercise strategies. This was studied by Pham [46] who obtained a pricing decomposition formula and some properties of the exercise boundary.

In conclusion, analytic properties of the American option value function have been used to build numerical methods of pricing and to get some theoretical properties. Although the variational point of view is better for understanding the discretization of American options, it is less explicit than the probabilistic methods. We can remark that a specific region of $[0, T] \times [0, +\infty)^d$ appears in these two approaches: the so-called exercise region

$$\mathcal{E} = \{(t, x) \in [0, T) \times [0, +\infty)^d : P(t, x) = f(x)\} \quad (30)$$

If we knew $\mathcal{E}$, on the one hand, we would be able to determine the law of optimal stopping times and, on the other hand, the option pricing problem would be reduced to solving a partial differential equation in the complementary set of $\mathcal{E}$. In the following section, we recall some results on exercise regions and in particular we give a price decomposition, known as the *early exercise premium formula*, which involves the exercise region.

Exercise Region

Description

In the section Optimal Exercise, we have already presented a brief description of the exercise region of an American put option on a single underlying following the Black–Scholes model. These results are still true in the local volatility model introduced in the section Analytic Properties of American Options. Hence, for a put option with maturity T and strike price K , we have

$$\mathcal{E} = \{(t, x) \in [0, T) \times [0, +\infty); x \leq s(t)\} \text{ with} \\ s(t) = \inf\{x \in [0, +\infty) : P(t, x) = f(x)\} \quad (31)$$

Using the integral equation satisfied by P in the Black–Scholes model, we can apply general results proved in [27] for free boundary problems and assert that s is continuously differentiable on $[0, T)$. It has been shown that this is still true in the local volatility model using some blow-up techniques and monotonicity formulas [7, 12]. Moreover, Kim [33] proved that $\lim_{t \rightarrow T} s(t) = \min(K, \frac{rK}{\delta})$ if S is solution to $dS_t = S_t((r - \delta)dt + \sigma(t, S_t)dW_t)$. We will see that the behavior of s near maturity has been extensively studied.

The description of exercise region for options on several assets is more interesting because in high dimension numerical methods are less efficient and it helps to have a better understanding of these products. Broadie and Detemple were the first to investigate this problem [10]. They give precise descriptions of the exercise region shapes for the most traded options on several assets. Their results were completed by Villeneuve [49]. In particular, he gives a characterization of the nonemptiness of the exercise region. We just quote here the main results concerning a call option on the maximum on two

assets but the same kinds of results exist for many others options.

We denote by $\mathcal{E}_t$ the temporal section of the exercise region. For a call option on the maximum of two assets, S^1 and S^2 $\mathcal{E}_t$ can be decomposed in two regions: $\mathcal{E}_t^1 = \mathcal{E}_t \cap \{(x_1, x_2) \in [0, +\infty)^2 : x_2 \leq x_1\}$ and $\mathcal{E}_t^2 = \mathcal{E}_t \cap \{(x_1, x_2) \in [0, +\infty)^2 : x_1 \leq x_2\}$. These two regions are convex and can be rewritten as follows:

$$\mathcal{E}_t^1 = \{(x_1, x_2) \in [0, +\infty)^2 : s_1(t, x_2) \leq x_1\} \text{ and} \\ \mathcal{E}_t^2 = \{(x_1, x_2) \in [0, +\infty)^2 : s_2(t, x_1) \leq x_2\} \quad (32)$$

where s_1 and s_2 are the respective continuous boundaries of $\mathcal{E}_1(t)$ and $\mathcal{E}_2(t)$.

To compute these boundaries, we can use the early exercise premium formula given in the following section.

Early Exercise Premium Formula

About the same time, many authors have exhibited a decomposition formula for the American option price [14, 30, 43]. This formula is very enlightening from a financial point of view because it consists in writing that $P_t = P_t^e + a_t$ where P_t^e is the corresponding European option price and a is a nonnegative function of time corresponding to the premium the option buyer has to pay to get the right of early exercise. If the exercise region is known, a closed formula allows us to compute this premium. We recall this formula for a put option on a dividend-paying asset following the Black–Scholes model:

$$P(t, x) = P^e(t, x) + \int_t^T \mathbb{E}_{\mathbb{P}^*} [e^{-r(u-t)} (\delta S_u - rK) \\ \times \mathbb{1}_{\{S_u \geq s(u)\}} | S_t = x] du \quad (33)$$

where $\delta > 0$ is the dividend rate and $P^e(t, x) = \mathbb{E}_{\mathbb{P}^*} [e^{-r(T-t)} (K - S_T)^+ | S_t = x]$. This formula is equally interesting from a theoretical point of view as it leads to an integral equation for the critical price:

$$K - s(t) = P^e(t, s(t)) + \int_t^T \mathbb{E}_{\mathbb{P}^*} [e^{-ru} (\delta S_u - rK) \\ \times \mathbb{1}_{\{S_u \geq s(u)\}} | S_t = s(t)] du \quad (34)$$

This formula has been extended in [10] to American options on several assets. For the call on the maximum on two assets, we get

$$\begin{aligned}
 P(t, x) = & P^e(t, x) + \int_t^T \mathbb{E}_{\mathbb{P}^*} [e^{-r(u-t)} (\delta_1 S_u^1 - rK) \\
 & \times \zeta_{\{S_u^1 \geq s_1(u, S_u^2)\}} | S_t = x] du \\
 & + \int_t^T \mathbb{E}_{\mathbb{P}^*} [e^{-r(u-t)} (\delta_2 S_u^2 - rK) \\
 & \times \zeta_{\{S_u^2 \geq s_2(u, S_u^1)\}} | S_t = x] du \quad (35)
 \end{aligned}$$

Once again an integral equation could be derived for (s_1, s_2) .

We can also use this formula and the integral equation satisfied by the free boundary to study the behavior of the exercise region for short maturity. This is a crucial point for numerical methods. Indeed, we have seen that both the value function and the free boundary present irregularities near maturity, which implies instability in numerical methods.

Behavior Near Maturity

The behavior of the exercise region near maturity has been extensively studied when there is only one underlying asset. In his pioneering work, Van Moerbeke conjectured a parabolic behavior for the boundary near maturity [48]. However, when the asset does not distribute dividends, it has been shown that there is an extra logarithmic factor [4]. Lamberton and Vileneuve have then proved that, in the Black–Scholes model, the free boundary has a parabolic behavior if its limit is a point of regularity for the pay-off function, else a logarithmic factor appears [39]. This result has been extended to local volatility model in [17]. In a recent paper [16], new approximations are provided for the location of the free boundary by using integral equation satisfied by P and s . However, this technique cannot be extended to the case of options on several assets. When there are several underlying assets, the behavior of the exercise boundary near maturity has been studied by Nyström [45], who has proved that the convergence rate, when time to maturity goes to 0, is faster than parabolic.

References

- [1] Bally, V. & Pagès, G. (2003). Error analysis of the quantization algorithm for obstacle problems, *Stochastic Processes and their Applications* **106**, 1–40.
- [2] Bally, V., Pagès, G. & Printems, J. (2005). A quantization method for pricing and hedging multi-dimensional American style options, *Mathematical Finance* **15**, 119–168.
- [3] Bally, V., Caramellino, L. & Zanette, A. (2005). Pricing American options by Monte Carlo methods using a Malliavin Calculus approach, *Monte Carlo Methods and Applications* **11**, 97–133.
- [4] Barles, G., Burdau, J., Romano, M. & Sansoen, N. (1995). Critical stock price near expiration, *Mathematical finance* **5**, 77–95.
- [5] Bensoussan, A. (1984). On the theory of option pricing, *Acta Applicandae Mathematicae* **2**, 139–158.
- [6] Bensoussan, A. & Lions, J.L. (1982). *Applications of Variational Inequalities in Stochastic Control*, North-Holland.
- [7] Blanchet, A. (2006). On the regularity of the free boundary in the parabolic obstacle problem. Application to American options, *Nonlinear Analysis* **65**(7), 1362–1378.
- [8] Bouchard, B., Ekeland, I. & Touzi, N. (2004). On the Malliavin approach to Monte-Carlo approximation of conditional expectations, *Finance and Stochastics* **8**(1), 45–71.
- [9] Bouchard, B., El Karoui, N. & Touzi, N. (2005). Maturity randomisation for stochastic control problems, *Annals of Applied Probability* **15**(4), 2575–2605.
- [10] Broadie, M. & Detemple, J.B. (1997). The valuation of American options on multiple assets, *Mathematical Finance* **7**, 241–286.
- [11] Broadie, M. & Glasserman, P. (1997). Pricing American-style securities using simulation, *Journal of Economic Dynamics and Control* **21**, 1323–1352.
- [12] Caffarelli, L., Petrosyan, A. & Shahgholian, H. (2004). Regularity of a free boundary in parabolic potential theory, *Journal of the American Mathematical Society* **17**(4), 827–869.
- [13] Carr, P. (1998). Randomization and the American Put, *The Review of Financial Studies* **11**, 597–626.
- [14] Carr, P., Jarrow, R. & Myneni, R. (1992). Alternative characterization of American put options, *Mathematical Finance* **2**, 87–106.
- [15] Carverhill, A.P. & Webber, N. (1990). *American options: theory and numerical analysis*, in *Options: Recent Advances in Theory and Practice*, S. Hodges, ed, Manchester University Press.
- [16] Chadam, J. & Chen, X. (2007). Analytical and numerical approximations for the early exercise boundary for American put options, to appear in *Dynamics of Continuous, Discrete and Impulsive Systems* **10**, 649–657.
- [17] Chevalier, E. (2005). Critical price near maturity for an American option on a dividend-paying stock in

- a local volatility model, *Mathematical Finance* **15**, 439–463.
- [18] Chevalier, E. (2007). Bermudean approximation of the free boundary associated with an American option, *Free Boundary Problems: Theory and Applications* **154**, 137–147.
- [19] Duffie, D. (1992). *Dynamic Asset Pricing Theory*, Princeton University Press, Princeton.
- [20] Dupire, B. (1994). Pricing with a smile, *Risk Magazine* **7**, 18–20.
- [21] Dupuis, P. & Wang, H. (2004). On the convergence from discrete to continuous time in an optimal stopping problem, *Annals of Applied Probability* **15**, 1339–1366.
- [22] Ekström, E. (2004). Convexity of the optimal stopping boundary for the American put option, *Journal of Mathematical Analysis and Applications* **299**, 147–156.
- [23] El Karoui, N. (1981). Les aspects probabilistes du contrôle stochastique, *Lecture Notes in Mathematics* **876**, 72–238. Springer-Verlag.
- [24] El Karoui, N., Jeanblanc-Piqué, M. & Shreve, S. (1998). Robustness of the Black-Scholes formula, *Mathematical Finance* **8**, 93–126.
- [25] Evans, J.D., Keller, R.J. & Kuske, R. (2002). American options on assets with dividends near expiry, *Mathematical Finance* **12**(3), 219–237.
- [26] Friedman, A. (1975). *Stochastic Differential Equations and Applications*, Academic Press, New York, Vol. 1.
- [27] Friedman, A. (1976). *Stochastic Differential Equations and Applications*, Academic Press, New York, Vol. 2.
- [28] Harrison, J.M. & Pliska, S.R. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and their Applications* **11**, 215–260.
- [29] Hobson, D. (1998). Volatility misspecification, option pricing and superreplication via coupling, *The Annals of Applied Probability* **8**(1), 193–205.
- [30] Jacka, S.D. (1991). Optimal stopping and the American put, *Mathematical Finance* **1**, 1–14.
- [31] Jaillet, P., Lamberton, D. & Lapeyre, B. (1990). Variational inequalities and the pricing of American options, *Acta Applicandae Mathematicae* **21**, 263–289.
- [32] Karatzas, I. (1988). On the pricing of American options, *Applied Mathematics and Optimization* **17**, 37–60.
- [33] Kim, I.J. (1990). The analytic valuation of American options, *Review of Financial Studies* **3**, 547–572.
- [34] Kushner, H.J. (1977). *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York.
- [35] Lamberton, D. (1998). *American Options, Statistics in Finance*, D. Hand & S. Jacka, Arnold Applications of Statistics Series. eds, Edward Arnold London.
- [36] Lamberton, D. (1998). Error estimates for the binomial approximation of American put options, *Annals of Applied Probability* **8**, 206–233.
- [37] Lamberton, D. & Lapeyre, B. (1996). *Introduction to Stochastic Calculus Applied to Finance*, Chapman and Hall, London.
- [38] Lamberton, D. & Pagès, G. (1990). Sur l'approximation des réduites, *Annales de l'I.H.P., Probabilités et Statistiques* **26**(2), 331–355.
- [39] Lamberton, D. & Villeneuve, S. (2003). Critical price for an American option on a dividend-paying stock, *The Annals of Applied Probability* **13**, 800–815.
- [40] Longstaff, F.A. & Schwartz, E.S. (2001). Valuing American options by simulations: a simple least squares approach, *Review of Financial Studies* **14**, 113–147.
- [41] McKean, H.P. Jr. (1965). Appendix: a free boundary problem for the heat equation arising from a problem in mathematical economics, *Industrial Management Review* **6**, 32–39.
- [42] Merton, R.C. (1976). Option pricing when underlying stock returns are discontinuous, *Journal of Financial Economics* **3**, 125–144.
- [43] Myneni, R. (1992). The pricing of the American option, *Annals of Applied Probability* **2**, 1–23.
- [44] Neveu, J. (1975). *Discrete-Parameter Martingales*, North Holland, Amsterdam.
- [45] Nyström, K. (2007). On the behaviour near expiry for multi-dimensional American options, to appear in *Journal of Mathematical Analysis and Applications* **339**, 664–654.
- [46] Pham, H. (1997). Optimal stopping, free boundary and American option in a jump-diffusion model, *Applied Mathematics and Optimization* **35**, 145–164.
- [47] Tsitsiklis, J.N. & Van Roy, B. (2001). Regression methods for pricing complex American-Style options, *IEEE Transactions on Neural Networks* **12**(4), 694–703.
- [48] Van Moerbeke, P. (1976). On optimal stopping and free boundary problems, *Archive for Rational Mechanics and Analysis* **20**, 101–148.
- [49] Villeneuve, S. (1999). Exercice region of American options on several assets, *Finance and Stochastics* **3**, 295–322.
- [50] Zhang, X.L. (1997). Numerical analysis of American option pricing in a jump-diffusion model, *Mathematics of Operations Research* **22**, 668–690.

Further Reading

- Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- Dalang, R.C., Morton, A. & Willinger, W. (1990). Equivalent martingale measures and no-arbitrage in stochastic securities market models, *Stochastics and Stochastics Reports* **29**(2), 185–202.

- Detemple, J. (2005). *American-Style Derivatives: Valuation and Computation*, Financial Mathematics Series, Chapman & Hall/CRC, New York.
- Detemple, J., Feng, S. & Tian, W. (2003). The valuation of American call options on the minimum of two dividend-paying assets, *Annals of Applied Probability* **13**, 953–983.
- Friedman, A. (1964). *Partial Differential Equations of Parabolic Type*, Prentice-Hall: Englewood Cliffs, New Jersey.
- Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.

Related Articles

Bermudan Options; Bermudan Swaptions and Callable Libor Exotics; Early Exercise Options: Upper Bounds; Exercise Boundary Optimization Methods; Finite Difference Methods for Early Exercise Options; Integral Equation Methods for Free Boundaries; Point Processes; Swing Options.

ETIENNE CHEVALIER

Asian Options

An Asian option is also known as a *fixed-strike Asian option* or an *average price* or *average rate option*. These options have a payoff based on the average of an underlying asset price over a specified time period. The Asian option has a payoff dependent on the average of the asset price and a strike that is fixed in advance. The other type of Asian option is the average strike option (or floating strike), where the payoff is determined by the difference between the underlying asset price and its average (see **Average Strike Options**). Asian options are path-dependent options as their payoff depends on the asset price path rather than just on the terminal value.

If the average is computed using a finite sample of asset price observations taken at a set of regularly spaced time points, we have a discrete Asian option. A continuous time Asian option is obtained by computing the average *via* the integral of the price path over an interval of time. In reality, contracts are based on discrete averaging; however, if there are a large number of averaging dates, there are advantages in working in continuous time. The average itself can be defined to be geometric or arithmetic. When the geometric average is used, the Asian option has a closed-form solution for the price, whereas the option with arithmetic average does not have a known closed-form solution.

One of the reasons Asian options were invented was to avoid price manipulation toward the end of the option's life. By making the payoff depend on the average price rather than on the price itself, such manipulations have little effect on the option value. For this reason, Asian options are usually of European style. The possibility of exercise before the expiration date would make the option more vulnerable to price manipulation; see [11]. The payoff of an Asian option cannot be obtained by combining other instruments such as vanilla options, forwards, or futures.

Asian options are commonly used for currencies, interest rates, and commodities, and more recently in energy markets. They are useful in corporate hedging situations, for instance, a company exchanging foreign currency for domestic currency at regular intervals. Each transaction could be hedged separately with derivatives or a single Asian option could hedge the "average" rate over the period during which the currency is transferred.

An advantage for the buyer of an Asian option is that it is often less expensive than an equivalent vanilla option. This is because the volatility of the average is lower than the volatility of the asset itself. Another advantage in thinly traded markets is that the payoff does not depend only on the price of the asset on a particular day.

Consider the standard Black–Scholes economy with a risky asset (stock) and a money market account. We also assume the existence of a risk-neutral probability measure Q (equivalent to the real-world measure P) under which discounted asset prices are martingales. Under measure Q we denote the expectation by E , and under Q , the stock price follows:

$$\frac{dS_t}{S_t} = (r - \delta) dt + \sigma dW_t \quad (1)$$

where r is the constant continuously compounded interest rate, δ is a continuous dividend yield, σ is the instantaneous volatility of asset return, and W is a Q -Brownian motion. The reader is referred to **Black–Scholes Formula** for details on the Black–Scholes model and **Risk-neutral Pricing** for a discussion of risk-neutral pricing.

The Asian contract is written at time 0 and expires at $T > t_0$. The averaging begins at time $0 \leq t_0$ and occurs over the period $[t_0, T]$. (It is possible to have contracts where the averaging period finishes before maturity T , but this case is not covered here.) It is of interest to calculate the price of the option at the current time t , where $0 \leq t \leq T$. The position of t compared to the start of the averaging, t_0 , may vary. If $t \leq t_0$, the option is "forward starting". The special case $t = t_0$ is called a *starting option* here. If $t > t_0$, the option is termed *in progress* as the averaging has begun.

We consider an Asian contract that is based on the value A_t^l , where we denote $l = c$ for continuous averaging and $l = d$ for discrete averaging. The continuous arithmetic average is given as

$$A_t^c = \frac{1}{t - t_0} \int_{t_0}^t S_u du, \quad t > t_0 \quad (2)$$

and by continuity, we define $A_{t_0}^c = S_{t_0}$. For the discrete arithmetic average, denote $0 \leq t_0 < t_1 < \dots < t_n = T$, and for current time $t_m \leq t < t_{m+1}$ (for integer $0 \leq m \leq n$),

$$A_t^d = \frac{1}{m+1} \sum_{0 \leq i \leq m} S_{t_i} \quad (3)$$

The corresponding geometric average $G_t^l; l = c, d$ is defined to be

$$G_t^c = \exp\left(\frac{1}{t-t_0} \int_{t_0}^t \ln S_u du\right) \quad (4)$$

for continuous averaging and

$$G_t^d = (S_{t_0} S_{t_1} \dots S_{t_m})^{1/(m+1)} \quad (5)$$

for discrete averaging.

The payoff of an Asian call with arithmetic averaging is given as

$$(A_T^l - K)^+ \quad (6)$$

and the payoff of an Asian put with arithmetic averaging is given as

$$(K - A_T^l)^+ \quad (7)$$

where K is the fixed strike. Option payoffs depending on the geometric average are identical with A_T^l replaced by G_T^l .

By standard arbitrage arguments, the time- t price of the Asian call is

$$e^{-r(T-t)} \mathbb{E}[(A_T^l - K)^+ | \mathcal{F}_t] \quad (8)$$

and the price of the put is

$$e^{-r(T-t)} \mathbb{E}[(K - A_T^l)^+ | \mathcal{F}_t] \quad (9)$$

It is worth noting that in pricing the Asian option, we need to consider only those cases where $t \leq t_0$. For $t > t_0$, the option is in progress, and we can write $e^{-r(T-t)} \mathbb{E}[(A_T - K)^+ | \mathcal{F}_t]$ as

(documented in many papers including Levy [13]), so it is enough to consider the call option and derive the price of the put from this.

The main difficulty in pricing and hedging the Asian option is that the random variable A_T does not have a lognormal distribution. This makes the pricing very involved, and an explicit formula does not exist to date. This is an interesting mathematical problem and many research papers have been and still are written on the topic. The first of these was by Boyle and Emmanuel [3] in 1980.

Early methods for pricing the Asian option with arithmetic average involved replacing the arithmetic average A_T with the geometric average G_T , which is lognormally distributed; see [5, 10, 11, 15, 17]. This gives a simple formula, but it underprices the call significantly. However, it is worth noting that the formula leads to a scaling known as the $1/\sqrt{3}$ rule, since for $t > t_0$, the volatility is scaled down by this factor. That is, the formula involves the term $\sigma \frac{1}{\sqrt{3}} \sqrt{T-t}$. This is a particularly useful observation if the averaging period is quite short relative to the life of the option. See [12], among others, for a description and more details.

The second class of methods used is to approximate the true distribution of the arithmetic average using an approximate distribution, usually lognormal with appropriate parameters. True moments for A_T are equated with those implied by a lognormal model, so

$$\mathbb{E}A_T^n = e^{n\alpha + 1/2n^2v^2} \quad (11)$$

for any integer n and α, v are the mean and standard deviation of a normally distributed variable. This idea was used in a number of papers including

$$\begin{aligned} & e^{-r(T-t)} \mathbb{E}_t \left[\left(\frac{1}{T-t_0} \int_{t_0}^T S_u du + \frac{t-t_0}{T-t_0} A_t - K \right)^+ \right] \\ &= \frac{T-t}{T-t_0} e^{-r(T-t)} \mathbb{E}_t \left[\left(\frac{1}{T-t} \int_t^T S_u du - \left(\frac{T-t_0}{T-t} K - \frac{t-t_0}{T-t} A_t \right) \right)^+ \right] \end{aligned} \quad (10)$$

where $\mathbb{E}_t$ denotes the expectation conditional on information at time t . This is now the time t price of $\frac{T-t}{T-t_0}$ times an Asian option with averaging beginning at time t , with modified strike $\left(\frac{T-t_0}{T-t} K - \frac{t-t_0}{T-t} A_t \right)$. The prices of Asian options also satisfy a put-call parity

[13]. Turnbull and Wakeman [19] also corrected for skew and kurtosis by expanding about the lognormal. The practical advantage of such approximations is their ease of implementation; however, typically these

methods work well for some parameter values but not for others.

A further analytical technique in approximating the price of the Asian option is to establish price bounds. Curran [6] and Rogers and Shi [16] used conditioning to obtain a two-dimensional integral, which proves to be a tight lower bound for the option.

Much work has been done on pricing the Asian option using quasi-analytic methods. Geman and Yor [8] derived a closed-form solution for an “in-the-money” Asian call and a Laplace transform for “at-the-money” and “out-of-the-money” cases. Their methods are based on a relationship between geometric Brownian motion and time-changed Bessel processes. To price the option, one must invert the Laplace transform numerically; see [7]. Shaw [18] demonstrated that the inversion can be done quickly and efficiently for all reasonable parameter choices in Mathematica, making this a fast and effective approach. Linetsky [14] produced a quasi-analytic pricing formula using eigenfunction methods, with highly accurate results, also employing a package such as Mathematica.

Direct numerical methods such as Monte Carlo or quasi-Monte Carlo simulation and finite-difference partial differential equation (PDE) methods can be used to price the Asian option (see **Lattice Methods for Path-dependent Options**). In fact, given the popularity of such techniques, these methods were probably amongst the first used by practitioners (and remain popular today). Monte Carlo simulation was used to price Asian options by Broadie and Glasserman [4] and Kemna and Vorst [11], among many other more recent researchers. Simulation methods have the advantage of being widely used by practitioners to price derivatives, so no “new” method is required. Additional practical features such as stochastic volatility or interest rates can be incorporated without a significant increase in complexity. Control variates can often be used (e.g., using a geometric Asian option when pricing an arithmetic option). Additionally, simulation is often used as a benchmark price against which other methods are tested. The disadvantages are that it is computationally expensive, even when variance reduction techniques are used. Lapeyre and Temam [12] showed that Monte Carlo simulation can be competitive under the more advanced schemes they propose and with variance reduction techniques.

The Asian option is an exotic path-dependent option since the value at any point in time depends on the history of the underlying asset price. Specifically, the value of the option at t depends on the current level of the underlying asset S_t , time to expiry $T - t$, and the average level of the underlying up to t , A_t . Zvan *et al.* [21] presented numerical methods for solving this PDE. It turns out that the problem can be reduced to two variables (one state and the other time). Rogers and Shi [16], Alziary *et al.* [1], and Andreasen [2] formulated a one-dimensional PDE. The PDE approach is flexible in that it can handle market realities, but it is difficult to solve numerically as the diffusion term is very small for values of interest on the finite-difference grid. Vecer [20] reformulated the problem using analogies to passport options [9] to obtain an unconditionally stable PDE, which is more easily solved.

Methods based on discrete sampling become more appropriate when there are relatively few averaging dates. One simplistic approach is a scaling correction to volatility as described. Other possibilities include a Monte Carlo simulation or numerical solution of a sequence of PDEs [2]. Monte Carlo simulation can be quite efficient when there are only a small number of averaging dates, since the first “step” can take one straight to the averaging period (under the usual exponential Brownian motion model). Andreasen [2] priced discretely sampled Asian options using finite-difference schemes on a sequence of PDEs. This is particularly efficient if the averaging period is short and hence there are only a small number of PDEs to solve. He compared his PDE results to that of Monte Carlo simulation and showed that the finite-difference schemes get within a penny accuracy of the Monte Carlo simulation in less than a second of CPU time.

To conclude, there has been ongoing research into the methods for pricing the Asian option. It seems, however, the current-state-of-the-art pricing methods (good implementation of inversion of Laplace transform, eigenfunction and other expansions, stable PDE, and Monte Carlo simulation where appropriate) are fast, accurate, and adequate for most uses.

References

- [1] Alziary, B., Decamps, J.P. & Koehl, P.F. (1997). A PDE approach to Asian options: analytical and numerical evidence, *Journal of Banking and Finance* **21**(5), 613–640.

- [2] Andreassen, J. (1998). The pricing of discretely sampled Asian and lookback options: a change of numeraire approach, *Journal of Computational Finance* **2**(1), 5–30.
- [3] Boyle, P. & Emanuel, D. (1980). *The Pricing of Options on the Generalized Mean*. Working paper, University of British Columbia.
- [4] Broadie, M. & Glasserman, P. (1996). Estimating security price derivatives using simulation, *Management Science* **42**, 269–285.
- [5] Conze, A. & Viswanathan, R. (1991). European path dependent options: the case of geometric averages, *Finance* **12**(1), 7–22.
- [6] Curran, M. (1992). Beyond average intelligence, *Risk* **5**, 60.
- [7] Fu, M., Madan, D. & Wang, T. (1999). Pricing continuous Asian options: a comparison of Monte Carlo and Laplace transform inversion methods, *Journal of Computational Finance* **2**(2), 49–74.
- [8] Geman, H. & Yor, M. (1993). Bessel processes, Asian options and perpetuities, *Mathematical Finance* **3**, 349–375.
- [9] Henderson, V. & Hobson, D. (2000). Local time, coupling and the passport option, *Finance and Stochastics* **4**(1), 69–80.
- [10] Jarrow, R.A. & Rudd, A. (1983). *Option Pricing*. Irwin, IL.
- [11] Kemna, A.G.Z. & Vorst, A.C.F. (1990). A pricing method for options based on average asset values, *Journal of Banking and Finance* **14**, 113–129.
- [12] Lapeyre, B. & Temam, E. (2000). Competitive Monte Carlo methods for the pricing of Asian options, *Journal of Computational Finance* **5**, 39–57.
- [13] Levy, E. (1992). Pricing European average rate currency options, *Journal of International Money and Finance* **11**(5), 474–491.
- [14] Linetsky, V. (2004). Spectral expansions for Asian (average price) options, *Operations Research* **52**(6), 856–867.
- [15] Ritchken, P., Sankarasubramanian, L. & Vijn, A.M. (1993). The valuation of path-dependent contracts on the average, *Management Science* **39**(10), 1202–1213.
- [16] Rogers, L.C.G. & Shi, Z. (1995). The value of an Asian option, *Journal of Applied Probability* **32**, 1077–1088.
- [17] Ruttiens, A. (1990). Classical replica, *Risk* February, 33–36.
- [18] Shaw, W. (2000). *A Reply to Pricing Continuous Asian Options by Fu, Madan and Wang*, Working paper.
- [19] Turnbull, S.M. & Wakeman, L.M. (1991). A quick algorithm for pricing European average options, *Journal of Financial and Quantitative Analysis* **26**(3), 377–389.
- [20] Vecer, J. (2001). A new PDE approach for pricing arithmetic average Asian options, *Journal of Computational Finance* **4**(4), 105–113.
- [21] Zvan, R., Forsyth, P. & Vetzal, K. (1998). Robust numerical methods for PDE models of Asian options, *Journal of Computational Finance* **2**, 39–78.

Related Articles

Average Strike Options; Black–Scholes Formula; Lattice Methods for Path-dependent Options; Risk-neutral Pricing.

VICKY HENDERSON

Arbitrage Bounds

A key question in option pricing concerns how to incorporate information about the prices of existing, liquidly traded options into the prices of exotic options. In the classical **Black–Scholes model**, where there is only one parameter to choose, this question becomes: what do existing prices tell us about the volatility? Since the Black–Scholes model lacks the flexibility to capture all the market information, a wide variety of pricing models have been proposed. Rather than specifying a model and pricing with respect to this model, an alternative approach is to construct model-free *arbitrage* bounds on the price of exotic options. Arbitrage bounds are constraints on the price of an option, due to the absence of **arbitrage strategies**. These strategies are typically derived from relationships between the payoff of an option, and the payoff of a simple trading strategy constructed from other related derivatives—for example, the strategy might be a buy-and-hold strategy. If such a simple trading strategy can be shown to be worth at least as much as the corresponding option at maturity in every possible outcome, then the initial cost of the trading strategy must be more than the cost of the option, or else there exists a simple arbitrage. An important feature of these bounds is that they are often valid for a very wide class of models.

Arbitrage Bounds for Call Prices

Perhaps the earliest and simplest example of arbitrage bounds are the following inequalities, which are described in the seminal paper [29]:

$$\max\{0, S_0 - B(T)K\} \leq C(K, T) \leq S_0 \quad (1)$$

where $C(K, T)$ is the time-0 price of a *European call* option on the asset $(S_t)_{t \geq 0}$ with strike K and maturity T , and $B(T)$ is the time-0 price of a bond that is worth \$1 at time T . These bounds can be derived from the following simple arbitrages:

1. Suppose $C(K, T) > S_0$. Then we can construct an arbitrage by selling the call option and buying the asset. We receive an initial positive cash flow, while at maturity the option is worth $(S_T - K)_+$, which is less than S_T , the value of the asset we hold.
2. Suppose $C(K, T) < S_0 - B(T)K$. Then we can construct an arbitrage by buying the call option with strike K , selling short the asset, and buying K units of the bond that pays \$1 at time T . At time 0, we receive the cash amount

$$S_0 - B(T)K - C(K, T) \quad (2)$$

which, by assumption, is strictly positive. At maturity, writing $x_+ = \max\{x, 0\}$, we hold a portfolio whose value is

$$(S_T - K)_+ - (S_T - K) \quad (3)$$

which is positive.

3. Finally, it is clear that the call option must have a positive value (i.e., $C(T, K) \geq 0$), but this can also be considered a consequence of the arbitrage strategy of “buying” the derivative (for a negative price), and hence receiving positive cash flows both initially and at maturity.

There are some key features of the above example that are repeated in other similar applications. Note, first of all, that the inequalities make no modeling assumptions—the final value of the arbitrage portfolios will be larger/smaller than the call option for any final value of the asset, so these bounds are truly independent of any model for the underlying asset. Secondly, the bounds are the best we can do in the following sense: it can be shown that there are arbitrage-free models for the asset price under which the bounds are tight. For example, if the interest rates are deterministic, and the asset price satisfies $S_t = S_0 B(t)$, then the lower bounds hold for all strikes, and there is no arbitrage in the market. Alternatively, the upper and lower bounds can be shown to be the Black–Scholes price of an option in the limit as $\sigma \rightarrow \infty$ and $\sigma \rightarrow 0$, respectively.

In practice, these bounds are far too wide for most practical purposes, although they can be useful as a check that a pricing algorithm is producing sensible numerical results. Part of the reason for this wide range of values concerns the relatively small amount of information that is being used in deriving the bound. In general, one would expect to have some information about the behavior of the market. A natural place to look for further information is in the market prices of other vanilla options: in model-specific pricing, this information is commonly used for calibration of the model. However, the

information contained in these prices can also be used to provide arbitrage bounds on the prices of other exotic derivatives through the formulation of appropriate portfolios.

Breeden–Litzenberger Formula

One of the initial works to consider the pricing implications of vanilla options on exotic options is [6]. Here, the authors suppose that the value of calls at all strikes and a given maturity are known, and observe that

$$p(x) = \frac{1}{B(T)} \frac{\partial^2 C(K, T)}{\partial K^2} \Big|_{K=x} \quad (4)$$

can be thought of as the density of a random variable. The value at time-0 of an option whose payoff is only a function of the terminal value of the asset, $f(S_T)$, can then be shown to be

$$B(T) \int f(x) p(x) dx \quad (5)$$

or, intuitively, the discounted expectation under the density implied by the call prices. We can see this by noting that (at least for twice-differentiable functions f), we have

$$f(S) = f(0) + Sf'(0) + \int_0^\infty f''(K)(S - K)_+ dK \quad (6)$$

and therefore may replicate the contract $f(S)$ exactly by holding $f(0)$ in cash, buying $f'(0)$ units of the asset, and “holding” a continuous portfolio of calls consisting of $f''(K) dK$ units of call options with strikes in $[K, K + dK]$. Since this portfolio replicates the exotic option exactly, by an arbitrage argument, the prices must agree. The price of the portfolio of calls can be shown to be equation (5). In practice, some discrete approximation of such a portfolio is necessary, and this is generally possible provided the calls trade at a suitably large range of strikes.

One of the interesting consequences of this result is that we have a representation for the price of the exotic option as a discounted expectation. A key result in modern mathematical finance is the **fundamental theorem of asset pricing**, which allows us to deduce from the assumption of no arbitrage

that the price of an option may be written as a discounted expectation under a suitable probability measure. However, an assumption of the fundamental theorem of asset pricing is that there is a (known) model for the underlying asset. In the situation we wish to consider, there is no such measure. It is therefore not immediate that we can say anything about any probabilistic structure that might help us. One of the interesting consequences of this result is that it does provide some information about the underlying probabilistic structure: namely, that the call prices “imply” a risk-neutral distribution for the asset price, and that there are arbitrage relationships that ensure that any other option whose payoff depends only on the final value of the asset also has the price implied by this probability measure.

Arbitrage Bounds for Exotic Options

A general approach that is implied by the above examples is the following: suppose we know the prices of (and can trade in) a set of “vanilla” derivatives. Consider also an exotic option, for example, a **barrier option**. Without making any (strong) assumptions about a model for the underlying asset, what does arbitrage imply about the price of the barrier option? Through a suitable set of trades in the underlying and vanilla options, we should be able to construct portfolios and self-financing trading strategies that either dominate, or are dominated by, the payoff of the exotic option. If we can find a portfolio that dominates the exotic option, then the initial cost of this portfolio (which is known) must be at least as much as the price of the exotic option, or else there will be an arbitrage from buying the portfolio and selling the exotic option. The price of this portfolio therefore provides an upper bound on the price of the option. In a similar manner, we may also find a lower bound for the price of the option by looking for portfolios and trading strategies in the underlying and vanilla options that result in a terminal value that is always dominated by the exotic option. Note that we are, in general, interested in the least upper bound and also the greatest lower bound that can be attained, since these will give the tightest possible bounds.

We have been vague about two concepts here: first, we said that we would not want to make any “strong” assumptions about the model of the underlying asset.

The exact assumptions that different examples make about the underlying models vary from case to case, but typically we might assume, perhaps, that the underlying asset price is continuous (or at least, that it continuously crosses a barrier), or that the price process satisfies some symmetry assumption. Secondly, we have not specified what types of trading strategies we wish to consider: this is because, in part, this depends heavily on the assumptions on the price process—for example, trading strategies that involve a trade when the asset first crosses a barrier often assume that the underlying crosses the barrier continuously; the assumption on the symmetry of the asset price results in identities connecting the prices of call and put options. However, the important point to note here is that we work typically in a class of price processes that are too large to be able to hedge dynamically in any meaningful way, so that continuously rebalancing the portfolio is not an option. Two important classes of strategies are *static* strategies, which involve purchasing an initial portfolio of the underlying and vanilla options, and holding this to maturity (see **Static Hedging**), and *semistatic* strategies, which involve a fixed position in the options, and some trading in the underlying asset, often at hitting times of certain levels or sets.

Consistency of Vanilla Options

Since we are looking for arbitrage in the market when we add an exotic option, it is important that the initial prices of the vanilla options do include an arbitrage. In the case of equity markets, where the underlying vanilla options are call options, written on a given set of strikes and maturities, this is a question that has been studied by a number of authors [9, 11, 13, 15, 18]. The fundamental conclusion that may be arrived at from all these works is the following: the prices of calls are arbitrage free if and only if there exists a model under which the prices agree with the discounted expectation under the model. Moreover, the existence of the model has a relatively straightforward characterization in terms of the properties of the call prices, so that for a given set of call prices, the conditions may be checked with relative ease. Moreover, some practical concerns can be included in the models: [15] allows the inclusion of default of the asset, while [18] also allows for the inclusion of dividends.

Of course, not all markets fit naturally into this framework, and so other settings should also be considered, as, for example, in [27], where arbitrage bounds for fixed income markets are considered.

Barrier Options

One of the simplest classes of options that can be considered are the various types of barrier options, and one of the simplest of these options is the *one-touch* barrier option: this is an option that pays \$1 at maturity if the barrier is breached during the lifetime of the contract, and expires worthless if the barrier is not hit before maturity. Suppose that the price process is continuous, and suppose further that the riskless interest rate is zero. Then [7] provides an upper bound on the price of the option, $OT(R, T)$, where R is the level of the barrier, $R > S_0$, and T is the maturity of the option. The bound that is derived in [7] is

$$OT(R, T) \leq \inf_{x \leq R} \frac{C(x, T)}{R - x} \quad (7)$$

The bound can be most clearly seen by noting the corresponding arbitrage strategy: suppose that the bound does not hold, then we can find an x for which

$$OT(R, T) > \frac{C(x, T)}{R - x} \quad (8)$$

We sell the one-touch option, and buy $\frac{1}{R-x}$ units of the call with strike x and maturity T . If the barrier at R is not hit, the one-touch option expires worthless, and our call option may have positive value. Alternatively, suppose that at some time, the barrier is hit. At this time, we enter into a forward contract on the asset. Specifically, we sell $\frac{1}{R-x}$ units of a forward struck at R . Since the current value of the asset is R , and we have assumed that the interest rates are zero, we may enter into such a contract for free. At maturity, the value of our position in the forward will be $\frac{R-S_T}{R-x}$, and the total value of our position in the call and the forward is

$$\begin{aligned} & \frac{1}{R-x}(S_T - x)_+ + \frac{R - S_T}{R-x} = \frac{1}{R-x} \\ & \times [(S_T - x) + (x - S_T)_+ + (R - S_T)] \\ & = -1 + \frac{(x - S_T)_+}{R-x} \end{aligned} \quad (9)$$

where we write $x_+ = \max\{x, 0\}$. Since the value of the portfolio is now greater than the value of the one-touch option, we have an arbitrage.

It can also be shown that the bound here is the best that can be attained: specifically, it can be shown that there exists a model under which there is equality in the identity (7). By considering the form of the hedge, we can also say something about the extremal model. For equality to be there in equation (7), we must always have equality between the payoff of the one-touch option, and the value of the hedging portfolio. The case where the barrier is not hit requires that

$$0 = \frac{(S_T - x)_+}{R - x} \quad (10)$$

or, equivalently, that S_T is always below x . The case where the barrier is struck requires that

$$1 = 1 + \frac{(x - S_T)_+}{R - x} \quad (11)$$

or that S_T is always above x . In other words, in the extremal model, the paths that hit the barrier will, at maturity, finish above the minimizing value of x , while those that do not hit the barrier will always end up below x .

A similar approach allows us to find a lower bound. In this case, the hedging portfolio consists of a digital call struck at the barrier, so that the payoff of this option is simply \$1 if the asset ends up above the barrier, and put options are struck at the barrier, at some $y < R$. Note that the digital call can, in theory at least, be arbitrarily closely approximated by buying a suitably large number of calls just below the strike, and selling the same number of calls at the strike, so that we can deduce the price of the digital call from the prices of the vanilla call options. The prices of the puts can be deduced from **put-call parity**. In a manner similar to the above, we can find the “best” bound by finding the value of y that corresponds to the most expensive portfolio. Again, the bound is tight, in the sense that there exists a model under which we attain equality. We can also describe the behavior in this model: the paths that hit the barrier will end up either below y or above R . Those that do not hit the barrier will finish between y and R .

Using extensions of these ideas, similar bounds can be found for other common barrier options, for example, down-and-in calls. Full details can be found in [7].

There are a number of observations that we can make about the solution to the above problem, and which extend more generally. First, the extension to nonzero interest rates is nontrivial—one of the assumptions that was made in constructing the trading strategy was that, when the barrier is struck, we would be able to enter into a forward contract with a strike at the barrier. If there are nonzero interest rates, we will not be able to enter into such a contract at no cost. Consequently, these results are only generally valid in cases where there is zero cost of carry, for example, where the underlying is a forward price, in foreign exchange markets where both currencies have the same interest rate, or commodities where the interest rate is the same as the convenience yield. Secondly, recall that the only assumption we made on the paths was continuity. This assumption is key to knowing that we can sell forward as we hit the barrier. In fact, the upper bound will still hold if the path is not continuous, provided we sell forward the first time that we go above the barrier, at which point, we can enter into a forward contract that is at least as good for our purposes. Note, however, that under the model for which the bound is tight, we must cross the barrier continuously. The same is not true of the lower bound, which fails if the asset price does not cross the barrier continuously. If the path is not assumed continuous, a new bound can be derived, which corresponds to the asset jumping immediately to its final value. The third aspect to note about these constructions is that there is a natural extension to the case where calls are available at finitely many strikes. Consider the upper bound on the one-touch option, and suppose that calls trade at a finite set of strikes $K_1, K_2, \dots, K_n$. Rather than taking the infimum over x where $x < R$, to get an upper bound, we can take the minimum over the strikes at which calls are available

$$OT(R, T) \leq \min_{i: K_i < R} \frac{C(K_i, T)}{R - K_i} \quad (12)$$

The previous arguments can be applied directly to show that this is an upper bound. It can also be shown that there is a model that fits with the call prices, and under which this bound is attained, so the resulting bounds are the best possible. Details of this extension can be found in [7].

Put–Call Symmetry

An alternative approach to the pricing of barrier options using the above techniques is to introduce the concept of “put–call symmetry”. Following [5], we say that put–call symmetry holds if the value of a call struck at $K > S_t$, and a put struck at $H < S_t$, satisfy

$$C(K)K^{-1/2} = P(H)H^{-1/2} \quad (13)$$

where the current asset price S_0 is the geometric mean of H and K : $(KH)^{1/2} = S_0$. While this is a more general concept, in the context of a **local volatility model**, this assumption can be interpreted in terms of a symmetry condition on the volatility: $\sigma(S_t, t) = \sigma(S_0^2/S_t, t)$. In particular, this is an assumption that is satisfied whenever the volatility is a deterministic function of time. Alternatively, if we graph the **implied volatility** smile against $\log(K/S_t)$, the smile should be symmetric. Note that, as above, we still require either the interest rate to be zero, or, for example, to be working with a forward price.

Under the assumption that put–call symmetry holds at all future times, we can construct replicating portfolios for many types of barrier options. Consider the case of a down-and-in call (see **Barrier Options**), with a barrier at R and strike K , so $R < S_0$. Then we may hedge the option simply by purchasing initially K/R puts at H , where $H = R^2/K$. If the asset never reaches the barrier, both the down-and-in call and the put expire worthless, so we consider the behavior at the barrier. When the asset is at the barrier, put–call symmetry implies

$$C(K) = \frac{K}{R} P(H) \quad (14)$$

and so we may sell the puts and buy a call with strike K . Thus this portfolio exactly replicates the down-and-in call.

The results described above were initially introduced in [5], where, in addition to considering knock-in and knockout calls, and the one-touch option above, the authors also included the **lookback** option by expressing it as a portfolio of suitable down-and-in options. Further developments can be found in [10], which considers the replication of more general options in this framework, and [12], which extends to double knockout calls, rolldown calls, and ratchet calls. Further extensions to these ideas, where the volatility is assumed to be a known function of the

underlying and time, can be found in [2]. A different approach to static hedging is given in [20].

Arbitrage Bounds via Skorokhod Embeddings

As shown by Dupire [21], if prices of calls at all strikes *and* all maturities are known, there is a unique diffusion model, the **local volatility model**, which matches those call prices. If we drop the diffusion assumption, we are led to follow the line of reasoning from [6]. One of the conclusions from this work is that knowing the call prices at all strikes at a fixed future maturity implies the law of the asset price under the risk-neutral measure at this fixed future date. Further, as a consequence of the assumption of no arbitrage, we believe that under the risk-neutral measure, the discounted asset price should be a martingale. In this manner, we should be able to restrict the class of possible (discounted, risk-neutral) price processes to the class of martingales that have a given terminal distribution. If we now also wish to infer information about the price of an exotic option, we can ask the question: what is the largest/smallest price implied by the martingale price processes in this class? Moreover, we might hope to find an arbitrage if the option trades outside this range.

One of the simplest examples to consider is the one-touch option above: under the risk-neutral measure, the price of the call is the discounted probability that the price process goes above the barrier before the expiry date. By restricting ourselves to the class of martingales with a given terminal law, we should be able to deduce some information about the possible values of this probability, and thus of the price of the option. The key to using this approach efficiently is to find a suitable representation of the set of martingales with the given terminal distribution.

A classical result from probability theory, the Dambis–Dubins–Schwartz Theorem, states that any continuous martingale may be written as the time change of a Brownian motion (see, for example, [33, Chapter V]), and this is essentially true if the martingale is only right continuous [30]. Hence, if the discounted asset price is a martingale, one would expect it to be a time change of a Brownian motion—that is, we would expect to be able to write

$$B(t)S_t = W_{\tau(t)} \quad (15)$$

where W_t is a Brownian motion, $\tau(t)$ is increasing in t and is a stopping time for all t . As a consequence, any martingale price process should be a time change of a Brownian motion. If, in addition, we know that the law of S_T under the risk-neutral measure is implied by the call prices, we also know that $W_{\tau(T)}$ has a given law. Finally, suppose that the time change is continuous (as it will be if the price process is continuous), then many of the properties in which we are interested remain unaffected by the exact form of the time change. For example, consider the probability of whether the discounted asset price goes above a barrier R before time T . This is the same as the probability that the Brownian motion W_t with $W_{\tau(t)} = B(t)S_t$ goes above the barrier before time $\tau(T)$. Moreover, consider two time changes $\tau(t)$ and $\tilde{\tau}(t)$, such that we always have $\tau(T) = \tilde{\tau}(T)$. Then the probability of whether the barrier has been breached will be the same for the price processes corresponding to the time change τ and the time change $\tilde{\tau}$. Consequently, if we are concerned with such path properties of the underlying price process, when we look in the Brownian setting, we need only differentiate between different final stopping times $\tau(T)$, and not different time changes.

The argument then goes as follows: suppose we know call prices at all strikes at time T . From this information, we may deduce the law of the discounted asset price $B(T)S_T$, which we assume to be a time change of a Brownian motion, and whose value at some stopping time τ therefore has the same law. Since the time change in the intermediate time is assumed to be continuous, and its exact form will not impact the quantities of interest, we get a one-to-one correspondence between possible price processes and the class of stopping times of a Brownian motion that have a given law. This line of reasoning is of interest, since the problem of finding a stopping time with a given terminal law has a long history in the probabilistic literature, where it is known as the **Skorokhod embedding problem**. In particular, given a distribution μ , we say that a stopping time τ is a (Skorokhod) embedding of μ , if W_τ has law μ . The recent survey paper [31] contains a comprehensive survey of the probabilistic literature on the Skorokhod embedding problem.

Getting back to the one-touch option, we see that the upper bound will correspond to the stopping time that maximizes the probability of being larger than the barrier within the class of embeddings,

and the minimum will correspond to stopping time that minimizes the probability within this class. The construction of arbitrage bounds for the price of the option is therefore equivalent to the identification of extremal Skorokhod embeddings for the law implied by the call prices at maturity, as seen in [7]. The construction that attains this maximum is due to Azema and Yor [3], while the construction that attains the minimum is due to Perkins [32], and it can be shown that these embeddings do indeed have the behavior that was hypothesized previously: for the upper bound, those paths that hit the barrier remain above the level x derived in the bound, while in the lower bound, those paths that hit the barrier all either finish above the barrier, or stop below y .

The Skorokhod embedding approach was initially explored in [23]. In this work, it is shown that the upper bound on the price of a lookback option can be computed in terms of the available call prices. Moreover, Hobson [23] has constructed a trading strategy that will result in an arbitrage should the lookback option trade above the given bound. In this case, the strategy involves constructing an initial portfolio of calls (purchased at the specified prices) and then selling these calls appropriately as the price process sets new maxima. The price at which the calls can be sold will be at least the intrinsic value of the call, and it can be shown that the profit from selling off the calls appropriately will be at least the payoff from the lookback option. A simple lower bound is also derived, but without assuming any continuity. For discontinuous asset prices, the lower bound is attained by the price process that jumps immediately to its final value. In terms of the corresponding Skorokhod embeddings, the upper bound has close connections with the embedding due to Azema and Yor [3]; this can be shown to maximize the law of the maximum over the class of embeddings. Further, it can be shown that if we use the price process that corresponds to the stopping time constructed in [3], then the trading strategy dominating the lookback option actually attains equality demonstrating that the upper bound is the best possible. This connection between an extremal Skorokhod embedding and a corresponding bound on the price of a connected exotic option has been exploited a number of times: in [8], these techniques are used to generalize the above results to the case where the call prices at an intermediate time are also known; in [24] the embedding due to Perkins [32] is generalized to

provide a lower bound on the price of a forward start digital option, under the assumption that the price process is continuous; in [16] the embedding of Vallois [35] is used to provide an upper bound on products related to corridor variance options.

A related development of these ideas is considered in [28], wherein the problem of fitting martingales to marginal distributions specified at all maturities is presented, and some solutions corresponding to the different Skorokhod embedding approaches, the local volatility models of Dupire [21], and processes with independent increments are discussed.

Advantages and Disadvantages

From a theoretical point of view, the results described above provide a clear, satisfactory picture: for a relatively large class of options, a range of model-free prices, or even exact prices, can be established. Where there is a range of prices, the upper and lower bounds can usually be shown to be tight, and trading strategies produced that result in arbitrages, should the bounds be violated.

However, the results have often been produced under strong restrictions on the mechanics of the market—typically, the cost of carry has been assumed to be zero, and factors such as transaction costs have been ignored. To some extent, these factors can be added into the bounds, although this is at the expense of wider bounds. Moreover, the bounds that result from the model-free techniques have a tendency to be rather wide. Figure 1 illustrates the resulting bounds for the one-touch option, comparing the upper and lower bounds described earlier, with the actual price derived from a Black–Scholes model. The range of the bounds is, for interesting values, of the order of 5% of the final payoff above the Black–Scholes price, and as much as 15% below the Black–Scholes price. These ranges are of much too high an order to be helpful for pricing purposes.

How else might these techniques be of use in practice? One important feature is the tendency to produce simple hedging portfolios. These allow a trader to cover a position in a derivative with a portfolio that needs little or no ongoing management, and through which they have a guaranteed lower bound on any possible hedging error. Several authors, for example, [22, 34], have produced comparisons between static or semistatic and dynamic

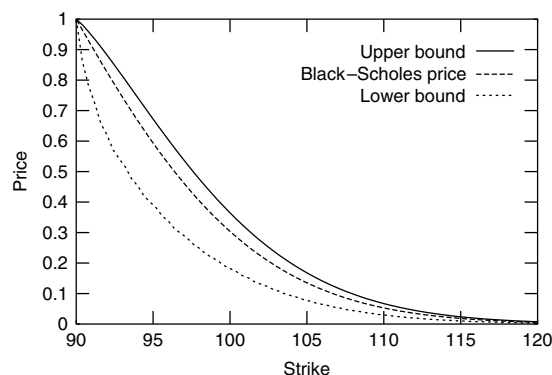


Figure 1 Upper and lower model-free bounds on the price of a one-touch option, as a function of the strike, compared with the Black–Scholes price. The interest rate is 0, the asset price is \$90, and $\sigma = 15\%$

hedging. In [34], there is no clear outperformance by either strategy, but in some circumstances the static or semistatic hedging strategy outperforms the dynamic strategy. In [22], the authors consider barrier options, and find that some static hedging strategies for barrier options appear to outperform dynamic strategies. Another useful observation is that by identifying the extremal models, one can identify the key model properties that influence the price of the option: for example, in finding bounds for the one-touch barrier, the extremal models were identified as those models that either hit the barrier and stay close, or those models that hit the barrier and end up far away. Knowledge of these extremes might help in deciding where the real price might lie in relation to the arbitrage bounds, or how prices of the option might react to large structural changes to the market. Finally, arbitrage bounds can also be considered as a special case of the **good-deal bounds** of [14]. Good-deal bounds provide a range of prices, outside of which there exists a trading strategy whose payoff may be considered a “good deal”, which is not necessarily an arbitrage, but is sufficiently close to one to be very desirable for an investor.

Additional Resources

There are a number of papers [17, 25, 26] that consider deriving bounds on the price of **basket options**, where the payoff of the option depends

on the value of a weighted sum of a number of assets, and where calls are traded on each of the underlying assets. There are also connections to [1], where bounds on the prices of **Asian options** are derived.

Another class of options where similar hedging techniques have been considered are installment options [19], which are options similar to a European call, but where the holder pays for the option in a set number of installments, and has the option to stop paying the installments at any point before maturity, thereby losing the final payoff for the contract.

A common complication that arises in constructing many of the bounds and their respective hedging portfolios is that there can be some nontrivial optimization problems, typically, large linear programming problems [4, 17, 25].

References

- [1] Albrecher, H., Mayer, P.A. & Schoutens, W. (2008). General lower bounds for arithmetic Asian option prices, *Applied Mathematical Finance* **15**(2), 123–149.
- [2] Andersen, L.B.G., Andreasen, J. & Eliezer, D. (2002). Static replication of barrier options: some general results, *Journal of Computational Finance* **5**(4), 1–25.
- [3] Azéma, J. & Yor, M. (1979). Une solution simple au problème de Skorokhod, in *Séminaire de Probabilités, XIII (Univ. Strasbourg, Strasbourg, 1977/78), Lecture Notes in Mathematics*, Springer, Berlin, Vol. 721, pp. 90–115.
- [4] Bertsimas, D. & Popescu, I. (2002). On the relation between option and stock prices: a convex optimization approach, *Operations Research* **50**(2), 358–374.
- [5] Bowie, J. & Carr, P. (1994). Static simplicity, *Risk* **7**(8), 45–49.
- [6] Breeden, D.T. & Litzenberger, R.H. (1978). Prices of state-contingent claims implicit in option prices, *Journal of Business* **51**(4), 621–651.
- [7] Brown, H., Hobson, D. & Rogers, L.C.G. (2001a). Robust hedging of barrier options, *Mathematical Finance* **11**(3), 285–314.
- [8] Brown, H., Hobson, D. & Rogers, L.C.G. (2001b). The maximum maximum of a martingale constrained by an intermediate law, *Probability Theory and Related Fields* **119**(4), 558–578.
- [9] Buehler, H. (2006). Expensive martingales, *Quantitative Finance* **6**(3), 207–218.
- [10] Carr, P. & Chou, A. (1997). Breaking barriers, *Risk* **10**(9), 139–145.
- [11] Carr, P. & Madan, D.B. (2005). A note on sufficient conditions for no arbitrage, *Finance Research Letters* **2**, 125–130.
- [12] Carr, P., Ellis, K. & Gupta, V. (1998). Static hedging of exotic options, *Journal of Finance* **53**(3), 1165–1190.
- [13] Carr, P., Geman, H., Madan, D.B. & Yor, M. (2003). Stochastic volatility for Lévy processes, *Mathematical Finance* **13**(3), 345–382.
- [14] Cerny, A. & Hodges, S.D. (1999). *The theory of good-deal pricing in financial markets*, FORC preprint, No. 98/90.
- [15] Cousot, L. (2007). Conditions on option prices for absence of arbitrage and exact calibration, *Journal of Banking and Finance* **31**, 3377–3397.
- [16] Cox, A.M.G., Hobson, D.G. & Obloj, J. (2008). Pathwise inequalities for local time: applications to Skorokhod embeddings and optimal stopping, *Annals of Applied Probability* **18**(5), 1870–1896.
- [17] d’Aspremont, A. & El Ghaoui, L. (2006). Static arbitrage bounds on basket option prices, *Mathematical Programming* **106**(3), Series A, 467–489.
- [18] Davis, M.H.A. & Hobson, D.G. (2007). The range of traded option prices, *Mathematical Finance* **17**(1), 1–14.
- [19] Davis, M.H.A., Schachermayer, W. & Tompkins, R.G., (2001). Installment options and static hedging, in *Mathematical Finance* (Konstanz, 2000), Trends in Mathematics, Birkhäuser, Basel, pp. 131–139.
- [20] Derman, E., Ergener, D. & Kani, I. (1995). Static options replication, *Journal of Derivatives* **2**, 78–95.
- [21] Dupire, B. (1994). Pricing with a smile, *Risk* **7**, 32–39.
- [22] Engelmann, B., Fengler, M.R., Nalholm, M. & Schwender, P. (2006). Static versus dynamic hedges: an empirical comparison for barrier options, *Review of Derivatives Research* **9**(3), 239–264.
- [23] Hobson, D.G. (1998). Robust hedging of the lookback option, *Finance and Stochastics* **2**(4), 329–347.
- [24] Hobson, D.G. & Pedersen, J.L. (2002). The minimum maximum of a continuous martingale with given initial and terminal laws, *Annals of Probability* **30**(2), 978–999.
- [25] Hobson, D., Laurence, P. & Wang, T. (2005a). Static-arbitrage upper bounds for the prices of basket options, *Quantitative Finance* **5**(4), 329–342.
- [26] Hobson, D., Laurence, P. & Wang, T. (2005b). Static-arbitrage optimal subreplicating strategies for basket options, *Insurance, Mathematics & Economics* **37**(3), 553–572.
- [27] Jaschke, S.R. (1997). Arbitrage bounds for the term structure of interest rates, *Finance and Stochastics* **2**(1), 29–40.
- [28] Madan, D.B. & Yor, M. (2002). Making Markov martingales meet marginals: with explicit constructions, *Bernoulli* **8**(4), 509–536.
- [29] Merton, R.C. (1973). Theory of rational option pricing, *The Bell Journal of Economics and Management Science* **4**(1), 141–183.
- [30] Monroe, I. (1972). On embedding right continuous martingales in Brownian motion, *Annals of Mathematical Statistics* **43**, 1293–1311.

-
- [31] Oblój, J. (2004). The Skorokhod embedding problem and its offspring, *Probability Surveys* **1**, 321–390, electronic.
 - [32] Perkins, E. (1986). The Cereteli-Davis solution to the H^1 -embedding problem and an optimal embedding in Brownian motion, in *Seminar on Stochastic Processes, 1985* (Gainesville, Fla., 1985), Progress in Probability and Statistics, Birkhäuser Boston, Boston, Vol. 12, pp. 172–223.
 - [33] Revuz, D. & Yor, M. (1999). *Continuous Martingales and Brownian Motion*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 3rd Edition, Springer-Verlag, Berlin, Vol. 293.
 - [34] Tompkins, R. (1997). Static versus dynamic hedging of exotic options: an evaluation of hedge performance via simulation, *Netexposure* **1**, 1–28.
 - [35] Vallois, P. (1983). Le problème de Skorokhod sur $\mathbf{R}$: une approche avec le temps local, in *Seminar on Probability, XVII*, Lecture Notes in Mathematics, Springer, Berlin, Vol. 986, pp. 227–239.

Related Articles

Arbitrage Strategy; Barrier Options; Dupire Equation; Good-deal Bounds; Hedging; Model Calibration; Skorokhod Embedding; Static Hedging.

ALEXANDER COX

Average Strike Options

An average strike option is also known as an *Asian option* with floating strike. These options have a payoff based on the difference between the terminal asset price and the average of an underlying asset price over a specified time period. The other type of Asian option is the fixed-strike option, where the payoff is determined by the average of an underlying asset price and a fixed strike set in advance (*see Asian Options*).

If the average is computed using a finite sample of asset price observations taken at a set of regularly spaced time points, we have a discrete average strike option. A continuous time option is obtained by computing the average *via* the integral of the price path over an interval of time. The average itself can be defined to be geometric or arithmetic. As for the fixed strike Asian option, when the geometric average is used, the average strike option has a closed-form solution for the price, whereas the option with arithmetic average does not have a known closed-form solution.

We concentrate on the continuous time, average strike option of European style with arithmetic averaging. A discussion of the uses and rationale for introducing Asian contracts is given in **Asian Options**. Average strike options are closely related to these options, but are less commonly used in practice.

Consider the standard Black–Scholes economy with a risky asset (stock) and a money market account. We also assume the existence of a risk-neutral probability measure Q (equivalent to the real-world measure P) under which discounted asset prices are martingales. We denote expectation under measure Q by E , and the stock price follows

$$\frac{dS_t}{S_t} = (r - \delta) dt + \sigma dW_t \quad (1)$$

where r is the constant continuously compounded interest rate, δ is a continuous dividend yield, σ is the instantaneous volatility of asset return, and W is a Q -Brownian motion. The reader is referred to **Black–Scholes Formula** for details on the Black–Scholes model and **Risk-neutral Pricing** for a discussion of risk-neutral pricing.

We consider a contract that is based on the value A_T , where $(A_t)_{t \geq t_0}$ is the arithmetic average

$$A_t = \frac{1}{t - t_0} \int_{t_0}^t S_u du, \quad t > t_0 \quad (2)$$

and by continuity, we define $A_{t_0} = S_{t_0}$. The corresponding geometric average G_t is defined as

$$G_t = \exp \left(\frac{1}{t - t_0} \int_{t_0}^t \ln S_u du \right) \quad (3)$$

The contract is written at time 0 (with $0 \leq t_0$) and expires at $T > t_0$. It is of interest to calculate the price of the option at the current time t , where $0 \leq t \leq T$. The position of t compared to the start of the averaging, t_0 may vary, as described in **Asian Options**.

The payoff of an average strike call with arithmetic averaging is given as

$$(S_T - A_T)^+ \quad (4)$$

and the payoff of an average strike put with arithmetic averaging is

$$(A_T - S_T)^+ \quad (5)$$

Average strike option payoffs with geometric averaging are identical, with A_T replaced by G_T . The buyer of an average strike call is able to exchange the terminal asset price for the average of the asset price over a given period. For this reason, it is sometimes referred to as a *lookback* on the average (*see Lookback Options* for a discussion of the lookback option).

By standard arbitrage arguments, the time- t price of the average strike call is

$$e^{-r(T-t)} \mathbb{E}[(S_T - A_T)^+ | \mathcal{F}_t] \quad (6)$$

and the price of the put is

$$e^{-r(T-t)} \mathbb{E}[(A_T - S_T)^+ | \mathcal{F}_t] \quad (7)$$

It turns out that we need to consider only the case $t \geq t_0$, where the option is “in progress”. The forward starting case ($t < t_0$) can be rewritten as a modified option with averaging starting at t , today. This is in contrast to the Asian option with fixed strike, where the difficult case was when the option was forward

starting. As for the Asian option, the average strike option satisfies a put–call parity; see [1] for details.

The average strike option is an exotic path-dependent option, as the price depends on the path of the underlying asset *via* the average. The distribution of the average A_T is not lognormal, if the asset price is lognormal, and pricing is difficult because the joint law of A_T and S_T is needed. This is in contrast to the Asian option, which required only the law of the average. Perhaps because of this increased complexity, or their lesser popularity in practice, fewer methods exist for the pricing of average strike options. Just as for the Asian option, there are no closed-form solutions for the price of the average strike option.

Many of the methods that we discuss here for pricing are similar to those used to price the Asian option. An early technique to give an approximate price for the average strike option was to replace the arithmetic average A_T with the geometric average G_T . Since G_T has a lognormal distribution, the (approximate) pricing problem becomes (for a call)

$$e^{-r(T-t)} \mathbb{E}[(S_T - G_T)^+ | \mathcal{F}_t] \quad (8)$$

We recognize that this is exactly an exchange option (*see Exchange Options*), which can be priced *via* a change of measure, as in [9]. Levy and Turnbull [8] mentioned this connection to exchange options, but it was Conze and Viswanathan [3] who presented the results of this computation.

Other analytical approximations can be obtained by approximating the true joint distribution of the arithmetic average and asset price using an approximate distribution, usually jointly lognormal with appropriate parameters. Chung *et al.* [2] extended the linear approximations of Bouaziz *et al.* [1], Levy [7], and Ritchken *et al.* [10] (approximating distribution of $\{A_T, S_T\}$ by joint lognormal) to include quadratic terms. Their approximation is no longer based on a geometric-type approximation.

Recently, symmetries of a similar style to that of the put-call symmetry have been found between fixed strike Asian options and average strike options. For forward starting average strike options, Henderson *et al.* [4] gave a symmetry with a starting Asian option. If the average strike option is starting, the special case of Henderson and Wojakowski [5] is recovered. If the average strike option is in progress, it cannot be rewritten as an Asian option, and Henderson *et al.* [4] derived an upper bound for the

price of the average strike option. This bound is in terms of an Asian option with fixed strike and a vanilla option. The method gives an “exact” bound for forward starting and starting options and when expiry is reached.

Numerical methods can be used to price the average strike option. The discussion of Monte Carlo simulation in **Asian Options** is also relevant here, as simulation is often used as a benchmark price. Ingersoll [6] was the first to recognize that it is possible to reduce the dimension of the pricing problem for the average strike option using a transformation of variables. Despite the value of the average strike option at t depending on the current asset price, current value of the average, and time to expiry, a one-dimensional partial differential equation (PDE) can be derived by using Ingersoll’s reduction of variables. However, the drawback is that the Dirac delta function appears as a coefficient of the PDE, making it prone to instabilities. Vecer’s [12] PDE method for Asian options with fixed strike also applies to average strike options and gives a stable one-dimensional PDE. Some testing of this method for the average strike option is given in [11].

To conclude, research into pricing the average strike option is ongoing, with current PDE and bound methods being very efficient.

References

- [1] Bouaziz, L., Briys, E. & Crouhy, M. (1994). The pricing of forward starting asian options, *Journal of Banking and Finance* **18**(5), 823–839.
- [2] Chung, S., Shackleton, M. & Wojakowski, R. (2003). Efficient quadratic approximation of floating strike Asian option values, *Finance* **24**(1), 49–62.
- [3] Conze, A. & Viswanathan, R. (1991). European path dependent options: the case of geometric averages, *Finance* **12**(1), 7–22.
- [4] Henderson, V., Hobson, D., Shaw, W. & Wojakowski, R. (2007). Bounds for in-progress floating-strike Asian options using symmetry, *Annals of Operations Research* **151**, 81–98.
- [5] Henderson, V. & Wojakowski, R. (2002). On the equivalence of fixed and floating-strike Asian options, *Journal of Applied Probability* **39**(2), 391–394.
- [6] Ingersoll, J. (1987). *Theory of Financial Decision Making*, Rowman and Littlefield Publishers, New Jersey.

-
- [7] Levy, E. (1992). Pricing European average rate currency options, *Journal of International Money and Finance* **11**(5), 474–491.
- [8] Levy, E. & Turnbull, S. (1992). Average intelligence, *Risk* **5**, 2.
- [9] Margrabe, W. (1978). The value of an option to exchange one asset for another, *Journal of Finance* **33**, 177–186.
- [10] Ritchken, P., Sankarasubramanian, L. & Vijn, A.M. (1993). The valuation of path-dependent contracts on the average, *Management Science* **39**(10), 1202–1213.
- [11] Shiuan, Y.J. (2001). *Pricing Floating-Strike Asian Options*. MSc dissertation, University of Warwick.
- [12] Vecer, J. (2001). A new pde approach for pricing arithmetic average Asian options, *Journal of Computational Finance* **4**(4), 105–113.

Related Articles

Asian Options; Black–Scholes Formula; Exchange Options; Lookback Options; Risk-neutral Pricing.

VICKY HENDERSON

Foreign Exchange Markets

The foreign exchange (FX) market has two major functionalities, one related to hedging and the other to investment.

In the age of globalization, it is essential for corporates and multinationals to hedge their FX exposure due to export/import activities. In addition, fund managers (institutional) need to hedge their FX risk in stocks or bonds if the stocks/bonds are quoted in a foreign currency. With hedging instruments, the FX exposure can be reduced and one can even benefit from certain market scenarios. This kind of participation brings us to the important class of investor-oriented products where the coupon depends on an FX rate or, at maturity, the pay-off (amount, currency) will be determined by an FX rate. This kind of product can be issued as a note, certificate, or bond.

For the major currencies such as USD, EUR, JPY, GBP, CHF, AUD, CAD, and NZD, the market has become more transparent over the last few years. For plain vanilla options, market data, especially volatilities for maturities below 1 year, are published by brokers or banks and are shown on Reuters pages (e.g., TTKLINDEX10, ICAPFXOP, GFIVOLS). For exotic products, new pricing tools, such as Superderivatives, LPA, Bloomberg, ICY, Fenics, and so on, are available for users, but the premium of the option will depend on the pricing model and the adjustments used. For the emerging market currencies such as PLN (Polish zloty), HUF (Hungarian forint), ZAR (South African rand), and so on, which are freely tradable but less liquid, the market data are less transparent. Currencies that are not freely tradable (the currency cannot be cash-settled offshore) such as BRL (Brazilian real) or CNY (Chinese yuan renmimbi) can be traded as a nondeliverable forward (NDF) or as a nondeliverable option (NDO) against a tradable currency. The NDF is a cash-settled product without exchange of notionals, which means that the intrinsic value at maturity will be paid in the free tradable currency based on a fixing source. The underlying of an NDO is the NDF, meaning that exercising the NDO results in an NDF, which will also be cash-settled. Another class of currencies is that of the fully cash-settled pegged ones, which means that their

exchange rate is 100% correlated to a major currency, mostly the USD. If one expects that this peg will continue, hedges should be done in the correlated major currency. In the case of SAR (Saudi riyal) or AED (United Arab Emirates dirham), discussion has been ongoing about depegging the currencies. In case this is done, there could be an increasing interest in SAR- or AED-linked investments. This opens an increasing interest in SAR- or AED-linked investments to participate in the case that these currencies are depegged. For the “more exotic currencies” such as the GHC (Ghanaian cedi), there is no options market.

Quotation

The exchange rate can be defined as the amount of domestic currency one gets if one sells one unit of foreign currency. If we take a look at an example of the EUR/USD exchange rate, the default quotation is EUR-USD, where USD is the domestic currency and EUR is the foreign currency. The terms *domestic* and *foreign* are not related to the location of the trader or any country, but it is more a question of the definition. Domestic and base are synonyms as are foreign and underlying. The common way is to denote the currency pair with a slash (/) and the quotation with a dash (–). The slash (/) does not mean a division.

For example, the currency pair EUR/USD can be quoted either in EUR-USD, which means how many USD one gets for selling one EUR, or in USD-EUR, which then means how many EUR one gets for selling one USD. There are certain market standard quotations; some of them are listed in Table 1.

In the FX market, two currencies are involved, which means that one needs to specify on which currency a particular call or put option is written. For

Table 1 Market convention of some major currency pairs with sample spot price

Currency pair	Quotation	Quote
EUR/USD	EUR-USD	1.4400
GBP/USD	GBP-USD	1.9800
USD/JPY	USD-JPY	114.00
USD/CHF	USD-CHF	1.1500
EUR/CHF	EUR-CHF	1.6600
EUR/JPY	EUR-JPY	165.00
EUR/GBP	EUR-GBP	0.7300
USD/CAD	USD-CAD	0.9800

instance, in the currency pair EUR/USD, there can be a EUR call, which is equivalent to a USD put, or a EUR put, which is equivalent to a USD call.

FX Terminology

In the FX market, a million is called a *buck* and a billion a *yard*. This is because the word “billion” has different meanings in different languages. In French and German, it represents 10^{12} and in English it stands for 10^9 .

Certain currency pairs have their own names in the market. For instance, GBP/USD is called a *cable*, because the exchange rate information used to be sent between England and America through a telephone cable in the Atlantic Ocean. EUR/JPY is called the *cross*, because it is the cross rate of the more liquidly traded USD/JPY and EUR/USD.

Some currency pairs also have their own names to make them short and unique in communication. New Zealand dollar, which is NZD/USD, is called *Kiwi*, and the Australian dollar, which is AUD/USD, is called *Aussie*. Among the Scandinavian currencies, NOK (Norwegian krone) is called *Noki*, SEK (Swedish krona) is called *Stoki*, and in combination with DKK (Danish krone) the three are called *Scandies*.

The exchange rates are usually quoted in five relevant figures, for example, in EUR-USD we would get a quote of 1.4567. Sometimes one can get a quote up to six figures, but for the time being we focus on five figures. The last digit “7” is called the *pip* and the middle digit “5” is called the *big figure*, because the interbank spot trading tools show this digit in bigger size since it is the most important information. The figure to the left of the big figure is known anyway and the pips to the right of the big figures are sometimes “negligible”. For example, a rise of EUR-JPY 165.00 by 40 pips is 165.40 or a rise by 3 big figures would be 168.00.

Quotation of Option Prices

Plain vanilla option prices are usually quoted in terms of implied volatility. If an option is priced in volatility, a delta exchange is necessary. The advantage is that the volatility does not usually move as quickly as the spot rate and one has the chance to

Table 2 Standard market quotation types for option premiums

Symbol	Description of symbol	Result of example
d pips	Domestic per unit foreign	208.42 USD pips per EUR
f pips	Foreign per unit domestic	97.17 EUR pips per USD
$\%f$	Foreign per unit foreign	1.4575% EUR
$\%d$	Domestic per unit domestic	1.3895% USD
d	Domestic amount	20 842 USD
f	Foreign amount	14 575 EUR

Foreign = EUR, domestic = USD, $S_0 = 1.4300$, $r_d = 5.0\%$, $r_f = 4.5\%$, volatility = 8.0%, $K = 1.5000$, $T = 365$ days, EUR call USD put, notional = 1 000 000 EUR = 1 500 000 USD

compare the prices, especially in the broker market. On the basis of the spot rate on which the delta exchange is done, the premium of the plain vanilla option is calculated *via* the Black–Scholes formula. For exotic options, a price in volatility is not possible because each bank has its own pricing model for these.

The premium, value, or prices of options can be quoted in six different ways (Table 2). The Black–Scholes formula quotes in domestic pips per one unit of foreign notional. The others can be retrieved in the following manner:

$$d \text{ pips} \xrightarrow{\times \frac{1}{SK}} f \text{ pips} \xrightarrow{\times K} \%f \xrightarrow{\times \frac{S}{K}} \%d \quad (1)$$

Delta and Premium Convention

The spot delta of a plain vanilla option can be retrieved in a straightforward way by using the Black–Scholes formula. It is called the *raw spot delta*, δ_{raw} . One retrieves it in percentage of the foreign currency, but the delta in the second involved currency $\delta_{\text{raw}}^{\text{opposite}}$ can be computed in the following manner:

$$\delta_{\text{raw}}^{\text{opposite}} = -\frac{S}{K} \delta_{\text{raw}} \quad (2)$$

The delta multiplied with the corresponding notional determines the amount that has to be bought

or sold to hedge the spot risk of the option up to the first order.

An important question is whether the premium of the option needs to be included in the delta or not? An example, EUR-USD, for investigation is considered here. In this quotation, USD is the domestic currency and EUR is the foreign one. The Black–Scholes formula calculates the premium in domestic per 1 unit foreign currency, which in our example is in USD per 1 EUR. This premium is denoted by p . If the premium is paid in EUR, which means in the foreign currency, it includes an FX risk. The premium p in USD is equivalent to $\frac{p}{S}EUR$, which means that the amount of EUR that has to be bought to hedge the option needs to be reduced by this EUR premium and is given as

$$\delta_{\text{raw}} - \frac{p}{S}EUR \quad (3)$$

We denoted USD as domestic currency and EUR as foreign currency, but do all banks or trading places have this notion? What is the notional currency of the option and what is the premium currency? In the interbank market, there exists a fixed notion of the delta of the currency pair. Normally, it is the LHS delta in Fenics^a if the option is traded in the LHS premium, which is mostly used, for example, for EUR/USD, USD/JPY, and EUR/JPY, and the RHS delta if it is the RHS premium, for example, for GBP/USD and AUD/USD. Most of the options traded in the market are out-of-the-money; therefore, the premium does not create a critical FX risk for the trader.

For the banks where the base currency is considered the risk-free currency, the market value of the option is in the base currency, and if the premium is in the risky currency, the premium needs to be included in the hedge. If the premium is in the risk-free (or the base) currency, the premium will be offset by the market value of the option. In the opposite case, where the risk-free currency is the underlying currency, if the premium is in the risky currency, the premium will be offset by the market value of the option. Only in the case of premium in risk-free currency, the amount needs to be included in the hedge.

Therefore, the delta hedge is invariant with respect to the risky currency notion of the bank; for example, for both banks, one is based in USD and the other in EUR, and the delta is the same.

Table 3 One-year EUR call USD put, the strike is 1.4300 for a EUR-based bank

Delta currency	Premium currency	Fenics	Hedge	Delta
%EUR	EUR	LHS	$\delta_{\text{raw}} - P$	48.35
%EUR	USD	RHS	δ_{raw}	51.64
%USD	EUR	RHS + F4	$-(\delta_{\text{raw}} - P) \frac{S}{K}$	-48.35
%USD	USD	LHS + F4	$-\delta_{\text{raw}} S/K$	-51.64

$S = 1.4300$, $r_d = 5.0\%$, $r_f = 4.5\%$, volatility = 8.0%, $K = 1.4300$

Table 4 One-year EUR call USD put, the strike is 1.5000 for a EUR-based bank

Delta currency	Premium currency	Fenics	Hedge	Delta
%EUR	EUR	LHS	$\delta_{\text{raw}} - P$	28.22
%EUR	USD	RHS	δ_{raw}	29.69
%USD	EUR	RHS + F4	$-(\delta_{\text{raw}} - P) \frac{S}{K}$	-26.91
%USD	USD	LHS + F4	$-\delta_{\text{raw}} S/K$	-28.30

$S = 1.4300$, $r_d = 5.0\%$, $r_f = 4.5\%$, volatility = 8.0%, $K = 1.5000$

Examples

To see the different deltas used in practice, consider two examples discussed in Tables 3 and 4.

Implied Volatility and Delta for a Given Strike

Implied volatility is not constant across strikes (*see Foreign Exchange Smiles*). The volatility σ depends on the corresponding delta of the option, but the delta depends on the price of the option and therefore on the used volatility. How can we retrieve the correct volatility for a given strike? For sure it is an iterative process. Initially, one uses the at-the-money (ATM) volatility σ_0 and calculates the delta Δ_1 . On the basis of Δ_1 , a new volatility σ_1 can be retrieved from the volatility matrix. This new volatility leads to a new delta and so on. Now one can define a convergence criterion to stop the iteration. In practice, a fixed number of iterations is used, usually five steps.

4 Foreign Exchange Markets

Table 5 Vega matrix for standard maturities and delta values, expressed in percent of foreign notional

Mat/ Δ	50%	45%	40%	35%	30%	25%	20%	15%	10%	5%
<i>O/N</i>	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01
1W	0.06	0.06	0.05	0.05	0.05	0.04	0.04	0.03	0.02	0.01
1M	0.11	0.11	0.11	0.10	0.10	0.09	0.08	0.07	0.05	0.03
2M	0.16	0.16	0.15	0.15	0.14	0.13	0.11	0.09	0.07	0.04
3M	0.20	0.02	0.19	0.19	0.17	0.16	0.14	0.12	0.09	0.05
6M	0.28	0.28	0.27	0.26	0.25	0.23	0.20	0.17	0.13	0.07
9M	0.33	0.33	0.33	0.32	0.30	0.28	0.24	0.20	0.15	0.09
1Y	0.38	0.38	0.38	0.37	0.35	0.32	0.28	0.24	0.18	0.10
2Y	0.51	0.51	0.51	0.50	0.48	0.44	0.40	0.33	0.25	0.15
3Y	0.60	0.60	0.60	0.60	0.57	0.54	0.48	0.40	0.31	0.18

The matrix shows, for example, that 2Y EUR call USD put 35 delta can be hedged with two times 6M EUR call USD put 30 delta

Mapping of Delta on Vega

From the Black–Scholes formula, it is clear that for a fixed delta, the vega P_σ does not depend on volatility or r_d , and P_σ is therefore a function of only r_f , maturity, and delta. This gives the trader the advantage of a moderately stable vega matrix. Such a matrix is shown in Table 5, with $r_f = 4.5\%$.

FX Smile

The FX smile surface has a different setup or construction in comparison to equity.

Plain vanilla options with different maturities have different implied volatilities. This is called the *term structure* of a currency pair.

Plain vanilla options are quoted in terms of volatility for a given delta. The smile curve is usually set up on some fixed pillars and the points between these pillars are interpolated to get a smooth surface. In the direction of the term structure, the easiest way is to interpolate linearly in the variance. In addition, weights are introduced to highlight or lower the importance of some dates, for example, the release of nonfarm payrolls, local holiday, or a day before or after a weekend. In the direction of the moneyness or delta, one method of interpolation is the cubic spline. The pillars in that direction are 10-delta put, 25-delta put, ATM, 25-delta call, and 10-delta call. Sometimes the 35-delta put and 35-delta call are also used. Unlike equity, in the FX market, the smile surface is decomposed into the symmetric part by using butterflies or strangles and the skew part by using the risk reversals for the fixed deltas. This is

because of the liquidity of these products in the FX market.

Risk Reversal (RR)

For instance, a 25-delta risk reversal is a combination of buying a 25-delta call and selling a 25-delta put. The payout profile is shown in Figure 1.

Butterfly (BF)

In the case of a 25-delta butterfly, it is the combination of buying 25-delta put, buying 25-delta call, selling ATM put, and selling ATM put (alternatively, 25-delta strangle is a 25-delta put and a 25-delta call). Payout profiles are shown in Figure 2.

The decomposition of a smile curve inspired by these products is shown in Figure 3.

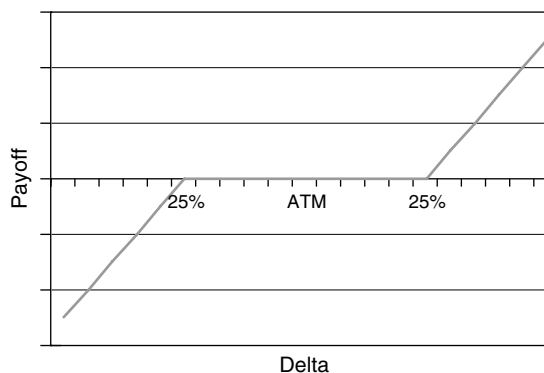


Figure 1 Payout profile of a risk reversal

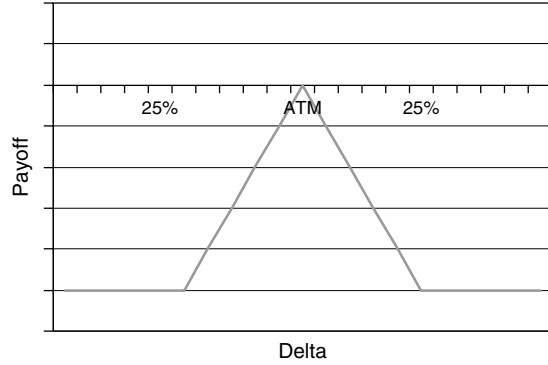


Figure 2 Payout profile of a butterfly

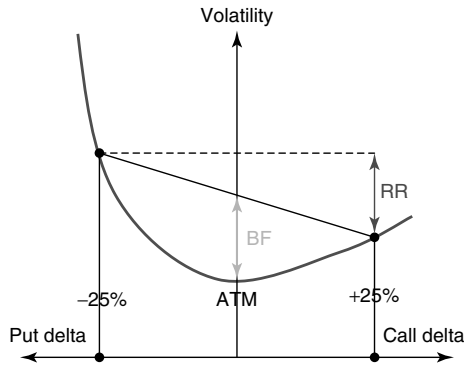


Figure 3 Decomposition of a smile curve

If we denote the ATM volatility by σ_0 , the 25-delta put volatility by σ_- , and the 25-delta call volatility by σ_+ , we get the following relationships:

$$RR = \sigma_+ - \sigma_- \quad (4)$$

$$BF = \frac{1}{2}(\sigma_+ + \sigma_-) - \sigma_0 \quad (5)$$

$$\sigma_+ = ATM + BF + \frac{1}{2}RR \quad (6)$$

$$\sigma_- = ATM + BF - \frac{1}{2}RR \quad (7)$$

It should be noted that the values RR and BF given above have nothing to do with the prices of actual risk reversal and butterfly contracts: rather they provide a convenient representation of the implied volatility smile in terms of its level (σ_0), convexity (BF), and skewness (RR).

Table 6 EUR/USD 25-delta risk reversal (in %)

Date	1 month	3 months	1 year
Dec 3, 2007	-0.6	-0.6	-0.6
Dec 4, 2007	-0.525	-0.55	-0.525
Dec 5, 2007	-0.525	-0.55	-0.525
Dec 6, 2007	-0.525	-0.55	-0.525
Dec 7, 2007	-0.6	-0.6	-0.6
Dec 10, 2007	-0.6	-0.6	-0.6

Table 7 EUR/USD 25-delta butterfly (in %)

Date	1 month	3 months	1 year
Dec 3, 2007	0.225	0.425	0.460
Dec 4, 2007	0.225	0.425	0.460
Dec 5, 2007	0.225	0.425	0.460
Dec 6, 2007	0.225	0.425	0.460
Dec 7, 2007	0.227	0.425	0.463
Dec 10, 2007	0.227	0.425	0.463

Table 6 shows the 25-delta risk reversal in EUR/USD on different trading dates and the corresponding butterflies are listed in Table 7.

For the setup of the smile surface in a risk management or pricing tool, it is important to know which convention is used for a certain currency pair in the option market to define the notion of ATM.

At-the-money Definition

There exist several definitions of ATM:

- ATM spot: Strike is equal to the spot.
- ATM forward: Strike is equal to the forward.
- Delta parity: The absolute value of the delta call is equal to the absolute value of the delta put.
- Fifty delta: Put delta is 50% and the call delta is 50%.
- Value parity: The premium of the delta put is equal to the delta call.

The most widely used one in the interbank market is the delta parity up to 1 year for the most liquid currencies. In emerging markets, (at-the-money-forward) (ATMF) is used. For long-term options such as USD/JPY 15 years, the ATMF convention is used, but since this results in a delta position, a forward delta exchange will be done.

End Notes

^aFenics is an FX option pricing tool owned by the broker GFI and used in the interbank market (www.fenics.com).

Further Reading

Hakala, J. & Wystup, U. (2002). *Foreign Exchange Risk*, Risk Publications, London.

Reiswich, D. & Wystup, U. (2009). *FX Volatility Smile Construction*, Research Report, Frankfurt School of Finance & Management, September 2009.

Wystup, U. (2006). *FX Options and Structured Products*, Wiley Finance.

Related Articles

Black–Scholes Formula; Exchange Options; Foreign Exchange Options; Foreign Exchange Options: Delta- and At-the-money Conventions; Foreign Exchange Smiles; Foreign Exchange Smile Interpolation.

MICHAEL BRAUN

Foreign Exchange Options

Market Overview

The importance of foreign exchange (FX) options for risk management and directional trades is gaining more and more recognition from companies and investors. Various banks have been adapting their products to this situation during the past years. Different risk and profit profiles can be generated with plain vanilla or exotic options as individual products, as well as in combination with various products such as structured products. Financial engineers call this as *playing with Lego bricks*. Linear combinations of basic products are used to build structured products. To price plain vanilla or exotic options and show their risk, many professional trading systems have been introduced and are being continuously developed. With these systems, the traders are able to evaluate the positions in the individual currency pair or in currency portfolios at any time. In the FX options market, options trading systems such as *Fenics*, *Murex*, or *SuperDerivates* are used. Owing to the very rapid development in this sector, some banks started developing and using systems of their own.

To comply with various customer requests, a successful trading desk in the interbank market is essential. This is mostly plain volatility trade. The market risk of a short-term FX options trading desk consists of changes in spot, volatility, and interest rates. Since spot risk is easily eliminated by **delta hedging** and the effect of rates is small compared to the risk of changing volatility in the short term up to two years, managing volatility risk is the main task of the trader. Since the relationship of volatility and price of call or put options is monotone, it is equivalent to quote the price of an option either by the price itself or by the volatility implied by the **Black–Scholes formula**. The established market standard is quoting this implied volatility, which is why it is often viewed as a traded quantity. In the case of plain vanilla options, a **vega** long position is given when buying (call or put); conversely, a vega short position is given when selling. Volatility difference between call and put with same expiry and same deltas is called a *risk reversal*. If the risk reversal is positive, the market is willing to pay more for calls than for puts; if the risk reversal is negative, the

market favors put options. The butterfly measures the convexity of the “smile” of the volatility, that is, the volatility for the out-of-the-money and the in-the-money-options (*see Foreign Exchange Markets* for details).

If the **delta hedge** is done with an interbank partner at the same time the option is traded, the trader can focus on the vega position in his book. The **delta hedge** neutralizes the change of the option price caused by changes of the underlying. For long-term options with an expiry longer than two years or options with high interest rate sensitivity, the **delta hedge** should be replaced by a forward hedge, as the risk of interest rate sensitivity is mostly higher than the volatility risk in this case. This means that instead of neutralizing spot risk by trading in the spot market, one would trade a forward contract with maturities matching those of the cash flows of the option. This would simultaneously take care of the spot and the rate risk.

First-generation Exotic Options

First-generation exotic options are all options beyond plain vanilla options that started trading in the 1990s, in particular, barrier options, digital and touch products, average rate or Asian options, and lookback and compound options. There is no strict separation between first- and second-generation exotics as the viewpoint on what is first and second varies by the person in charge. Exotic options are traded live in currency trading as opposed to plain vanilla options, which mostly trade through automated systems. Trading exotic options is done by quoting the bid and ask price of the product rather than the corresponding volatility, because the monotone relationship of volatility and price is often not guaranteed. When asking for a quote, the spot reference level is agreed upon at which the option is calculated and priced. This allows comparing quotes of the exotic options and is also the basis of the **delta hedge**. To keep the **vega** risk low when fixing a deal, a vega hedge can be done with the partner. In this case, plain vanilla options (calls/puts, at-the-money ATM straddles) are traded to offset the vega of the exotic option. The default vega hedge is done with a straddle—an out-of-the-money call and an out-of-the-money put—the reason being that this product does not have any delta, so one offsets the vega position without touching

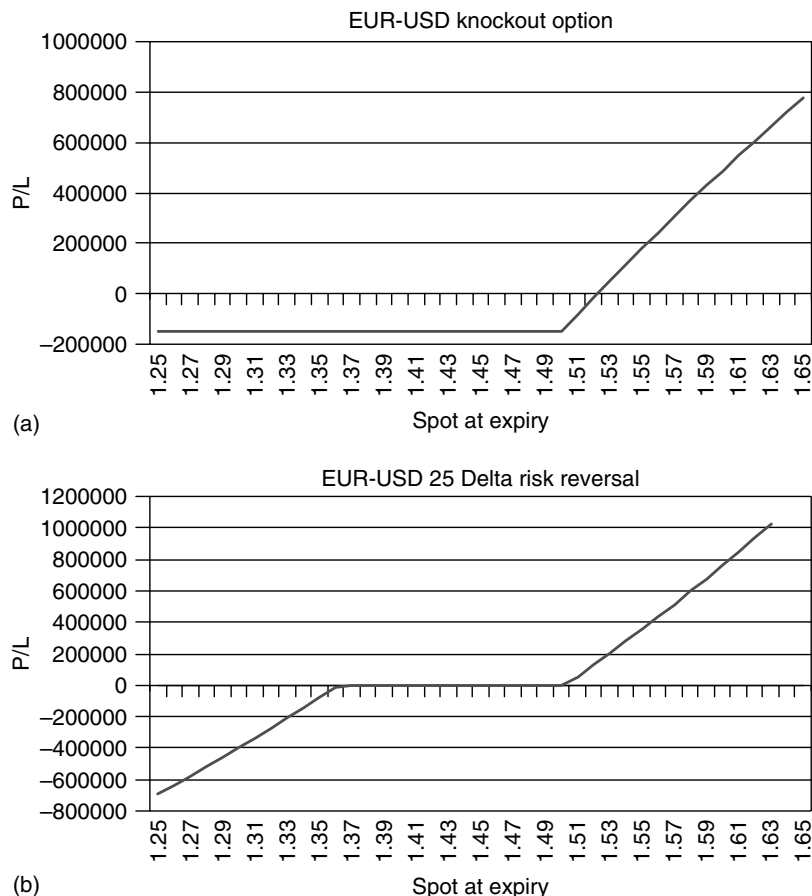


Figure 1 (a) Payoff profile of a EUR-USD knockout option that is not knocked out during its lifetime; (b) payoff profile of the EUR-USD risk reversal at expiry

the delta position. Normally, during the lifetime of the option, the risk is hedged dynamically across the entire option book.

The quoting bank (market maker) is the calculation agent. It stipulates the regulations under which predefined triggers are reached or how often the underlying is traded in certain predefined ranges. The market maker informs the market user about the trigger event.

Barrier Options

Barrier options are vanilla put and call options with additional barriers. In case of a knockout, the option expires worthless, if the spot ever trades at or beyond the prespecified barrier. In case of a knockin option,

the option is only activated if the spot ever trades at or beyond the prespecified barrier. The barrier is valid at all times between inception of the trade and the maturity time of the option.

One can further distinguish regular barrier options, where the barrier is out-of-the-money, and reverse-barrier options, where the barrier is in-the-money.

A regular knockout barrier option can basically be priced and semistatically hedged by a risk reversal (Lego-brick principle).

Figure 1 illustrates the example: EUR-USD Spot 1.4600 expiry six months, strike 1.5000, EUR CALL with regular knockout trigger at 1.4300.

Hedging a short regular knockout EUR call, we can go long a vanilla EUR call with the same strike and the same expiry and go short a vanilla EUR put

with a strike such that the value of the hedge portfolio is zero if the spot is at the barrier. The long call and short put is called a *risk reversal* and its market price can be used as a proxy for the price of the regular knockout call. In our example, it would be a 1.3650 EUR put. If the trigger is not reached, then the put expires worthless and the call offsets the knockout call payoff. If the trigger is reached, the risk reversal can be canceled with approximately zero value. The delta of a knockout option is higher than the delta of the corresponding plain vanilla option, and the higher it is, the closer the trigger is to the underlying spot.

Reverse knockout and reverse knockin are more difficult to price and hedge as the risk profile of these options is difficult to replicate with other options. In this case, the trigger is in the money. The volatility risk of first and second order arising from these options can be hedged dynamically with risk reversals and butterflies (see **Vanna–Volga Pricing**). However, all sensitivities take extreme values when getting closer to the trigger and closer to maturity. Delta positions can be a multiple of the notional amount. Therefore, it is difficult for the trader to perform dynamic hedging strategies. To manage these risks, short-term reverse knockout barrier options are often removed from the global books and are matched as individual positions, or are closed two to three weeks before expiry. The risk surcharge paid in this case is often smaller than the cost of keeping to such positions and hedging them individually.

Modifications and Extensions of Barrier Options

Standard extensions of barrier options are *double-barrier options*, where there is a barrier above and below the current spot. A double knockout option expires worthless if any of the two barriers are ever touched or crossed. A double knockin option only becomes a vanilla option if at least one of the two barriers is touched or crossed in the underlying.

A further modification of barrier options is called the *knockin/knockout (KIKO)* option. This option can knockout at any time; however, it must knockin to become alive. A short KIKO option can be statically hedged with a long knockout option and a short double knockout option, if the spot value is between the triggers, and with a long knockout option and a

short knockout option, if the spot value is above both triggers (Lego-brick principle).

Window barriers (partial barriers) are additional modifications of barrier options. In case of a window-barrier option, the trigger is valid only within a certain period of time. Commonly, this period of time is from inception of the trade until a specific date (*early ending*) or from a specific date during validity until expiry date of the option (*deferred start*). Arbitrary time intervals are possible.

For *European barrier options*, the triggers are only valid at maturity. They can be statically hedged with plain vanilla options and European digital options (Lego-brick principle).

Binary Options/Digital Options

Digital or binary options pay a fixed amount in a currency to be specified if the spot trades at or beyond a prespecified barrier or trigger. For European digitals, the trigger is valid only at maturity, whereas for American digitals, the trigger is valid during the entire lifetime of the trade. In FX-interbank trade American digitals are also called *one-touch* (if the fixed amount is paid at maturity) or *instant one-touch* (if the fixed amount is paid at first hitting time) options. Further touch options are the so-called *no-touch* options, *double no-touch* options, and *double one-touch* options. A no-touch pays only if the spot never touches or crosses the prespecified trigger. A double no-touch pays only if neither the upper trigger nor the lower trigger is ever touched or crossed during the lifetime of the contract. A double one-touch pays only if at least one of the upper or the lower triggers is touched. When buying a double no-touch option, a vega short position is generated. This means that double no-touch options are cheap in phases of high volatility.

European digital options can be replicated with bull or bear spreads with large amounts. Their market price can thus be approximated by liquid vanilla options. However, this type of option is difficult to hedge as the delta hedge close to expiry is zero almost everywhere.

General Features When Pricing Exotic Options

Most commercial software packages calculate the “theoretical value (TV)” of the exotic options, which

is the value of the product in a **Black–Scholes model** with constant parameters.

Knowing the TV is important for trading partners as it serves as a checksum to ensure that both parties talk about the same product. The market value, however, often deviates from this value because of so-called overhedge costs, which arise when hedging the exotic option. Every trader must be aware of the risk arising from these options and should be able to control this risk dynamically in his books via the Greeks (price sensitivity with respect to market and model parameters). If a gain is generated by performing this hedge, the price of an exotic option must be higher than the TV. Conversely, if the hedge leads to a loss, the market price of the exotic option should be above TV.

A very important issue when trading exotic options is placing automatic spot orders at spot levels that could lead to a knockout or expiry of the option. This order eliminates the **delta hedge** of the option automatically when reaching the trigger. This explains the occasional very heavy spot movements during specific trigger events in the market.

The following vega structure is often found in options books as it stems from most of the **structured products** offered today in the FX range: ATM vega long and wing vega short. This is the reason for a long phase of low volatility and high butterflies for the past years. See also **Foreign Exchange Smiles**.

Second-generation Exotic Options

We consider every exotic option as second generation if it is not a vanilla and not a first-generation product.

Some of the common examples in FX markets are *range accruals* and *faders*.

A range accrual is a sum of digital call spreads and pays an amount of a prespecified currency that depends on the number of currency fixings that come to fall inside a prespecified range. A fader is any basic option product like a vanilla or barrier option, whose notional amount depends on the number of currency fixings that come to fall inside a prespecified range. We distinguish fade-in products, where the notional grows with each fixing inside the range and fade-out products, where the notional decreases with each fixing inside the range.

Further extensions are target redemption products, whose notional amount increases until a certain gain is reached. A common example is a **target redemption forward** (TRF). We provide a description and an example here: We consider a TRF in which a counterpart sells EUR and buys USD at a much higher rate than current spot or forward rates. The key feature in this product is that counterpart has a total target profit that, once hit, knocks out all future settlements (in the example below, all weekly settlements), locking the gains registered until then.

The idea is to place the strike over 5.5 big figures above spot to allow the counterpart to quickly accumulate profits and have the trade knocked out after five or six weeks. The counterpart will start losing money if EUR-USD starts fixing above the strike. On a spot reference of 1.4760, consider a one year TRF, in which the counterpart sells 1 EUR 1 million per week at 1.5335, subject to a knockout condition: if the sum of the counterpart profits reaches the target, all future settlements are canceled. We let the target be 0.30 (i.e., 30 big figures), measured weekly as $\text{Profit} = \text{Max}(0, 1.5335 - \text{EUR-USD Spot Fixing})$. As usual, this type of forward is also traded at zero cost:

Week 1 Fixing = 1.4800 Profit = 0.0535 Max (1.5335–1.4800, 0)

Week 2 Fixing = 1.4750 Profit = 0.0585 Accumulated profit = 0.1120

Week 3 Fixing = 1.4825 Profit = 0.0510 Accumulated profit = 0.1630

Week 4 Fixing = 1.4900 Profit = 0.0435 Accumulated profit = 0.2065

Week 5 Fixing = 1.4775 Profit = 0.0560 Accumulated profit = 0.2625

Week 6 Fixing = 1.4850 Profit = 0.0485 Accumulated profit = 0.3110

The profit is capped at 0.30, so the counterpart only accumulates the last 3.75 big figures and the trade knocks out.

Each forward will be settled physically every week until trade knocks out (if the target is reached).

Another popular FX product is the *time option*, which is essentially a forward contract of American style, that is, the buyer is entitled and obliged to trade a prespecified amount at a prespecified strike, but can choose the time within a prespecified time interval.

The market is likely to continue to develop fast. Besides *Bermudan style options*, where early exercise is allowed at certain prespecified times, **basket options** and the corresponding structures are very much in demand in the market. Hybrid structures are exotic options whose payoff depends on underlying spots across different market sectors. We refer the reader to [1].

References

- [1] Wystup, U. (2006). *FX Options and Structured Products*, John Wiley & Sons.

MARKUS CEKAN, ARMIN WENDEL &
UWE WYSTUP

Currency Forward Contracts

Executive Summary

Structured forwards use combinations of options to replicate forwards like payout profiles. The main use of structured forwards is for corporate and institutional clients trying to hedge their foreign exchange exposures. While standard forwards lock in a fixed exchange rate, structured forwards give the user the advantage of the possibility of an improved exchange rate, while still guaranteeing a worst-case rate. As standard forwards, structured forwards usually have no upfront premium requirements (zero-cost strategies). Having the chance of an improved exchange rate for no upfront premium suggests that structured forwards must have a guaranteed worst-case exchange rate that is worse than the prevailing forward rate. This is the risk involved when entering into structured forward transactions.

Forward Contract

A foreign exchange forward transaction involves two parties, who enter into a contract, whereby one counterparty agrees to sell a specified amount of a currency A in exchange for a specified amount of another currency B on a specified date. The other counterparty agrees to buy the specified amount of that currency A in exchange of the other currency B.

The Characteristics of Forward Contracts

Both counterparties have the obligation to fulfill the contract (as opposed to an option transaction where only one of the parties has the obligation, the option seller, while the other counterparty, the option buyer, has the right, but no obligation). As both currency amounts are fixed on the day the contract is entered into, the exchange rate between the two currencies is fixed. Hence, the parties to the contract know from the beginning at what exchange rate they are obliged to buy or sell the specified currency.

For corporate and institutional clients, this can be a useful information, as they can use this exchange rate to calculate the cost of production of a given product

or service. Another positive feature of forwards is that there is no upfront premium to be paid by either party. As both parties to the contract have an obligation to deliver and the contract is struck at the prevailing market forward rate, the transaction is by definition a zero-cost strategy. This is because the definition of the market forward rate is a future exchange rate of two currencies at a rate that demands no upfront payment from either party.

How can one calculate this market forward rate and what are the influencing factors?

Calculating the Market Forward Rate

The following example helps to determine the forward exchange rate of a given currency pair.

Market Information.

- Company X: London-based manufacturer exporting to the United States.
- The importing company: New York-based company importing from the United Kingdom.
- The bank: it is the other counterparty to the forward transaction.
- Company X sells its goods to the importing company. The sale is agreed in USD, and the payment of USD 100 000 is expected six months after the contract is signed. Therefore, the London-based company X has a foreign exchange exposure, as the change in the foreign exchange rate has an effect on its income in GBP.
- Current GBP/USD exchange rate: 2.0000. This means that 1 GBP is worth 2 USD.
- Current GBP interest rate: 6% per annum. This is the interest rate company X can borrow and lend in GBP.
- Current USD interest rate: 3% per annum. This is the interest rate company X can borrow and lend in USD.

What can company X do to eliminate its foreign exchange exposure? We know that company X will receive USD 100 000 in six months time. In fact they could already sell this USD 100 000 at the prevailing market rate (spot rate), but they don't have it yet. The solution is to go to the bank and borrow the USD 100 000. To be precise, they need to borrow less than USD 100 000, because they need to pay interest on the loan to the bank. So the exact amount to borrow

2 Currency Forward Contracts

is the net present value (NPV) of USD 100 000. To calculate this, we use the following formula:

$$NPV = \frac{N}{1 + r^*d/dc} \quad (1)$$

where

- N is the amount for which one wants to calculate the NPV. In this example, it is USD 100 000.
- r is the interest rate, expressed as percentage per annum, for the currency in which N is denominated.
- d is the duration of the deposit or loan in days. In this example, it is 180 days (i.e., six months).
- dc is the day-count fraction. This is usually 360, except for GBP deposits or loans where it is 365.

We are now able to calculate the amount company X has to borrow: $\text{USD } 100\,000 / (1 + 0.03 \times 180/360) = \text{USD } 98,522.17$. If they borrow this money, they have to pay back exactly USD 100 000 in six months time including the interest charge. This is the amount company X is due to receive in six months time from the sales of its goods to the import-export company.

If company X now sells the borrowed USD in the spot market and buys GBP, they receive GBP 49 261.08. This is calculated by dividing the borrowed USD amount by the current GBP/USD exchange rate (2.0000 in this example).

Company X now has the GBP and has eliminated the foreign exchange exposure. They can take GBP and deposit it with their bank at the current interest rate (6% in this example). The amount they get back after six months is equal to GBP 50 718.67—this is calculated as follows: $\text{GBP } 49\,261.08 \times (1 + 0.06 \times 180/365)$.

After these series of transactions company X is left with no cash position at the beginning of the transaction. They receive GBP 50 718.67 after six months and have to pay USD 100 000 in exchange. The exchange rate that is implied from the above-mentioned two amounts is 1.9717 (calculated as USD 100 000 divided by GBP 50 718.67).

What happens when a forward transaction is entered into? Exactly the same:

- At the beginning of the transaction:
 - company X has no cash position;
 - company X agrees to sell USD 100 000 for GBP at the market forward rate.

- At the end of the transaction:
 - company X pays USD 100 000 for the GBP amount exchanged at the agreed forward rate.

Because the two approaches described earlier have the same outcome, the GBP amount received for the USD 100 000 has to be the same; otherwise there would be an arbitrage opportunity. Therefore, the market forward rate in this example has to be 1.9717.

Generally, the forward rate can be calculated with a single and easy formula:

$$F_{\text{GBP/USD}} = S_{\text{GBP/USD}} \frac{1 + r_{\text{USD}}^*d/dc_{\text{USD}}}{1 + r_{\text{GBP}}^*d/dc_{\text{GBP}}} \quad (2)$$

where $F_{\text{GBP/USD}}$, forward rate for GBP/USD; $S_{\text{GBP/USD}}$, spot exchange rate for GBP/USD; r_{USD} , USD interest rate expressed in percentage per annum; r_{GBP} , GBP interest rate expressed in percentage per annum; d , duration of the deposit or loan in days; dc_{USD} , day-count fraction for USD (360); and dc_{GBP} , day-count fraction for GBP (365).

As the formula suggests, the market forward rate is a function of only the current (spot) exchange rate and the interest rates of the two currencies for the specified forward period. Hence, it is not market expectations, or any other factor that determines the arbitrage-free forward rate.

Structured Forwards

The previous section helped us to understand how a foreign exchange exposure resulting from a cross-border transaction can be eliminated and hedged through a forward transaction. It showed that the forward exchange rate was fixed right at the beginning of the contract and hence the uncertainty about exchange rate movements was turned into a known rate with which companies can calculate their cost of production. The example also demonstrated that there is no cash flow at the beginning of a forward transaction and there is no premium or any other fee associated with it. A forward transaction is by definition a zero-cost strategy.

The Difference between Forwards and Structured Forwards

The disadvantage of forwards is that favorable exchange rate moves are also eliminated when the

exchange rate is fixed. In the previous example, the forward rate was calculated to be 1.9717. This is the rate at which company X has to buy GBP and sell the USD. If in six months time the GBP/USD exchange rate falls below 1.9717, company X would be better off without hedging the GBP purchase through a forward.

Structured forwards allow just this. They are more flexible, because favorable exchange rate moves, and, in fact, any market view can be incorporated into the transaction to enhance the rate at which a currency is exchanged for another.

As with forwards, structured forwards offer the worst-case exchange rate. This rate is fixed at the beginning of the contract and similar to a regular forward, it offers the benefit of certainty about the exchange rate that can be used for financial planning.

Similar to standard forward contracts, most structured forward contracts are zero-cost strategies, that is, no upfront premium is required.

We all know that there is no such a thing as a “free lunch”. Therefore, to have the benefit of an improved exchange rate, a fixed worst-case rate, and a zero-cost strategy, the company entering into a structured forward transaction needs to take on certain risks. This risk is usually structured so that the guaranteed worst-case exchange rate is set at a rate that is worse than the prevailing market forward rate. The hedging counterparty accepts this worse guaranteed rate for the chance of receiving a better rate, in case a predefined condition is met. As the examples in the following section demonstrate, these predefined conditions can take many forms and may incorporate the market view of the counterparty entering into the structured forward transaction.

Examples of Structured Forwards

As mentioned in the previous section, structured forwards offer the possibility to incorporate one’s market view into a forward transaction. This view might be the appreciation or depreciation of a currency or even the view that a currency pair remains in a certain range over a given period of time. The following examples demonstrate how these different market views can be expressed with currency options that can be structured into the forward transaction. As a reminder: all examples follow the basic assumptions that the structured forward has a worst-case buying

(or selling) rate and no upfront premium must enter into the transaction.

Forward Plus. The forward plus is the simplest of all structured forwards. It offers the possibility to take advantage of favorable market movements up to a certain point, while still having a certain worst-case hedged rate.

How does it work: by accepting a worst-case hedge rate that is less favorable than the prevailing market forward rate, we create excess cash. Remember, trading at the market forward rate is zero cost by definition. If one trades on a rate that is worse than the market rate, one can expect some compensation. The cash generated is used to buy an option that pays out, if the underlying currency pair moves favorably. To make this a zero-cost strategy, we need to introduce a barrier, or knockout. This has the effect that options cease to exist (are knocked out) if the barrier is reached. For our strategy, it means that we can participate in a favorable market move, but only up to a certain point, namely, the predefined barrier level. If the barrier is reached we are locked into a forward transaction with a rate equal to the worst-case rate.

Let us continue the previous example with company X: We calculated the market forward rate to purchase GBP against USD in six months time to be 1.9717. A forward plus could have a worst-case buying rate of 1.9850. This rate is 0.0133 worse than the market forward rate. As compensation for accepting this hedge rate, company X has the opportunity to buy GBP at the prevailing spot rate in six months time as long as the barrier of 1.8875 is not reached or breached during the life of the contract. As the barrier is observed continuously during the entire life of the transaction, we call this barrier an American style barrier (this is not to be confused with an American style option that is exercisable during the life of the option). So what does this right to buy the GBP at the prevailing market spot rate in six months time give to company X? Imagine that the barrier was never reached and the spot rate in six months time is at 1.9000. In this case, company X may buy the GBP at 1.9000 and it will outperform the forward transaction that would have forced it to buy the GBP at 1.9717. However, if the spot rate ever trades at or below the barrier of 1.8875, company X has to buy the GBP at the worst-case rate of 1.9850.

Table 1 and Figure 1 demonstrate possible scenarios with assumed spot rates after six months.

4 Currency Forward Contracts

Table 1 Forward plus scenario analysis

Spot rate in six months time	Forward plus buying rate		Market forward rate
	Barrier never reached	Barrier reached	
2.0200	1.9850	1.9850	1.9717
2.0100	1.9850	1.9850	1.9717
2.0000	1.9850	1.9850	1.9717
1.9900	1.9850	1.9850	1.9717
1.9850	1.9850	1.9850	1.9717
1.9750	1.9750	1.9850	1.9717
1.9700	1.9700	1.9850	1.9717
1.9650	1.9650	1.9850	1.9717
1.9600	1.9600	1.9850	1.9717
1.9550	1.9550	1.9850	1.9717
1.9500	1.9500	1.9850	1.9717
1.9450	1.9450	1.9850	1.9717
1.9400	1.9400	1.9850	1.9717
1.9350	1.9350	1.9850	1.9717
1.9300	1.9300	1.9850	1.9717
1.9250	1.9250	1.9850	1.9717
1.9200	1.9200	1.9850	1.9717
1.9150	1.9150	1.9850	1.9717
1.9100	1.9100	1.9850	1.9717
1.9050	1.9050	1.9850	1.9717
1.9000	1.9000	1.9850	1.9717
1.8950	1.8950	1.9850	1.9717
1.8876	1.8876	1.9850	1.9717
1.8875	1.9850	1.9850	1.9717
1.8800	1.9850	1.9850	1.9717
1.8750	1.9850	1.9850	1.9717
1.8700	1.9850	1.9850	1.9717

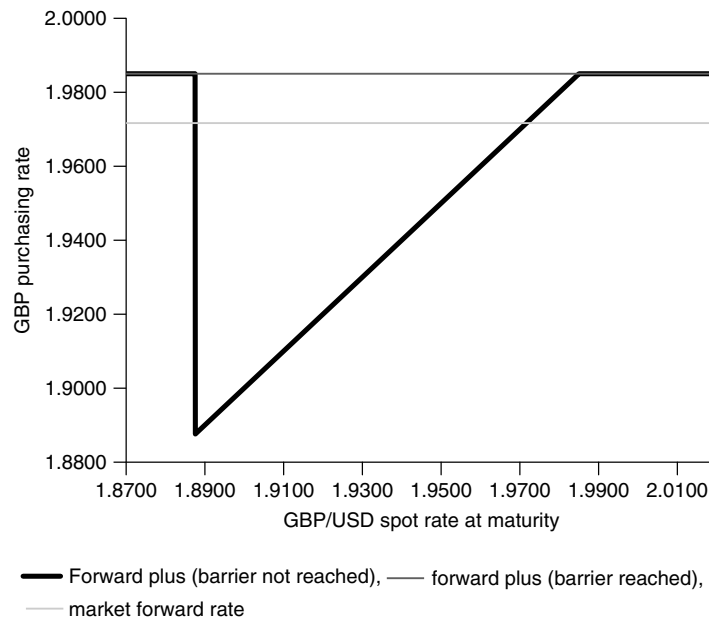


Figure 1 Forward plus scenario analysis

As Figure 1 demonstrates, the forward plus outperforms the market forward rate, if the barrier is never reached and the GBP/USD spot rate at maturity is below 1.9717.

If we set the worst-case scenario even higher than 1.9850, we can set the barrier further down. Taking advantage of this flexibility, each company entering into a forward plus can create a product that suits its risk appetite.

Range Forward. The following example uses another market view to try to outperform the forward rate. In this case, we expect the underlying currency pair to trade within a predefined range during the life of the contract.

Like with the forward plus (and with nearly all other structured forwards), the worst-case hedge rate is less favorable than the prevailing market forward rate. The generated excess cash is spent on an option

that pays out if the range holds. The payout of the option is then used to improve the worst-case rate.

Here is an example: we calculated the market forward rate to purchase GBP against USD in six months time to be 1.9717. A range forward could have a worst-case buying rate of 1.9850. This rate is 0.0133 worse than the market forward rate. In compensation for accepting this hedge rate, company X can buy GBP at 1.8850 (0.0867 better than the forward rate), if the GBP/USD exchange rate remains within the 2.0700–1.9400 range during the entire six-month period. If at any time during the life of the contract, the underlying currency pair trades outside the range, company X has to buy the GBP at the worst-case rate of 1.9850.

Table 2 and Figure 2 demonstrate possible scenarios with assumed spot rates after six months.

As Figure 2 demonstrates, the range forward outperforms the market forward rate, if the range holds, even if spot rate closes above the forward rate.

Table 2 Range forward scenario analysis

Spot rate in six months time	Range forward buying rate		Market forward rate
	Barriers never reached	Barrier reached	
2.1000	1.9850	1.9850	1.9717
2.0700	1.9850	1.9850	1.9717
2.0699	1.8850	1.9850	1.9717
2.0500	1.8850	1.9850	1.9717
2.0300	1.8850	1.9850	1.9717
2.0250	1.8850	1.9850	1.9717
2.0200	1.8850	1.9850	1.9717
2.0150	1.8850	1.9850	1.9717
2.0100	1.8850	1.9850	1.9717
2.0050	1.8850	1.9850	1.9717
2.0000	1.8850	1.9850	1.9717
1.9950	1.8850	1.9850	1.9717
1.9900	1.8850	1.9850	1.9717
1.9850	1.8850	1.9850	1.9717
1.9800	1.8850	1.9850	1.9717
1.9750	1.8850	1.9850	1.9717
1.9700	1.8850	1.9850	1.9717
1.9650	1.8850	1.9850	1.9717
1.9600	1.8850	1.9850	1.9717
1.9550	1.8850	1.9850	1.9717
1.9500	1.8850	1.9850	1.9717
1.9450	1.8850	1.9850	1.9717
1.9401	1.8850	1.9850	1.9717
1.9400	1.9850	1.9850	1.9717
1.9300	1.9850	1.9850	1.9717
1.9250	1.9850	1.9850	1.9717
1.9200	1.9850	1.9850	1.9717

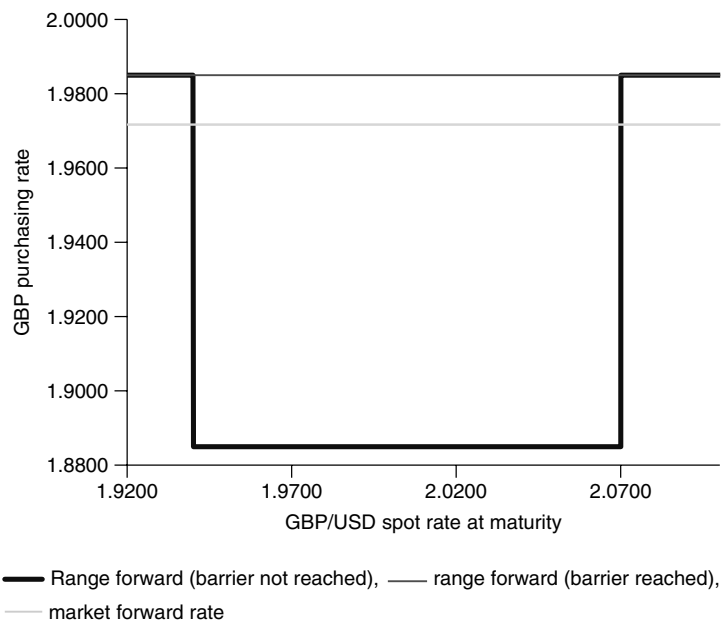


Figure 2 Range forward scenario analysis

If we set the worst-case scenario even higher than 1.9850, we can widen the range or improve the best-case buying rate. Taking advantage of this flexibility, each company entering into a range forward can create a product that suits its risk appetite.

References

- [1] Wystup, U. (2006). *FX Options and Structured Products*, John Wiley & Sons.
- [2] Weithers, T. (2006). *A Practical Guide to the FX Markets*, John Wiley & Sons.
- [3] Villanueva, O.M. (2007). Spot-forward cointegration, structural breaks and FX market unbiasedness *Journal of International Financial Markets Institutions & Money* **17**, 58–78.

- [4] Chisholm, A.M. (2004). *Derivatives Demystified: A Step-by-Step Guide to Forwards, Futures, Swaps and Options*, John Wiley & Sons.

Related Articles

Barrier Options; Forwards and Futures; Pricing Formulae for Foreign Exchange Options.

TAMÁS KORCHMÁROS

Pricing Formulae for Foreign Exchange Options

The foreign exchange options market is highly competitive, even for products beyond vanilla call and put options. This means that pricing and risk management systems always need to have the fastest possible method to compute values and sensitivities for all the products in the book. Only then can a trader or risk manager know the current position and risk of his book. The ideal solution is to use pricing formulae in closed form. However, this is often only possible in the **Black–Scholes model**.

General Model Assumptions and Abbreviations

Throughout this article, we denote the current value of the spot S_t by x and use the abbreviations listed in Table 1.

The pricing follows the usual procedures of **Arbitrage pricing theory** and the **Fundamental theorem of asset pricing**. In a foreign exchange market, this means that we model the underlying exchange rate by a *geometric Brownian motion*

$$dS_t = (r_d - r_f)S_t dt + \sigma S_t dW_t \quad (1)$$

where r_d denotes the domestic interest rate, σ the volatility, and W_t the standard Brownian motion; see **Foreign Exchange Symmetries** for details. Most importantly, we note that there is a foreign interest rate r_f . As in **Option Pricing: General Principles**, one can compute closed-form solutions for many options types with payoff $F(S_T)$ at maturity T directly via

$$\begin{aligned} v(t, x) &= e^{-r_d T} \mathbb{E}[F(S_T) | S_t = x] \\ &= e^{-r_d T} \mathbb{E} \left[F \left(x e^{(r_d - r_f - \frac{1}{2}\sigma^2)\tau + \sigma\sqrt{\tau}Z} \right) \right] \end{aligned} \quad (2)$$

where $v(t, x)$ denotes the value of the derivative with payoff F at time t if the spot is at x . The random variable Z represents the continuous returns, which are modeled as standard normal in the **Black–Scholes model**. In this model, we can proceed as

Table 1 Abbreviations used for the pricing formulae of FX options

$\tau \triangleq T - t$	$\theta_{\pm} \triangleq \frac{r_d - r_f}{\sigma} \pm \frac{\sigma}{2}$
$D_d \triangleq e^{-r_d \tau}$	$d_{\pm} \triangleq \frac{\ln\left(\frac{x}{K}\right) + \sigma\theta_{\pm}\tau}{\sigma\sqrt{\tau}}$
$D_f \triangleq e^{-r_f \tau}$	$x_{\pm} \triangleq \frac{\ln\left(\frac{x}{B}\right) + \sigma\theta_{\pm}\tau}{\sigma\sqrt{\tau}}$
$n(t) \triangleq \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$	$z_{\pm} \triangleq \frac{\ln\left(\frac{B^2}{xK}\right) + \sigma\theta_{\pm}\tau}{\sigma\sqrt{\tau}}$
$\mathcal{N}(x) \triangleq \int_{-\infty}^x n(t) dt$	$y_{\pm} \triangleq \frac{\ln\left(\frac{B}{x}\right) + \sigma\theta_{\pm}\tau}{\sigma\sqrt{\tau}}$
$\phi = +1$ for call options	$\phi = -1$ for put options
t : current time	T : maturity time
K : strike	B, L, H : barriers

$$\begin{aligned} v(t, x) &= e^{-r_d \tau} \int_{-\infty}^{+\infty} F \left(x e^{(r_d - r_f - \frac{1}{2}\sigma^2)\tau + \sigma\sqrt{\tau}z} \right) n(z) dz \\ &= D_d \int_{-\infty}^{+\infty} F \left(x e^{\sigma\theta_{-}\tau + \sigma\sqrt{\tau}z} \right) n(z) dz \end{aligned} \quad (3)$$

The rest is working out the integration. In other models, one would replace the normal density by another density function such as a t -density. However, in many other models densities are not explicitly known, or even if they are, the integration becomes cumbersome.

For the resulting pricing formulae, there are many sources, for example, [7, 11, 17]. Many general books on **Option Pricing** also contain formulae in a context outside the foreign exchange, for example, [8, 18]. Obviously, we cannot cover all possible formulae in this section. We give an overview of several relevant examples and refer to **Foreign Exchange Basket Options; Margrabe Formula; Quanto Options** for more. FX vanilla options are covered in **Foreign Exchange Symmetries**.

Barrier Options

We consider the payoff for single-barrier knock-out options

$$\begin{aligned} & [\phi(S_T - K)]^+ \mathbb{I}_{\{\eta S_t > \eta B, 0 \leq t \leq T\}} \\ &= [\phi(S_T - K)]^+ \mathbb{I}_{\{\min_{t \in [0, T]}(\eta S_t) > \eta B\}} \end{aligned} \quad (4)$$

where the binary variable η takes the value +1 if the barrier B is approached from above (down-and-out) and -1 if the barrier is approached from below (up-and-out).

To price knock-in options paying

$$[\phi(S_T - K)]^+ \mathbb{I}_{\{\min_{t \in [0, T]}(\eta S_t) \leq \eta B\}} \quad (5)$$

we use the fact that

$$\text{knock-in} + \text{knock-out} = \text{vanilla} \quad (6)$$

Computing the value of a barrier option in the Black–Scholes model boils down to knowing the joint density $f(x, y)$ for a Brownian motion with drift and its running extremum ($\eta = +1$ for a minimum and $\eta = -1$ for a maximum),

$$\left(W(T) + \theta_- T, \eta \min_{0 \leq t \leq T} [W(t) + \theta_- t] \right) \quad (7)$$

which is derived, for example, in [15], and can be written as

$$\begin{aligned} & f(x, y) \\ &= -\eta e^{\theta_- x - \frac{1}{2}\theta_-^2 T} \frac{2(2y - x)}{T\sqrt{2\pi T}} \exp \left\{ -\frac{(2y - x)^2}{2T} \right\}, \quad (8) \\ & \quad \eta y \leq \min(0, \eta x) \end{aligned}$$

Using the density (8), the value of a barrier option can be written as the following integral

$$\begin{aligned} & \text{barrier}(S_0, \sigma, r_d, r_f, K, B, T) \\ &= e^{-r_d T} \mathbb{E} \left[[\phi(S_T - K)]^+ \mathbb{I}_{\{\eta S_t > \eta B, 0 \leq t \leq T\}} \right] \quad (9) \\ &= e^{-r_d T} \int_{x=-\infty}^{x=+\infty} \int_{\eta y \leq \min(0, \eta x)} [\phi(S_0 e^{\sigma x} - K)]^+ \\ & \quad \times \mathbb{I}_{\left\{ \eta y > \eta \frac{1}{\sigma} \ln \frac{B}{S_0} \right\}} f(x, y) dy dx \quad (10) \end{aligned}$$

Further details on how to evaluate this integral can be found in [15]. It results in four terms. We provide the four terms and summarize, in Table 2, how they are used to find the value function (see also [13] or [14]).

$$A_1 = \phi x D_f \mathcal{N}(\phi d_+) - \phi K D_d \mathcal{N}(\phi d_-) \quad (11)$$

$$A_2 = \phi x D_f \mathcal{N}(\phi x_+) - \phi K D_d \mathcal{N}(\phi x_-) \quad (12)$$

$$\begin{aligned} A_3 &= \phi \left(\frac{B}{x} \right)^{\frac{2\theta_-}{\sigma}} \\ & \quad \times \left[x D_f \left(\frac{B}{x} \right)^2 \mathcal{N}(\eta z_+) - K D_d \mathcal{N}(\eta z_-) \right] \quad (13) \end{aligned}$$

$$\begin{aligned} A_4 &= \phi \left(\frac{B}{x} \right)^{\frac{2\theta_-}{\sigma}} \\ & \quad \times \left[x D_f \left(\frac{B}{x} \right)^2 \mathcal{N}(\eta y_+) - K D_d \mathcal{N}(\eta y_-) \right] \quad (14) \end{aligned}$$

Table 2 The summands for the value of single barrier options

Option type	ϕ	η	In/Out	Reverse	Combination
Standard up-and-in call	+1	-1	in	$K > B$	A_1
Reverse up-and-in call	+1	-1	in	$K \leq B$	$A_2 - A_3 + A_4$
Reverse up-and-in put	-1	-1	in	$K > B$	$A_1 - A_2 + A_4$
Standard up-and-in put	-1	-1	in	$K \leq B$	A_3
Standard down-and-in call	+1	+1	in	$K > B$	A_3
Reverse down-and-in call	+1	+1	in	$K \leq B$	$A_1 - A_2 + A_4$
Reverse down-and-in put	-1	+1	in	$K > B$	$A_2 - A_3 + A_4$
Standard down-and-in put	-1	+1	in	$K \leq B$	A_1
Standard up-and-out call	+1	-1	out	$K > B$	0
Reverse up-and-out call	+1	-1	out	$K \leq B$	$A_1 - A_2 + A_3 - A_4$
Reverse up-and-out put	-1	-1	out	$K > B$	$A_2 - A_4$
Standard up-and-out put	-1	-1	out	$K \leq B$	$A_1 - A_3$
Standard down-and-out call	+1	+1	out	$K > B$	$A_1 - A_3$
Reverse down-and-out call	+1	+1	out	$K \leq B$	$A_2 - A_4$
Reverse down-and-out put	-1	+1	out	$K > B$	$A_1 - A_2 + A_3 - A_4$
Standard down-and-out put	-1	+1	out	$K \leq B$	0

Digital and Touch Options

Digital Options

Digital options have a payoff

$$v(T, S_T) = \mathbb{I}_{\{\phi S_T \geq \phi K\}} \text{ domestic paying} \quad (15)$$

$$w(T, S_T) = S_T \mathbb{I}_{\{\phi S_T \geq \phi K\}} \text{ foreign paying} \quad (16)$$

In the domestic paying case, the payment of the fixed amount is in domestic currency, whereas in the foreign paying case the payment is in foreign currency. We obtain for the value functions

$$v(t, x) = D_d \mathcal{N}(\phi d_-) \quad (17)$$

$$w(t, x) = x D_f \mathcal{N}(\phi d_+) \quad (18)$$

of the digital options paying one unit of domestic and paying one unit of foreign currency, respectively.

One-touch Options

The payoff of a one-touch is given by

$$R \mathbb{I}_{\{\tau_B \leq T\}} \quad (19)$$

$$\tau_B \triangleq \inf\{t \geq 0 : \eta S_t \leq \eta B\} \quad (20)$$

This type of option pays a domestic cash amount R if a barrier B is hit any time before the expiration time. We use the binary variable η to describe whether B is a lower barrier ($\eta = 1$) or an upper barrier ($\eta = -1$). The stopping time τ_B is called the *first hitting time*. In FX markets, an option with this payoff is usually called a *one-touch (option)*, *one-touch-digital*, or *hit option*. The modified payoff of a *no-touch (option)*, $R \mathbb{I}_{\{\tau_B \geq T\}}$ describes a rebate, which is paid if a knock-in-option is not knocked-in by the time it expires and can be valued similarly by exploiting the identity

$$R \mathbb{I}_{\{\tau_B \leq T\}} + R \mathbb{I}_{\{\tau_B > T\}} = R \quad (21)$$

Furthermore, we distinguish the time at which the rebate is paid and let

$$\omega = 0, \text{ if the rebate is paid at first hitting time } \tau_B \quad (22)$$

$$\omega = 1, \text{ if the rebate is paid at maturity time } T \quad (23)$$

where the former is also called *instant one-touch* and the latter is the default in FX options markets. It is important to mention that the payoff is one unit of the domestic currency. For a payment in the foreign currency EUR, one needs to exchange r_d and r_f , replace x and B by their reciprocal values, and change the sign of η ; see **Foreign Exchange Symmetries**.

For the one-touch, we use the abbreviations

$$\begin{aligned} \vartheta_- &\triangleq \sqrt{\theta_-^2 + 2(1-\omega)r_d} \quad \text{and} \\ e_{\pm} &\triangleq \frac{\pm \ln \frac{x}{B} - \sigma \vartheta_- \tau}{\sigma \sqrt{\tau}} \end{aligned} \quad (24)$$

The theoretical value of the one-touch turns out to be

$$\begin{aligned} v(t, x) &= R e^{-\omega r_d \tau} \\ &\times \left[\left(\frac{B}{x} \right)^{\frac{\theta_- + \vartheta_-}{\sigma}} \mathcal{N}(-\eta e_+) + \left(\frac{B}{x} \right)^{\frac{\theta_- - \vartheta_-}{\sigma}} \mathcal{N}(\eta e_-) \right] \end{aligned} \quad (25)$$

Note that $\vartheta_- = |\theta_-|$ for rebates paid-at-end ($\omega = 1$).

The risk-neutral probability of knocking out is given by

$$\mathbb{P}[\tau_B \leq T] = \mathbb{E}[\mathbb{I}_{\{\tau_B \leq T\}}] = \frac{1}{R} e^{r_d T} v(0, S_0) \quad (26)$$

Properties of the First Hitting Time τ_B . As derived, for example, in [15], the first hitting time

$$\tilde{\tau} \triangleq \inf\{t \geq 0 : \theta t + W(t) = x\} \quad (27)$$

of a Brownian motion with drift θ and hit level $x > 0$ has the density

$$\begin{aligned} \mathbb{P}[\tilde{\tau} \in dt] \\ = \frac{x}{t \sqrt{2\pi t}} \exp \left\{ -\frac{(x - \theta t)^2}{2t} \right\} dt, \quad t > 0 \end{aligned} \quad (28)$$

the cumulative distribution function

$$IP[\tilde{\tau} \leq t] = \mathcal{N}\left(\frac{\theta t - x}{\sqrt{t}}\right) + e^{2\theta x} \mathcal{N}\left(\frac{-\theta t - x}{\sqrt{t}}\right), \quad t > 0 \quad (29)$$

the Laplace transform

$$IEe^{-\alpha \tilde{\tau}} = \exp\left\{x\theta - x\sqrt{2\alpha + \theta^2}\right\}, \quad \alpha > 0, \quad x > 0 \quad (30)$$

and the property

$$IP[\tilde{\tau} < \infty] = \begin{cases} 1 & \text{if } \theta \geq 0 \\ e^{2\theta x} & \text{if } \theta < 0 \end{cases} \quad (31)$$

For upper barriers $B > S_0$, we can now rewrite the first passage time τ_B as

$$\begin{aligned} \tau_B &= \inf\{t \geq 0 : S_t = B\} \\ &= \inf\left\{t \geq 0 : W_t + \theta_- t = \frac{1}{\sigma} \ln\left(\frac{B}{S_0}\right)\right\} \end{aligned} \quad (32)$$

The density of τ_B is hence

$$\begin{aligned} IP[\tilde{\tau}_B \in dt] &= \frac{\frac{1}{\sigma} \ln\left(\frac{B}{S_0}\right)}{t\sqrt{2\pi t}} \\ &\times \exp\left\{-\frac{\left(\frac{1}{\sigma} \ln\left(\frac{B}{S_0}\right) - \theta_- t\right)^2}{2t}\right\}, \quad t > 0 \end{aligned} \quad (33)$$

Derivation of the value function. Using the density (33), the value of the paid-at-end ($\omega = 1$) upper rebate ($\eta = -1$) option can be written as the following integral:

$$\begin{aligned} v(T, S_0) &= Re^{-r_d T} IE\left[\mathbb{I}_{\{\tau_B \leq T\}}\right] \\ &= Re^{-r_d T} \int_0^T \frac{\frac{1}{\sigma} \ln\left(\frac{B}{S_0}\right)}{t\sqrt{2\pi t}} \\ &\times \exp\left\{-\frac{\left(\frac{1}{\sigma} \ln\left(\frac{B}{S_0}\right) - \theta_- t\right)^2}{2t}\right\} dt \end{aligned} \quad (34)$$

To evaluate this integral, we introduce the notation

$$e_{\pm}(t) \triangleq \frac{\pm \ln \frac{S_0}{B} - \sigma \theta_- t}{\sigma \sqrt{t}} \quad (35)$$

and list the properties

$$e_-(t) - e_+(t) = \frac{2}{\sqrt{t}} \frac{1}{\sigma} \ln\left(\frac{B}{S_0}\right) \quad (36)$$

$$n(e_+(t)) = \left(\frac{B}{S_0}\right)^{-\frac{2\theta_-}{\sigma}} n(e_-(t)) \quad (37)$$

$$\frac{\partial e_{\pm}(t)}{\partial t} = \frac{e_{\mp}(t)}{2t} \quad (38)$$

We evaluate the integral in equation (34) by rewriting the integrand in such a way that the coefficients of the exponentials are the inner derivatives of the exponentials using properties (36)–(38).

$$\begin{aligned} &\int_0^T \frac{\frac{1}{\sigma} \ln\left(\frac{B}{S_0}\right)}{t\sqrt{2\pi t}} \exp\left\{-\frac{\left(\frac{1}{\sigma} \ln\left(\frac{B}{S_0}\right) - \theta_- t\right)^2}{2t}\right\} dt \\ &= \frac{1}{\sigma} \ln\left(\frac{B}{S_0}\right) \int_0^T \frac{1}{t^{(3/2)}} n(e_-(t)) dt \\ &= \int_0^T \frac{1}{2t} n(e_-(t)) [e_-(t) - e_+(t)] dt \\ &= - \int_0^T n(e_-(t)) \frac{e_+(t)}{2t} + \left(\frac{B}{S_0}\right)^{\frac{2\theta_-}{\sigma}} n(e_+(t)) \frac{e_-(t)}{2t} dt \\ &= \left(\frac{B}{S_0}\right)^{\frac{2\theta_-}{\sigma}} \mathcal{N}(e_+(T)) + \mathcal{N}(-e_-(T)) \end{aligned} \quad (39)$$

The computation for lower barriers ($\eta = 1$) is similar.

Double-no-touch Options

A double-no-touch with payoff function

$$\mathbb{I}_{\{L < \min_{t \in [0, T]} S_t \leq \max_{t \in [0, T]} S_t < H\}} \quad (40)$$

pays one unit of domestic currency at maturity T , if the spot *never* touches any of the two barriers, where the lower barrier is denoted by L , the higher

barrier by H . A *double-one-touch* pays one unit of domestic currency at maturity, if the spot either touches or crosses any of the lower or higher barrier at least once between inception of the trade and maturity. This means that a portfolio of a double-one-touch and a double-no-touch is equivalent to a certain payment of one unit of domestic currency at maturity.

To compute the value, let us introduce the stopping time

$$\tau_{L,H} \triangleq \min \{ \inf \{ t \in [0, T] | S_t = L \text{ or } S_t = H \}, T \} \quad (41)$$

and the notation

$$\tilde{h} \triangleq \frac{1}{\sigma} \ln \frac{H}{S_t} \quad (42)$$

$$\tilde{l} \triangleq \frac{1}{\sigma} \ln \frac{L}{S_t} \quad (43)$$

$$\tilde{\theta}_{\pm} \triangleq \theta_{\pm} \sqrt{\tau} \quad (44)$$

$$h \triangleq \tilde{h} / \sqrt{\tau} \quad (45)$$

$$l \triangleq \tilde{l} / \sqrt{\tau} \quad (46)$$

$$\epsilon_{\pm} \triangleq \epsilon_{\pm}(j) = 2j(h-l) - \tilde{\theta}_{\pm} \quad (47)$$

$$n_T(x) \triangleq \frac{1}{\sqrt{2\pi T}} \exp \left(-\frac{x^2}{2T} \right) \quad (48)$$

At any time $t < \tau_{L,H}$, the value of the double-no-touch is

$$v(t) = \mathbb{E}^t \left[D_d \mathbb{I} \{ L < \min_{u \in [t, T]} S_u \leq \max_{u \in [t, T]} S_u < H \} \right] \quad (49)$$

and for $t \in [\tau_{L,H}, T]$,

$$v(t) = D_d \mathbb{I} \{ L < \min_{u \in [t, T]} S_u \leq \max_{u \in [t, T]} S_u < H \} \quad (50)$$

The joint distribution of the maximum and the minimum of a Brownian motion can be taken from [12] and is given by

$$\mathbb{P} \left[\tilde{l} < \min_{[0, T]} W_t \leq \max_{[0, T]} W_t < \tilde{h} \right] = \int_{\tilde{l}}^{\tilde{h}} k_T(x) dx \quad (51)$$

with

$$k_T(x) = \sum_{j=-\infty}^{\infty} \left[n_T(x + 2j(\tilde{h} - \tilde{l})) - n_T(x - 2\tilde{h} + 2j(\tilde{h} - \tilde{l})) \right] \quad (52)$$

One can use Girsanov's theorem (see **Equivalent Martingale Measures**) to deduce that the joint density of the maximum and the minimum of a Brownian motion with drift θ , $W_t^{\theta} \triangleq W_t + \theta t$, is then given by

$$k_T^{\theta}(x) \triangleq k_T(x) \exp \{ \theta x - \frac{1}{2} \theta^2 T \} \quad (53)$$

We obtain for the value of the double-no-touch at any time $t < \tau_{L,H}$

$$\begin{aligned} v(t, S_t) &= D_d \mathbb{E} \mathbb{I} \{ L < \min_{u \in [t, T]} S_u \leq \max_{u \in [t, T]} S_u < H \} \\ &= D_d \mathbb{E} \mathbb{I} \{ \tilde{l} < \min_{u \in [t, T]} W_u^{\theta_-} \leq \max_{u \in [t, T]} W_u^{\theta_-} < \tilde{h} \} \\ &= D_d \int_{\tilde{l}}^{\tilde{h}} k_{(T-t)}^{\theta_-}(x) dx \\ &= D_d \cdot \sum_{j=-\infty}^{\infty} \left[e^{-2j\tilde{\theta}_-(h-l)} \right. \\ &\quad \times \{ \mathcal{N}(h + \epsilon_-) - \mathcal{N}(l + \epsilon_-) \} \\ &\quad - e^{-2j\tilde{\theta}_-(h-l) + 2\tilde{\theta}_-h} \\ &\quad \times \{ \mathcal{N}(h - 2h + \epsilon_-) - \mathcal{N}(l - 2h + \epsilon_-) \} \left. \right] \end{aligned} \quad (54)$$

and for $t \in [\tau_{L,H}, T]$

$$v(t, S_t) = D_d \mathbb{I} \{ L < \min_{u \in [t, T]} S_u \leq \max_{u \in [t, T]} S_u < H \} \quad (55)$$

Of course, the value of the double-one-touch is given by

$$D_d - v(t, S_t) \quad (56)$$

To obtain a formula for a double-no-touch paying foreign currency, see **Foreign Exchange Symmetries**.

Lookback Options

Lookback options are path dependent. At expiration, the holder of the option can “look back” over the lifetime of the option and exercise based upon the optimal underlying value (extremum) achieved during

that period. Thus, lookback options (like **Asian options**) avoid the problem of European options that the underlying performed favorably throughout most of the option's lifetime but moves into a nonfavorable direction toward maturity. Moreover, (unlike **American Options**) **lookback options** optimize the market timing, because the investor gets, by definition, the most favorable underlying price. As summarized in Table 3, **lookback options** can be structured in two different types with the extremum representing either the strike price or the underlying value. Figure 1 shows the development of the payoff of **lookback options** depending on a sample price path. In detail, we define

$$M_T \triangleq \max_{0 \leq u \leq T} S(u) \quad \text{and} \quad m_T \triangleq \min_{0 \leq u \leq T} S(u) \quad (57)$$

Variations of lookback options include *partial lookback options*, where the monitoring period for the underlying is shorter than the lifetime of the option. Conze and Viswanathan [2] present further variations like *limited risk* and *American lookback options*.

In theory, Garman pointed out in [4], that lookback options can also add value for risk managers, because floating (fixed) strike lookback options are good means to solve the timing problem of market entries (exits) (see [9]). For instance, a minimum strike call is suitable for avoiding missing the best

Table 3 Types of lookback options. The contract parameters T and X are the time to maturity and the strike price, respectively, and S_T denotes the spot price at expiration time. Fixed strike lookback options are also called *hindsight options*

Payoff	Lookback type	Parameter used below in valuation
$M_T - S_T$	Floating strike put	$\phi = -1, \bar{\eta} = +1$
$S_T - m_T$	Floating strike call	$\phi = +1, \bar{\eta} = +1$
$(M_T - X)^+$	Fixed strike call	$\phi = +1, \bar{\eta} = -1$
$(X - m_T)^+$	Fixed strike put	$\phi = -1, \bar{\eta} = -1$

exchange rate in currency-linked security issues. However, this right is very expensive. Since one buys a guarantee for the best possible exchange rate ever, lookback options are generally too expensive and hardly ever trade. Exceptions are performance notes, where lookback and average features are mixed, for example, performance notes paying say 50% of the best of 36 monthly average gold price returns.

Valuation

We consider the example of the floating strike lookback call. Again, the value of the option is given by

$$\begin{aligned} v(0, S_0) &= \mathbb{E} [e^{-r_d T} (S_T - m_T)] \\ &= S_0 e^{-r_f T} - e^{-r_d T} \mathbb{E} [m_T] \end{aligned} \quad (58)$$

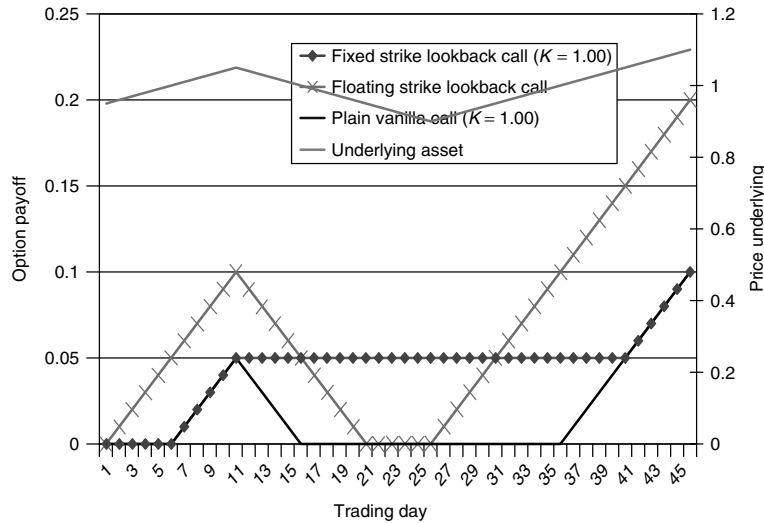


Figure 1 Payoff profile of lookback calls (sample underlying price path, $m = 20$ trading days)

In the standard Black–Scholes model (1), the value can be derived using the reflection principle and results in

$$v(t, x) = \phi \left\{ x D_f \mathcal{N}(\phi d_+) - K D_d \mathcal{N}(\phi d_-) \right. \\ \left. + \frac{1-\eta}{2} \phi D_d [\phi(R-X)]^+ \right. \\ \left. + \eta x D_d \frac{1}{h} \left[\left(\frac{x}{K} \right)^{-h} \mathcal{N}(-\eta \phi(d_+ - h\sigma\sqrt{\tau})) \right. \right. \\ \left. \left. - e^{(r_d - r_f)\tau} \mathcal{N}(-\eta \phi d_+) \right] \right\} \quad (59)$$

This value function has a removable discontinuity at $h = 0$ where it turns out to be

$$v(t, x) = \phi \left\{ x D_f \mathcal{N}(\phi d_+) - K D_d \mathcal{N}(\phi d_-) \right. \\ \left. + \frac{1-\eta}{2} \phi D_d [\phi(R-X)]^+ \right. \\ \left. + \eta x D_d \sigma \sqrt{\tau} [-d_+ \mathcal{N}(-\eta \phi d_+) \right. \\ \left. + \eta \phi n(d_+)] \right\} \quad (60)$$

The abbreviations we use are

$$h \triangleq \frac{2(r_d - r_f)}{\sigma^2} \quad (61)$$

$$R \triangleq \begin{array}{l} \text{running extremum: extremum observed} \\ \text{until valuation time} \end{array} \quad (62)$$

$$K \triangleq \begin{cases} R & \text{floating strike lookback} \\ -\phi \min(-\phi X, -\phi R) & \text{fixed strike lookback} \end{cases} \quad (63)$$

$$\eta \triangleq \begin{cases} +1 & \text{floating strike lookback} \\ -1 & \text{fixed strike lookback} \end{cases} \quad (64)$$

Note that this formula basically consists of that for a **call option** (the first two terms) plus another term. Conze and Viswanathan also show closed-form solutions for fixed strike lookback options and the variations mentioned above in [2]. Heynen and Kat develop equations for *partial fixed and floating strike*

Table 4 Sample values for lookback options. For the input data, we used spot $S_0 = 0.9800$, $r_d = 3\%$, $r_f = 6\%$, $\sigma = 10\%$, $\tau = 1/12$, running min $R = 0.9500$, running max $R = 0.9900$, and number of equidistant fixings $m = 22$

Payoff sampled	Discretely sampled Equations (67) and (68)	Continuously Equations (59) or (60)
$M_T - S_T$	0.0231	0.0255
$S_T - m_T$	0.0310	0.0320
$(M_T - 0.99)^+$	0.0107	0.0131
$(0.97 - m_T)^+$	0.0235	0.0246

lookback options in [10]. We list some sample results in Table 4.

Discrete Sampling

In practice, one cannot take the average over a continuum of exchange rates. The standard is to specify a *fixing calendar* and take only a finite number of fixings into account. Suppose there are m equidistant sample points left until expiration at which we evaluate the extremum. In this case, the value function v_m can be determined by an approximation described by Broadie *et al.* [1]. We set

$$\beta \triangleq -\zeta(1/2)/\sqrt{2\pi} \\ = 0.5826 \quad (\zeta \text{ being Riemann's } \zeta\text{-function}) \quad (65)$$

$$\alpha \triangleq e^{\phi\beta\sigma\sqrt{\tau/m}} \quad (66)$$

and obtain for fixed strike lookback options

$$v_m(t, x, r_d, r_f, \sigma, R, X, \phi, \eta) \\ = v(t, x, r_d, r_f, \sigma, \alpha R, \alpha X, \phi, \eta)/\alpha \quad (67)$$

and for floating strike lookback options

$$v_m(t, x, r_d, r_f, \sigma, R, X, \phi, \eta) \\ = av(t, x, r_d, r_f, \sigma, R/\alpha, X, \phi, \eta) - \phi(\alpha - 1)x D_f \quad (68)$$

Forward Start Options

Product Definition

A forward start vanilla option is just like a vanilla option, except that the strike is fixed on some future date $t_f \in (0, T)$, specified in the contract. The strike is fixed as αS_{t_f} , where $\alpha > 0$ is some contractually defined factor (very common one) and S_{t_f} is the spot at time t_f . It pays off

$$[\phi(S_T - \alpha S_{t_f})]^+ \quad (69)$$

The Value of Forward Start Options

Using the abbreviations

$$d_{\pm}(x) \triangleq \frac{\ln \frac{x}{K} + \sigma \theta_{\pm}(T - t_f)}{\sigma \sqrt{T - t_f}} \quad (70)$$

$$d_{\pm}^{\alpha} \triangleq \frac{-\ln \alpha + \sigma \theta_{\pm}(T - t_f)}{\sigma \sqrt{T - t_f}} \quad (71)$$

we recall the value of a vanilla option with strike K at time t_f as

$$\begin{aligned} & \text{vanilla}(t_f, x; K, T, \phi) \\ &= \phi [x e^{-r_f(T-t_f)} \mathcal{N}(\phi d_+(x)) - K e^{-r_d(T-t_f)} \mathcal{N}(\phi d_-(x))] \end{aligned} \quad (72)$$

For the value of a forward start vanilla option, we obtain

$$\begin{aligned} & v(0, S_0) \\ &= e^{-r_d t_f} \mathbb{E}[\text{vanilla}(t_f, S_{t_f}; K = \alpha S_{t_f}, T, \phi)] \\ &= \phi S_0 [e^{-r_f T} \mathcal{N}(\phi d_+^{\alpha}) - \alpha e^{(r_d - r_f)t_f} e^{-r_d T} \mathcal{N}(\phi d_-^{\alpha})] \end{aligned} \quad (73)$$

Noticeably, the value computation is easy here, because the strike K is set as a *multiple* of the future spot. If we were to choose to set the strike as a constant *difference* of the future spot, the integration would not work in closed form, and we would have to use numerical integration.

Table 5 Value of a forward start vanilla in USD on EUR/USD—spot of 0.9000, $\alpha = 99\%$, $\sigma = 12\%$, $r_d = 2\%$, $r_f = 3\%$, maturity $T = 186$ days, and strike set at $t_f = 90$ days

	Call	Put
Value	0.0251	0.0185

Example

We consider an example in Table 5.

Compound and Instalment Options

An instalment call option allows the holder to pay the premium of the **call option** in instalments spread over time. A first payment is made at the inception of the trade. On the following payment days, the holder of the instalment call can decide to prolong the contract, in which case, he has to pay the second instalment of the premium, or to terminate the contract by simply not paying any more. After the last instalment payment, the contract turns into a plain vanilla call.

Valuation in the Black–Scholes Model

The intention of this section is to obtain a closed-form formula for the n -variate instalment option in the **Black–Scholes model**. For the cases $n = 1$ and $n = 2$, the Black–Scholes formula and Geske’s compound option formula (see [5]) are well-known special cases.

Let $t_0 = 0$ be the instalment option inception date and $t_1, t_2, \dots, t_n = T$ a schedule of decision dates in the contract on which the option holder has to pay the premiums $k_1, k_2, \dots, k_{n-1}$ to keep the option alive. To compute the price of the instalment option, which is the upfront payment V_0 at t_0 to enter into the contract, we begin with the option payoff at maturity T

$$V_n(s) \triangleq [\phi_n(s - k_n)]^+ \triangleq \max[\phi_n(s - k_n), 0] \quad (74)$$

where $s = S_T$ is the price of the underlying asset at T and as usual $\phi_n = +1$ for a call option, $\phi_n = -1$ for a put option.

At time t_i , the option holder can either terminate the contract or pay k_i to continue. This means that the instalment option can be viewed as an option with strike k_1 on an option with strike $k_2 \cdots$ on an option with strike k_n . Therefore, by the **risk-neutral pricing** theory, the holding value is

$$e^{-r_d(t_{i+1}-t_i)} \mathbb{E}[V_{i+1}(S_{t_{i+1}}) | S_{t_i} = s],$$

for $i = 0, \dots, n-1$ (75)

where

$$V_i(s) = \begin{cases} [e^{-r_d(t_{i+1}-t_i)} \mathbb{E}[V_{i+1}(S_{t_{i+1}}) | S_{t_i} = s] - k_i]^+ & \text{for } i = 1, \dots, n-1 \\ V_n(s) & \text{for } i = n \end{cases}$$

(76)

Then the unique arbitrage-free time-zero value is

$$P \triangleq V_0(s) = e^{-r_d(t_1-t_0)} \mathbb{E}[V_1(S_{t_1}) | S_{t_0} = s] \quad (77)$$

Figure 2 illustrates this context.

One way of pricing this instalment option is to evaluate the nested expectations through multiple numerical integration of the payoff functions *via* backward iteration. Alternatively, one can derive a solution in closed form in terms of the n -variate cumulative normal, which is described in the following.

The Curnow and Dunnett Integral Reduction Technique

Denote the n -dimensional multivariate normal integral with upper limits $h_1, \dots, h_n$ and correlation

matrix $R_n \triangleq (\rho_{ij})_{i,j=1,\dots,n}$ by $\mathcal{N}_n(h_1, \dots, h_n; R_n)$. Let the correlation matrix be nonsingular and $\rho_{11} = 1$.

Under these conditions, Curnow and Dunnett [3] derived the following reduction formula for multivariate normal integrals:

$$\begin{aligned} \mathcal{N}_n(h_1, \dots, h_n; R_n) &= \int_{-\infty}^{h_1} \mathcal{N}_{n-1}\left(\frac{h_2 - \rho_{21}y}{(1 - \rho_{21}^2)^{1/2}}, \dots, \frac{h_n - \rho_{n1}y}{(1 - \rho_{n1}^2)^{1/2}}; R_{n-1}^*\right) n(y) dy, \\ R_{n-1}^* &\triangleq (\rho_{ij}^*)_{i,j=2,\dots,n}, \\ \rho_{ij}^* &\triangleq \frac{\rho_{ij} - \rho_{i1}\rho_{j1}}{(1 - \rho_{i1}^2)^{1/2}(1 - \rho_{j1}^2)^{1/2}} \end{aligned} \quad (78)$$

For example, to go from dimension 1 to dimension 2, this takes the form

$$\begin{aligned} \int_{-\infty}^x \mathcal{N}(az + B)n(z) dz &= \mathcal{N}_2\left(x, \frac{B}{\sqrt{1+a^2}}; \frac{-a}{\sqrt{1+a^2}}\right) \end{aligned} \quad (79)$$

or more generally,

$$\begin{aligned} \int_{-\infty}^x e^{Az} \mathcal{N}(az + B)n(z) dz &= e^{\frac{A^2}{2}} \mathcal{N}_2\left(x - A, \frac{aA + B}{\sqrt{1+a^2}}; \frac{-a}{\sqrt{1+a^2}}\right) \end{aligned} \quad (80)$$

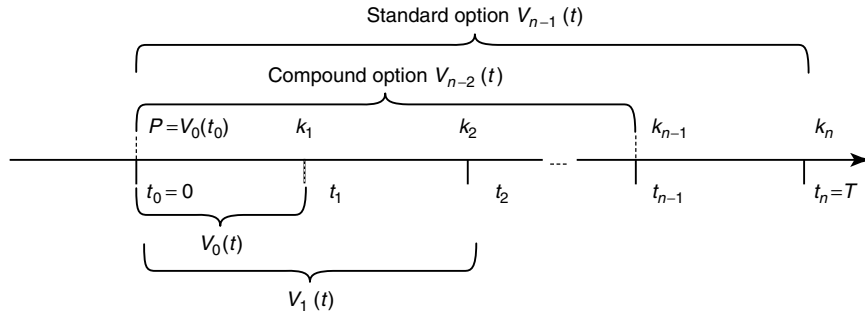


Figure 2 Lifetime of the options V_i

A Closed-form Solution for the Value of an Instalment Option

Heuristically, the formula which is given in the Theorem 1 has the structure of the **Black-Scholes formula** in higher dimensions, namely, $S_0 \mathcal{N}_n(\cdot) - k_n \mathcal{N}_n(\cdot)$ minus the later premium payments $k_i \mathcal{N}_i(\cdot)$

Theorem 1 Let $\mathbf{k} = (k_1, \dots, k_n)$ be the strike price vector, $\mathbf{t} = (t_1, \dots, t_n)$ the vector of the exercise dates of an n -variate instalment option and $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)$ the vector of the put/call-indicators of these n options.

The value function of an n -variate instalment option is given by

$$\begin{aligned}
 V_n(S_0, \mathbf{k}, \mathbf{t}, \boldsymbol{\phi}) &= e^{-r_f t_n} S_0 \phi_1 \cdots \phi_n \times \mathcal{N}_n \left[\phi_1 \frac{\ln \frac{S_0}{S_1^*} + \sigma \theta_+ t_1}{\sigma \sqrt{t_1}}, \phi_2 \frac{\ln \frac{S_0}{S_2^*} + \sigma \theta_+ t_2}{\sigma \sqrt{t_2}}, \dots, \phi_n \frac{\ln \frac{S_0}{S_n^*} + \sigma \theta_+ t_n}{\sigma \sqrt{t_n}}; R_n \right] \\
 &- e^{-r_d t_n} k_n \phi_1 \cdots \phi_n \times \mathcal{N}_n \left[\phi_1 \frac{\ln \frac{S_0}{S_1^*} + \sigma \theta_- t_1}{\sigma \sqrt{t_1}}, \phi_2 \frac{\ln \frac{S_0}{S_2^*} + \sigma \theta_- t_2}{\sigma \sqrt{t_2}}, \dots, \phi_n \frac{\ln \frac{S_0}{S_n^*} + \sigma \theta_- t_n}{\sigma \sqrt{t_n}}; R_n \right] \\
 &- e^{-r_d t_{n-1}} k_{n-1} \phi_1 \cdots \phi_{n-1} \times \mathcal{N}_{n-1} \left[\phi_1 \frac{\ln \frac{S_0}{S_1^*} + \sigma \theta_- t_1}{\sigma \sqrt{t_1}}, \phi_2 \frac{\ln \frac{S_0}{S_2^*} + \sigma \theta_- t_2}{\sigma \sqrt{t_2}}, \dots, \phi_{n-1} \frac{\ln \frac{S_0}{S_{n-1}^*} + \sigma \theta_- t_{n-1}}{\sigma \sqrt{t_{n-1}}}; R_{n-1} \right] \\
 &\vdots \\
 &- e^{-r_d t_2} k_2 \phi_1 \phi_2 \mathcal{N}_2 \left[\phi_1 \frac{\ln \frac{S_0}{S_1^*} + \sigma \theta_- t_1}{\sigma \sqrt{t_1}}, \phi_2 \frac{\ln \frac{S_0}{S_2^*} + \sigma \theta_- t_2}{\sigma \sqrt{t_2}}; \rho_{12} \right] \\
 &- e^{-r_d t_1} k_1 \phi_1 \mathcal{N} \left[\phi_1 \frac{\ln \frac{S_0}{S_1^*} + \sigma \theta_- t_1}{\sigma \sqrt{t_1}} \right]
 \end{aligned} \tag{81}$$

($i = 1, \dots, n-1$). This structure is a result of the integration of the vanilla option payoff, which is again integrated minus the next instalment, which in turn is integrated with the following instalment and so forth. By this iteration, the vanilla payoff is integrated with respect to the normal density function n times and the i th payment is integrated i times for $i = 1, \dots, n-1$.

The correlation coefficients ρ_{ij} of these normal distribution functions contained in the formula arise from the overlapping increments of the Brownian motion, which models the price process of the underlying S_t at the particular exercise dates t_i and t_j .

where S_i^* ($i = 1, \dots, n$) is to be determined as the spot price S_t for which the payoff of the corresponding i -variate instalment option ($i = 1, \dots, n$) is equal to 0, that is, $V_i(S_i^*, \mathbf{k}, \mathbf{t}, \boldsymbol{\phi}) = 0$. This has to be done numerically by a zero search.

The correlation coefficients in R_i of the i -variate normal distribution function can be expressed through the exercise dates t_i ,

$$\rho_{ij} = \sqrt{t_i/t_j} \text{ for } i, j = 1, \dots, n \text{ and } i < j \tag{82}$$

The proof is established with equation (78). Formula (81) has been independently derived by Thomassen and Wouve in [16] and Griebisch *et al.* in [6].

References

- [1] Broadie, M. Glasserman, P. & Kou, S.G. (1999). Connecting discrete and continuous path-dependent options, *Finance and Stochastics* **3**(1), 55–82.
- [2] Conze, A. & Viswanathan, R. (1991). Path dependent options: the case of lookback options, *The Journal of Finance*, **XLVI**(5), 1893–1907.
- [3] Curnow, R.N. & Dunnett, C.W. (1962). The numerical evaluation of certain multivariate normal integrals, *Annals of Mathematical Statistics* **33**, 571–579.
- [4] Garman, M. (1989). Recollection in tranquillity,, re-edited version, in *From Black Scholes to Black Holes*, Originally in Risk Publications, London, pp. 171–175.
- [5] Geske, R. (1979). The valuation of compound options, *Journal of Financial Economics* **7**, 63–81.
- [6] Griebisch, S.A. Kühn, C. & Wystup, U. (2008). Instalment options: a closed-form solution and the limiting case, in *Contribution to Mathematical Control Theory and Finance*, A. Sarychev, A. Shiryaev, M. Guerra, & M.R. Grossinho, eds, Springer, pp. 211–229.
- [7] Hakala, J. & Wystup, U. (2002). *Foreign Exchange Risk*, Risk Publications, London.
- [8] Haug, E.G. (1997). *Option Pricing Formulas*, McGraw Hill.
- [9] Heynen, R. & Kat, H. (1994). Selective memory: reducing the expense of lookback options by limiting their memory, re-edited version, in *Over the Rainbow: Developments in Exotic Options and Complex Swaps*, Risk Publications, London.
- [10] Heynen, R. & Kat, H. (1994). Crossing barriers, *Risk* **7**(6), 46–51.
- [11] Lipton, A. (2001). *Mathematical Methods for Foreign Exchange*, World Scientific, Singapore.
- [12] Revuz, D. & Yor, M. (1995). *Continuous Martingales and Brownian Motion*, 2nd Edition, Springer.
- [13] Rich, D. (1994). The mathematical foundations of barrier option pricing theory, *Advances in Futures and Options Research* **7**, 267–371.
- [14] Reiner, E. & Rubinstein, M. (1991). Breaking down the barriers, *Risk* **4**(8), 28–35.
- [15] Shreve, S.E. (2004). *Stochastic Calculus for Finance I+II*, Springer.
- [16] Thomassen, L. & van Wouwe, M. (2002). *A Sensitivity Analysis for the N-fold Compound Option*, Research Paper, Faculty of Applied Economics, University of Antwerpen.
- [17] Wystup, U. (2006). *FX Options and Structured Products*, Wiley Finance Series.
- [18] Zhang, P.G. (1998). *Exotic Options*, 2nd Edition, World Scientific, London.

Related Articles

Barrier Options; Black–Scholes Formula; Discretely Monitored Options; Foreign Exchange Markets; Foreign Exchange Options; Foreign Exchange Symmetries; Lookback Options.

ANDREAS WEBER & UWE WYSTUP

Foreign Exchange Symmetries

Motivation

The symmetries of the foreign exchange (FX) market are the key features that distinguish this market from all others. With an EUR-USD exchange rate of 1.2500 USD per EUR, there is an equivalent USD-EUR exchange rate of 0.8000 EUR per USD, which is just the reciprocal. Any model S_t for an exchange rate at time t should guarantee that $1/S_t$ is within the same model class. This is satisfied by the Black–Scholes model or local volatility models, but not for many stochastic volatility models.

The further symmetry is that both currencies pay interest, which we can assume to be continuously paid. The FX market is the only market where this really works.

In the EUR-USD market, any EUR call is equivalent to an USD put. Any premium in USD can be also paid in EUR, any delta hedge can be specified in the amount of USD to sell or, alternatively, the amount of EUR to buy.

Furthermore, if S_1 is a model for EUR-USD and S_2 is a model for USD–JPY, then $S_3 = S_1 \cdot S_2$ should be a model for EUR-JPY. Therefore, besides the reciprocals, the products of two modeling quantities should also remain within the same model class.

Finally, the smile of an FX options market is summarized by Risk Reversals and Butterflies, the skew-symmetric part and the symmetric part of a smile curve.

In no other market are symmetries so prominent and so heavily used as in FX. It is this special feature that makes it hard for many newcomers to capture the way FX options market participants think.

Geometric Brownian Motion Model for the Spot

We consider the model *geometric Brownian motion*

$$dS_t = (r_d - r_f)S_t dt + \sigma S_t dW_t \quad (1)$$

for the underlying exchange rate quoted in foreign–domestic (FOR–DOM), which means that 1 unit of the foreign currency costs FOR–DOM units of the

domestic currency. In the case of EUR-USD with a spot of 1.2000, this means that the price of 1 EUR is 1.2000 USD. The notions of *foreign* and *domestic* do not refer to the location of the trading entity, but only to this quotation convention. We denote the (continuous) foreign interest rate by r_f and the (continuous) domestic interest rate by r_d . In an equity scenario, r_f would represent a continuous dividend rate. The volatility is denoted by σ , and W_t is a standard Brownian motion.

We consider this standard model, not because it reflects the statistical properties of the exchange rate (in fact, it does not), but because it is widely used in practice and front office systems and mainly serves as a tool to communicate prices in FX options. These prices are generally quoted in terms of volatility in the sense of this model.

Vanilla Options

The payoff for a vanilla option (European put or call) is given by

$$F = [\phi(S_T - K)]^+ \quad (2)$$

where the contractual parameters are the strike K , the expiration time T and the type ϕ , a binary variable, which takes the value $+1$ in the case of a call and -1 in the case of a put. The symbol x^+ denotes the positive part of x , that is, $x^+ = \max(0, x) = 0 \vee x$.

Value

In the Black–Scholes model, the value of the payoff F at time t if the spot is at S is denoted by $v(t, S)$. The result is the *Black–Scholes formula*

$$\begin{aligned} v(S, K, T, t, \sigma, r_d, r_f, \phi) \\ = \phi e^{-r_d \tau} [f \mathcal{N}(\phi d_+) - K \mathcal{N}(\phi d_-)] \end{aligned} \quad (3)$$

We abbreviate

- S : current price of the underlying
- $\tau = T - t$: time to maturity
- $f = \mathbb{E}[S_T | S_t = S] = S e^{(r_d - r_f)\tau}$: forward price of the underlying
- $\theta_{\pm} = \frac{r_d - r_f}{\sigma} \pm \frac{\sigma}{2}$

- $d_{\pm} = \frac{\ln \frac{S}{K} + \sigma \theta_{\pm} \tau}{\sigma \sqrt{\tau}} = \frac{\ln \frac{f}{K} \pm \frac{\sigma^2}{2} \tau}{\sigma \sqrt{\tau}}$
- $n(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} = n(-t)$
- $\mathcal{N}(x) = \int_{-\infty}^x n(t) dt = 1 - \mathcal{N}(-x)$

A Note on the Forward

The forward price f is the strike that makes the time zero value of the forward contract

$$F = S_T - f \quad (4)$$

equal to zero. It follows that $f = E[S_T] = S e^{(r_d - r_f) \cdot T}$, that is, the forward price is the expected price of the underlying at time T in a risk-neutral setup (drift of the geometric Brownian motion is equal to the cost of carry $r_d - r_f$).

Greeks

Greeks are derivatives of the value function with respect to model and contract parameters. They are an important information for traders and have become standard information provided by front-office systems. More details on Greeks and the relations among Greeks are presented in [5] or [6]. We now list some of them for vanilla options:

- Spot delta. The spot delta shows how many units of FOR in the spot must be traded to hedge an option with 1 unit of FOR notional.

$$\frac{\partial v}{\partial S} = \phi e^{-r_f \tau} \mathcal{N}(\phi d_+) \quad (5)$$

- Forward delta. The forward delta shows how many units of FOR of a forward contract must be traded to hedge an option with 1 unit of FOR notional.

$$\phi \mathcal{N}(\phi d_+) = IP \left[\phi S_T \leq \phi \frac{f^2}{K} \right] \quad (6)$$

It is also equal to the risk-neutral exercise probability or in-the-money probability of the symmetric put (see Section 1.4.3).

- Gamma.

$$\frac{\partial^2 v}{\partial S^2} = e^{-r_f \tau} \frac{n(d_+)}{S \sigma \sqrt{\tau}} \quad (7)$$

- Vega.

$$\frac{\partial v}{\partial \sigma} = S e^{-r_f \tau} \sqrt{\tau} n(d_+) \quad (8)$$

- Dual delta.

$$\frac{\partial v}{\partial K} = -\phi e^{-r_d \tau} \mathcal{N}(\phi d_-) \quad (9)$$

The forward dual delta

$$\mathcal{N}(\phi d_-) = IP[\phi S_T \geq \phi K] \quad (10)$$

can also be viewed as the risk-neutral exercise probability.

- Dual gamma.

$$\frac{\partial^2 v}{\partial K^2} = e^{-r_d \tau} \frac{n(d_-)}{K \sigma \sqrt{\tau}} \quad (11)$$

Identities

Any computations with vanilla options often rely on the symmetry identities

$$\frac{\partial d_{\pm}}{\partial \sigma} = -\frac{d_{\mp}}{\sigma} \quad (12)$$

$$\frac{\partial d_{\pm}}{\partial r_d} = \frac{\sqrt{\tau}}{\sigma} \quad (13)$$

$$\frac{\partial d_{\pm}}{\partial r_f} = -\frac{\sqrt{\tau}}{\sigma} \quad (14)$$

$$S e^{-r_f \tau} n(d_+) = K e^{-r_d \tau} n(d_-) \quad (15)$$

Put–Call Parity

The put–call parity is the relationship,

$$\begin{aligned} & v(S, K, T, t, \sigma, r_d, r_f, +1) \\ & - v(S, K, T, t, \sigma, r_d, r_f, -1) \\ & = S e^{-r_f \tau} - K e^{-r_d \tau} \end{aligned} \quad (16)$$

which is just a more complicated way to write the trivial equation $x = x^+ - x^-$. Taking the strike K to be equal to the forward price f , we see that the call and put have the same value. The forward is the center of symmetry for vanilla call and put values. However, this does not imply that the *deltas* are also symmetric about the forward price.

The put–call delta parity is

$$\begin{aligned} & \frac{\partial v(S, K, T, t, \sigma, r_d, r_f, +1)}{\partial S} \\ & - \frac{\partial v(S, K, T, t, \sigma, r_d, r_f, -1)}{\partial S} = e^{-r_f \tau} \end{aligned} \quad (17)$$

In particular, we learn that the absolute value of a put delta and a call delta do not exactly add up to 1, but only to a positive number $e^{-r_f \tau}$. They add up to 1 approximately if either the time to expiration τ is short or if the foreign interest rate r_f is close to 0. For this reason, traders often prefer to work with forward deltas, because these are symmetric in the sense that a 25-delta call is a 75-delta put.

Although the choice $K = f$ produces identical values for call and put, we seek the *delta-symmetric strike* $\check{K}$, which produces absolutely identical deltas (spot, forward, or driftless). This condition implies $d_+ = 0$ and thus

$$\check{K} = f e^{\frac{\sigma^2}{2} \tau} \quad (18)$$

in which case the absolute delta is $e^{-r_f \tau}/2$. In particular, we learn that always $\check{K} > f$, that is, there cannot be a put and a call with identical values *and* deltas. This is natural as the payoffs of calls and puts are not symmetric to start with: the call has unlimited upside potential, whereas the put payoff is always bounded by the strike. Note that the strike $\check{K}$ is usually chosen as the middle strike when trading a straddle or a butterfly. Similarly the dual-delta-symmetric strike $\hat{K} = f e^{-\frac{\sigma^2}{2} \tau}$ can be derived from the condition $d_- = 0$.

Note that the delta-symmetric strike $\check{K}$ also maximizes gamma and vega of a vanilla option and is thus often considered as a center of symmetry.

Homogeneity-based Relationships

Space Homogeneity

We may wish to measure the value of the underlying in a different unit. This will obviously effect the option pricing formula as follows:

$$\begin{aligned} av(S, K, T, t, \sigma, r_d, r_f, \phi) \\ = v(aS, aK, T, t, \sigma, r_d, r_f, \phi) \text{ for all } a > 0 \end{aligned} \quad (19)$$

Differentiating both sides with respect to a and then setting $a = 1$ yields

$$v = S v_S + K v_K \quad (20)$$

Comparing the coefficients of S and K in equations (3) and (20) leads to suggestive results for the delta v_S and dual delta v_K . This *space-homogeneity* is the reason behind the simplicity of the delta formulas, whose tedious computation can be saved this way.

Time Homogeneity

We can perform a similar computation for the time-affected parameters and obtain the obvious equation

$$\begin{aligned} v(S, K, T, t, \sigma, r_d, r_f, \phi) \\ = v\left(S, K, \frac{T}{a}, \frac{t}{a}, \sqrt{a}\sigma, ar_d, ar_f, \phi\right) \\ \text{for all } a > 0 \end{aligned} \quad (21)$$

Differentiating both sides with respect to a and then setting $a = 1$ yields

$$0 = \tau v_t + \frac{1}{2} \sigma v_\sigma + r_d v_{r_d} + r_f v_{r_f} \quad (22)$$

Of course, this can also be verified by direct computation. The overall use of such equations is to generate double checking benchmarks when computing Greeks. These homogeneity methods can easily be extended to other more complex options.

Put–Call Symmetry

By *put–call symmetry*, we understand the relationship (see [1–4])

$$\begin{aligned} v(S, K, T, t, \sigma, r_d, r_f, +1) \\ = \frac{K}{f} v\left(S, \frac{f^2}{K}, T, t, \sigma, r_d, r_f, -1\right) \end{aligned} \quad (23)$$

The geometric mean of the strike of the put $\frac{f^2}{K}$ and the strike of the call K is equal to $\sqrt{\frac{f^2}{K} \cdot K} = f$, the outright forward rate. Therefore, the outright forward rate can be interpreted as a *geometric mirror* reflecting a call into a certain number of puts. Note that for at-the-money (forward) options ($K = f$) the put–call symmetry coincides with the special case of

the put–call parity where the call and the put have the same value.

Rates Symmetry

Direct computation shows that the *rates symmetry*

$$\frac{\partial v}{\partial r_d} + \frac{\partial v}{\partial r_f} = -\tau v \quad (24)$$

holds for vanilla options. This relationship, in fact, holds for all European options and a wide class of path-dependent options as shown in [6].

Foreign–Domestic Symmetry

One can directly verify the *FOR–DOM symmetry*

$$\begin{aligned} & \frac{1}{S} v(S, K, T, t, \sigma, r_d, r_f, \phi) \\ &= K v\left(\frac{1}{S}, \frac{1}{K}, T, t, \sigma, r_f, r_d, -\phi\right) \end{aligned} \quad (25)$$

This equality can be viewed as one of the faces of put–call symmetry. The reason is that the value of an option can be computed both in a domestic as well as in a foreign scenario. We consider the example of S_t modeling the exchange rate of EUR/USD. In New York, the call option $(S_T - K)^+$ costs $v(S, K, T, t, \sigma, r_{\text{usd}}, r_{\text{eur}}, 1)$ USD and hence $v(S, K, T, t, \sigma, r_{\text{usd}}, r_{\text{eur}}, 1)/S$ EUR. This EUR call option can also be viewed as an USD put option with payoff $K \left(\frac{1}{K} - \frac{1}{S_T}\right)^+$. This option costs $K v\left(\frac{1}{S}, \frac{1}{K}, T, t, \sigma, r_{\text{eur}}, r_{\text{usd}}, -1\right)$ EUR in Frankfurt, because S_t and $\frac{1}{S_t}$ have the same volatility. Of course, the New York value and the Frankfurt value must agree, which leads to equation (25). This can also be seen as a change of measure to the foreign discount bond as numeraire (see, e.g., in [7]).

Exotic Options

In FX markets, one can use many symmetry relationships for exotic options.

Digital Options

For example, let us define the payoff of digital options by

$$\text{DOM paying: } \mathbb{1}_{\{\phi S_T \geq \phi K\}} \quad (26)$$

$$\text{FOR paying: } S_T \mathbb{1}_{\{\phi S_T \geq \phi K\}} \quad (27)$$

where the contractual parameters are the strike K , the expiration time T , and the type ϕ , a binary variable, which takes the value $+1$ in the case of a call and -1 in the case of a put. Then we observe that a DOM-paying digital call in the currency pair FOR–DOM with a value of v_d units of domestic currency must be worth the same as a FOR-paying digital put in the currency pair DOM–FOR with a value of v_f units of foreign currency. And since we are looking at the same product, we conclude that $v_d = v_f \cdot S$, where S is the initial spot of FOR–DOM.

Touch Options

This key idea generalizes from the path-independent digitals to touch products. Consider the value function for one-touch in EUR-USD paying 1 USD. If we want to find the value function of a one-touch in EUR-USD paying 1 EUR, we can price the one-touch in USD-EUR paying 1 EUR using the known value function with rates r_d and r_f exchanged, volatility unchanged, using the formula for a one-touch in EUR-USD paying 1 USD. We also note that an upper one-touch in EUR-USD becomes a lower one-touch in USD-EUR. The result we get is in domestic currency, which is EUR in USD-EUR notation. To convert it into an USD price, we just multiply by the EUR-USD spot S .

Barrier Options

For a standard knockout barrier option, we let the value function be

$$v(S, r_d, r_f, \sigma, K, B, T, t, \phi, \eta) \quad (28)$$

where B denotes the barrier and the variable η takes the value $+1$ for a lower barrier and -1 for an upper barrier. With this notation at hand, we can state our FOR–DOM symmetry as

$$\begin{aligned} & v(S, r_d, r_f, \sigma, K, B, T, t, \phi, \eta) \\ &= v\left(\frac{1}{S}, r_f, r_d, \sigma, \frac{1}{K}, \frac{1}{B}, T, t, -\phi, -\eta\right) SK \end{aligned} \quad (29)$$

Note that the rates r_d and r_f have been interchanged on purpose. This implies that if we know how to price

barrier contracts with upper barriers, we can derive the formulas for lower barriers.

Quotation

Quotation of the Underlying Exchange Rate

Equation (1) is a model for the exchange rate. The quotation is a permanently confusing issue, so let us clarify this here. The exchange rate means how much of the *domestic* currency are needed to buy 1 unit of *foreign* currency. For example, if we take EUR/USD as an exchange rate, then the default quotation is EUR-USD, where USD is the domestic currency and EUR is the foreign currency. The term *domestic* is in no way related to the location of the trader or any country. It merely means the *numeraire* currency. The terms *domestic*, *numeraire*, or *base currency* are synonyms as are *foreign* and *underlying*. Commonly, we denote with the slash (/) the currency pair and with a dash (-) the quotation. The slash (/) does *not* mean a division. For instance, EUR/USD can also be quoted in either EUR-USD, which then means how many USD are needed to buy one EUR, or in USD-EUR, which then means how many EUR are needed to buy 1 USD. There are certain market standard quotations listed in Table 1.

Quotation of Option Prices

Values and prices of vanilla options may be quoted in the six ways explained in Table 2.

The Black–Scholes formula quotes an option value in units of domestic currency per unit of foreign notional. Since this is usually a small number, it is often multiplied with 10 000 and quoted as domestic

Table 1 Standard market quotation of major currency pairs with sample spot prices

Currency pair	Default quotation	Sample quote
GBP/USD	GPB-USD	1.8000
GBP/CHF	GBP-CHF	2.2500
EUR/USD	EUR-USD	1.2000
EUR/GBP	EUR-GBP	0.6900
EUR/JPY	EUR-JPY	135.00
EUR/CHF	EUR-CHF	1.5500
USD/JPY	USD-JPY	108.00
USD/CHF	USD-CHF	1.2800

piPs per foreign, in short, **d piPs**. The others can be computed using the following instruction:

$$\mathbf{d\ piPs} \xrightarrow{\times \frac{1}{S}} \% \mathbf{f} \xrightarrow{\times \frac{S}{K}} \% \mathbf{d} \xrightarrow{\times \frac{1}{S}} \mathbf{f\ piPs} \xrightarrow{\times SK} \mathbf{d\ piPs} \quad (30)$$

Delta and Premium Convention

The spot delta of a European option without premium-adjustment is well known. It will be called *raw spot delta* δ_{raw} now and denotes the amount of FOR to buy when selling an option with 1 unit of FOR notional. However, the same option can also be viewed as an option with K units of DOM notional. The delta that goes with the same option, but 1 unit of DOM notional and tells how many units of DOM currency must be sold for the delta hedge is denoted by $\delta_{\text{raw}}^{\text{reverse}}$. In the market, both deltas can be quoted in either of the two currencies involved. The relationship is

$$\delta_{\text{raw}}^{\text{reverse}} = -\delta_{\text{raw}} \frac{S}{K} \quad (31)$$

The delta is used to buy or sell spot in the corresponding amount in order to hedge the option up

Table 2 Standard market quotation types for option values. In the example, we take FOR = EUR, DOM = USD, $S = 1.2000$, $r_d = 3.0\%$, $r_f = 2.5\%$, $\sigma = 10\%$, $K = 1.2500$, $T = 1$ year, $\phi = +1$ (call), notional = 1 000 000 EUR = 1 250 000 USD. For the piPs, the quotation 291.48 USD piPs per EUR is also sometimes stated as 2.9148% USD per 1 EUR. Similarly, the 194.32 EUR piPs per USD can also be quoted as 1.9432% EUR per 1 USD

Name	Symbol	Value in units of	Example
Domestic cash	d	DOM	29 148 USD
Foreign cash	f	FOR	24 290 EUR
% domestic	% d	DOM per unit of DOM	2.3318% USD
% foreign	% f	FOR per unit of FOR	2.4290% EUR
Domestic piPs	d piPs	DOM per unit of FOR	291.48 USD piPs per EUR
Foreign piPs	f piPs	FOR per unit of DOM	194.32 EUR piPs per USD

to first order. To interpret this relationship, note that the minus sign refers to selling DOM instead of buying FOR, and the multiplication by S adjusts the amounts. Furthermore, we divide by the strike, because a call on 1 EUR corresponds to K USD puts. More details on delta conventions are contained in **Foreign Exchange Options: Delta- and At-the-money Conventions**

For consistency, the premium needs to be incorporated into the delta hedge, since a premium in foreign currency will already hedge part of the option's delta risk. To make this clear, let us consider EUR-USD. In the standard arbitrage theory, $v(S)$ denotes the value or premium in USD of an option with 1 EUR notional, if the spot is at S , and the raw delta v_S denotes the number of EUR to buy for the delta hedge. Therefore, Sv_S is the number of USD to sell. If now the premium is paid in EUR rather than in USD, then we already have $\frac{v}{S}$ EUR, and the number of EUR to buy has to be reduced by this amount, that is, if EUR is the premium currency, we need to buy $v_S - v/S$ EUR for the delta hedge or equivalently sell $Sv_S - v$ USD.

To quote an FX option, we need to first sort out which currency is domestic, which is foreign, what is the notional currency of the option, and what is the premium currency. Unfortunately, this is not symmetric, since the counterparty might have another notion of domestic currency for a given currency pair. Hence, in the professional interbank market, there is one notion of delta per currency pair. Normally, it is the left-hand side delta of the *Fenics*^a screen if the option is traded in left-hand side premium, which is normally the standard and right-hand side delta if it is traded with right-hand-side premium, for example, EUR/USD lhs, USD/JPY lhs, EUR/JPY lhs, AUD/USD rhs, and so on. Since OTM options are traded most of the time, the difference is not huge and hence does not create a huge spot risk.

Additionally, the standard delta per currency pair (left-hand-side delta in *Fenics* for most cases) is used to quote options in volatility. This has to be specified by currency pair.

This standard interbank notion must be adapted to the real delta risk of the bank for an automated trading system. For currency pairs where the risk-free currency of the bank is the domestic or base currency, it is clear that the delta is the raw delta of

the option, and for risky premium this premium must be included. In the opposite case, the risky premium and the market value must be taken into account for the base currency premium, such that these offset each other. And for premium in underlying currency of the contract, the market value needs to be taken into account. In this way, the delta hedge is invariant with respect to the risky currency notion of the bank, for example, the delta is the same for a USD-based bank and an EUR-based bank.

Example

We consider two examples in Tables 3 and 4 to compare the various versions of deltas that are used in practice.

Greeks in Terms of Deltas

In FX markets, the moneyness of vanilla options is always expressed in terms of deltas and prices

Table 3 1 y EUR call USD put strike $K = 0.9090$ for a EUR-based bank. Market data: spot $S = 0.9090$, volatility $\sigma = 12\%$, EUR rate $r_f = 3.96\%$, USD rate $r_d = 3.57\%$. The raw delta is 49.15%EUR and the value is 4.427%EUR

Delta currency	Prem currency	Fenics	Formula	Delta
% EUR	EUR	lhs	$\delta_{\text{raw}} - P$	44.72
% EUR	USD	rhs	δ_{raw}	49.15
% USD	EUR	rhs	$-(\delta_{\text{raw}} - P)S/K$	-44.72
% USD	USD	lhs	$-(\delta_{\text{raw}})S/K$	-49.15
		[flip F4]		

Table 4 1 y call EUR call USD put strike $K = 0.7000$ for a EUR-based bank. Market data: spot $S = 0.9090$, volatility $\sigma = 12\%$, EUR rate $r_f = 3.96\%$, USD rate $r_d = 3.57\%$. The raw delta is 94.82%EUR and the value is 21.88%EUR

Delta currency	Prem currency	Fenics	Formula	Delta
% EUR	EUR	lhs	$\delta_{\text{raw}} - P$	72.94
% EUR	USD	rhs	δ_{raw}	94.82
% USD	EUR	rhs	$-(\delta_{\text{raw}} - P)S/K$	-94.72
% USD	USD	lhs	$-(\delta_{\text{raw}})S/K$	-123.13
		[flip F4]		

are quoted in terms of volatility. This makes a 10-delta call a financial object as such independent of spot and strike. This method and the quotation in volatility makes objects and prices transparent in a very intelligent and user-friendly way. At this point, we list the Greeks in terms of deltas instead of spot and strike. Let us introduce the quantities

$$\Delta_+ \triangleq \phi e^{-r_f \tau} \mathcal{N}(\phi d_+) \text{ spot delta} \quad (32)$$

$$\Delta_- \triangleq -\phi e^{-r_d \tau} \mathcal{N}(\phi d_-) \text{ dual delta} \quad (33)$$

which we assume to be given. From these we can retrieve

$$d_+ = \phi \mathcal{N}^{-1}(\phi e^{r_f \tau} \Delta_+) \quad (34)$$

$$d_- = \phi \mathcal{N}^{-1}(-\phi e^{r_d \tau} \Delta_-) \quad (35)$$

Interpretation of Dual Delta

The dual delta introduced in equation (9) as the sensitivity with respect to strike has another—more practical—interpretation in an FX setup. Recall from equation (25) that for vanilla options the domestic value

$$v(S, K, \tau, \sigma, r_d, r_f, \phi) \quad (36)$$

corresponds to a foreign value

$$v\left(\frac{1}{S}, \frac{1}{K}, \tau, \sigma, r_f, r_d, -\phi\right) \quad (37)$$

up to an adjustment of the nominal amount by the factor SK . From a foreign point of view, the delta is

thus given by

$$\begin{aligned} & -\phi e^{-r_d \tau} \mathcal{N}\left(-\phi \frac{\ln\left(\frac{1/S}{1/K}\right) + \left(r_f - r_d + \frac{1}{2}\sigma^2\tau\right)}{\sigma\sqrt{\tau}}\right) \\ & = -\phi e^{-r_d \tau} \mathcal{N}\left(\phi \frac{\ln\left(\frac{S}{K}\right) + \left(r_d - r_f - \frac{1}{2}\sigma^2\tau\right)}{\sigma\sqrt{\tau}}\right) \\ & = \Delta_- \end{aligned} \quad (38)$$

which means that the dual delta is the delta from the foreign point of view.

Now, we list value, delta, and vega in terms of $S, \Delta_+, \Delta_-, r_d, r_f, \tau$, and ϕ .

- Value.

$$\begin{aligned} & v(S, \Delta_+, \Delta_-, r_d, r_f, \tau, \phi) \\ & = S\Delta_+ + S\Delta_- \frac{e^{-r_f \tau} n(d_+)}{e^{-r_d \tau} n(d_-)} \end{aligned} \quad (39)$$

- Spot delta.

$$\frac{\partial v}{\partial S} = \Delta_+ \quad (40)$$

- Vega.

$$\frac{\partial v}{\partial \sigma} = S e^{-r_f \tau} \sqrt{\tau} n(d_+) \quad (41)$$

Notice that vega does not require knowing the dual delta.

Table 5 Vega in terms of Delta for the standard maturity labels and various deltas. It shows that one can neutralize a vega position of a long 9M 35 delta call with 4 short 1M 20 delta puts. This offsetting, however, is not a static, but only a momentary hedge

Matrix/ Δ	50%	45%	40%	35%	30%	25%	20%	15%	10%	5%
1D	2	2	2	2	2	2	1	1	1	1
1W	6	5	5	5	5	4	4	3	2	1
1W	8	8	8	7	7	6	5	5	3	2
1M	11	11	11	11	10	9	8	7	5	3
2M	16	16	16	15	14	13	11	9	7	4
3M	20	20	19	18	17	16	14	12	9	5
6M	28	28	27	26	24	22	20	16	12	7
9M	34	34	33	32	30	27	24	20	15	9
1Y	39	39	38	36	34	31	28	23	17	10
2Y	53	53	52	50	48	44	39	32	24	14
3Y	63	63	62	60	57	53	47	39	30	18

Vega in Terms of Delta

The mapping

$$\Delta \mapsto v_\sigma = S e^{-r_f \tau} \sqrt{\tau} n(\mathcal{N}^{-1}(e^{r_f \tau} \Delta)) \quad (42)$$

is important for trading vanilla options. Observe that this function does not depend on r_d or σ , just on r_f . Quoting vega in % foreign will additionally remove the spot dependence. This means that for a moderately stable foreign term structure curve, traders will be able to use a moderately stable vega matrix. For $r_f = 3\%$, the vega matrix is presented in Table 5.

The most important result of this paragraph is the fact that vega can be written in terms of delta, which is the main reason why the FX market uses implied volatility quotation based on deltas in the first place.

End Notes

^a. Fenics is one of the standard tools for FX option pricing (see <http://www.fenics.com/>)

References

- [1] Bates, D. (1988). *Crashes, Options and International Asset Substitutability*. PhD Dissertation, Economics Department, Princeton University.
- [2] Bates, D. (1991). The crash of 1987—was it expected? The evidence from options markets, *The Journal of Finance* **46**, 1009–1044.
- [3] Bowie, J. & Carr, P. (1994). Static simplicity, *Risk Magazine* (7), 45–49. <http://www.riskpublications.com>
- [4] Carr, P. (1994). *European Put Call Symmetry*, Cornell University Working Paper.
- [5] Hakala, J. & Wystup, U. (2002). *Foreign Exchange Risk*, Risk Publications, London. <http://www.mathfinance.com/FXRiskBook/>.
- [6] Reiss, O. & Wystup, U. (2001). Efficient computation of option price sensitivities using homogeneity and other tricks, *The Journal of Derivatives* **9**(2), 41–53.
- [7] Shreve, S.E. (2004). *Stochastic Calculus for Finance II*. Springer.

Further Reading

Wystup, U. (2006). *FX Options and Structured Products*, Wiley Finance Series, Wiley. <http://fxoptions.mathfinance.com/>.

Related Articles

Black–Scholes Formula; Foreign Exchange Options; Delta- and At-the-money Conventions; Foreign Exchange Markets; Put–Call Parity.

UWE WYSTUP

Quanto Options

A quanto option can be any cash-settled option whose payoff is converted into a third currency at maturity at a prespecified rate, called the *quanto factor*. There can be quanto plain vanilla, quanto barriers, quanto forward starts, quanto corridors, and so on. The **arbitrage pricing theory** and the **fundamental theorem of asset pricing**, also covered for example in [3] and [2], allow the computation of option values. Other references include **Options: Basic Definitions; Option Pricing: General Principles; Foreign Exchange Markets**.

Foreign Exchange Quanto Drift Adjustment

We take the example of a gold contract with underlying XAU/USD in XAU–USD quotation that is quantoed into EUR. Since the payoff is in EUR, we let EUR be the numeraire or domestic or base currency and consider a **Black–Scholes model**

$$\begin{aligned} \text{XAU–EUR: } dS_t^{(3)} &= (r_{\text{EUR}} - r_{\text{XAU}})S_t^{(3)} dt \\ &+ \sigma_3 S_t^{(3)} dW_t^{(3)} \end{aligned} \quad (1)$$

$$\begin{aligned} \text{USD–EUR: } dS_t^{(2)} &= (r_{\text{EUR}} - r_{\text{USD}})S_t^{(2)} dt \\ &+ \sigma_2 S_t^{(2)} dW_t^{(2)} \end{aligned} \quad (2)$$

$$dW_t^{(3)} dW_t^{(2)} = -\rho_{23} dt \quad (3)$$

where we use a minus sign in front of the correlation, because both $S_t^{(3)}$ and $S_t^{(2)}$ have the same base currency (DOM), which is EUR in this case. The scenario is displayed in Figure 1. The actual underlying is then

$$\text{XAU–USD: } S_t^{(1)} = \frac{S_t^{(3)}}{S_t^{(2)}} \quad (4)$$

Using Itô's formula, we first obtain

$$\begin{aligned} d\frac{1}{S_t^{(2)}} &= -\frac{1}{(S_t^{(2)})^2} dS_t^{(2)} + \frac{1}{2} \cdot 2 \cdot \frac{1}{(S_t^{(2)})^3} (dS_t^{(2)})^2 \\ &= (r_{\text{USD}} - r_{\text{EUR}} + \sigma_2^2) \frac{1}{S_t^{(2)}} dt - \sigma_2 \frac{1}{S_t^{(2)}} dW_t^{(2)} \end{aligned} \quad (5)$$

and hence

$$\begin{aligned} dS_t^{(1)} &= \frac{1}{S_t^{(2)}} dS_t^{(3)} + S_t^{(3)} d\frac{1}{S_t^{(2)}} + dS_t^{(3)} d\frac{1}{S_t^{(2)}} \\ &= \frac{S_t^{(3)}}{S_t^{(2)}} (r_{\text{EUR}} - r_{\text{XAU}}) dt + \frac{S_t^{(3)}}{S_t^{(2)}} \sigma_3 dW_t^{(3)} \\ &\quad + \frac{S_t^{(3)}}{S_t^{(2)}} (r_{\text{USD}} - r_{\text{EUR}} + \sigma_2^2) dt \\ &\quad - \frac{S_t^{(3)}}{S_t^{(2)}} \sigma_2 dW_t^{(2)} + \frac{S_t^{(3)}}{S_t^{(2)}} \rho_{23} \sigma_2 \sigma_3 dt \\ &= (r_{\text{USD}} - r_{\text{XAU}} + \sigma_2^2 + \rho_{23} \sigma_2 \sigma_3) S_t^{(1)} dt \\ &\quad + S_t^{(1)} (\sigma_3 dW_t^{(3)} - \sigma_2 dW_t^{(2)}) \end{aligned} \quad (6)$$

Since $S_t^{(1)}$ is a geometric Brownian motion with volatility σ_1 , we introduce a new Brownian motion $W_t^{(1)}$ and find

$$\begin{aligned} dS_t^{(1)} &= (r_{\text{USD}} - r_{\text{XAU}} + \sigma_2^2 + \rho_{23} \sigma_2 \sigma_3) S_t^{(1)} dt \\ &\quad + \sigma_1 S_t^{(1)} dW_t^{(1)} \end{aligned} \quad (7)$$

Now Figure 1 and the *law of cosine* imply

$$\sigma_3^2 = \sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2 \quad (8)$$

$$\sigma_1^2 = \sigma_2^2 + \sigma_3^2 + 2\rho_{23}\sigma_2\sigma_3 \quad (9)$$

which yields

$$\sigma_2^2 + \rho_{23}\sigma_2\sigma_3 = \rho_{12}\sigma_1\sigma_2 \quad (10)$$

As explained in the *currency triangle* in Figure 1, ρ_{12} is the correlation between XAU–USD and USD–EUR, whence $\rho \triangleq -\rho_{12}$ is the correlation between XAU–USD and EUR–USD. Inserting this into equation (7), we obtain the usual formula for the drift adjustment

$$\begin{aligned} dS_t^{(1)} &= (r_{\text{USD}} - r_{\text{XAU}} - \rho\sigma_1\sigma_2) S_t^{(1)} dt \\ &\quad + \sigma_1 S_t^{(1)} dW_t^{(1)} \end{aligned} \quad (11)$$

This is the **risk neutral pricing** process that can be used for the valuation of any derivative depending on $S_t^{(1)}$, which is quantoed into EUR.

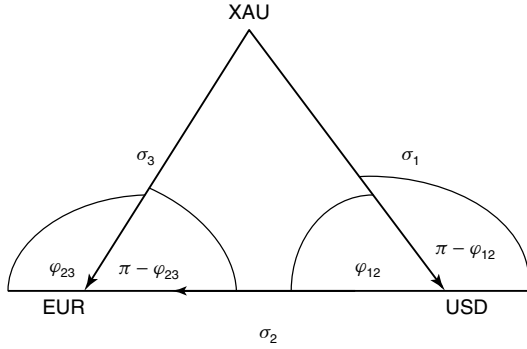


Figure 1 XAU-USD-EUR FX quanto triangle. The arrows point in the direction of the respective base currencies. The length of the edges represents the volatility. The cosine of the angles $\cos \phi_{ij} = \rho_{ij}$ represents the correlation of the currency pairs $S^{(i)}$ and $S^{(j)}$, if the base currency (DOM) of $S^{(i)}$ is the underlying currency (FOR) of $S^{(j)}$. If both $S^{(i)}$ and $S^{(j)}$ have the same base currency (DOM), then the correlation is denoted by $-\rho_{ij} = \cos(\pi - \phi_{ij})$

Extensions to Other Models

The previous derivation can be extended to the case of term-structure of volatility and correlation. However, introduction of volatility smile would distort the relationships. Nevertheless, accounting for smile effects is important in real-market scenarios. See **Foreign Exchange Smiles** and **Foreign Exchange Smile Interpolation** for details. To do this, one could, for example, capture the smile for a multicurrency model with a *weighted Monte Carlo technique* as described in [1]. This would still allow to use the previous result.

Quanto Vanilla

Common among **foreign exchange options** is a quanto plain vanilla paying

$$Q[\phi(S_T - K)]^+ \quad (12)$$

where K denotes the strike, T the expiration time, ϕ the usual put-call indicator taking the value +1 for a call and -1 for a put, S the underlying in FOR-DOM quotation, and Q the quanto factor from the domestic currency into the quanto currency. We let

$$\tilde{\mu} \triangleq r_d - r_f - \rho\sigma\tilde{\sigma} \quad (13)$$

be the *adjusted drift*, where r_d and r_f denote the risk-free rates of the domestic and foreign underlying currency pair, respectively, $\sigma = \sigma_1$ the volatility of this currency pair, $\tilde{\sigma} = \sigma_2$ the volatility of the currency pair DOM-QUANTO, and

$$\rho = \frac{\sigma_3^2 - \sigma^2 - \tilde{\sigma}^2}{2\sigma\tilde{\sigma}} \quad (14)$$

the correlation between the currency pairs FOR-DOM and DOM-QUANTO in this quotation. Furthermore, we let r_Q be the risk-free rate of the quanto currency. With the same principles as in **pricing formulae for foreign exchange options**, we can derive the formula for the value as

$$v = Qe^{-r_Q T} \phi[S_0 e^{\tilde{\mu} T} \mathcal{N}(\phi d_+) - K \mathcal{N}(\phi d_-)] \quad (15)$$

$$d_{\pm} = \frac{\ln \frac{S_0}{K} + (\tilde{\mu} \pm \frac{1}{2}\sigma^2) T}{\sigma \sqrt{T}} \quad (16)$$

where $\mathcal{N}$ denotes the cumulative standard normal distribution function and n its density.

Quanto Forward

Similarly, we can easily determine the value of a quanto forward paying

$$Q[\phi(S_T - K)] \quad (17)$$

where K denotes the strike, T the expiration time, ϕ the usual long-short indicator, S the underlying in FOR-DOM quotation, and Q the quanto factor from the domestic currency into the quanto currency. Then the formula for the value can be written as

$$v = Qe^{-r_Q T} \phi[S_0 e^{\tilde{\mu} T} - K] \quad (18)$$

This follows from the vanilla quanto value formula by taking both the normal probabilities to be 1. These normal probabilities are exercise probabilities under some measure. Since a forward contract is always exercised, both these probabilities must be equal to 1.

Quanto Digital

A European-style quanto digital pays

$$Q \mathbb{I}_{\{\phi S_T \geq \phi K\}} \quad (19)$$

Table 1 Example of a quanto digital put. The buyer receives 100 000 EUR if at maturity, the European Central Bank fixing for USD–JPY (computed *via* EUR–JPY and EUR–USD) is below 108.65. Terms were created on January 12, 2004 with the following market data: USD–JPY spot reference 106.60, USD–JPY at-the-money volatility 8.55%, EUR–JPY at-the-money volatility 6.69%, EUR–USD at-the-money volatility 10.99% (corresponding to a correlation of -27.89% for USD–JPY against JPY–EUR), USD rate 2.5%, JPY rate 0.1%, and EUR rate 4%

Notional	100 000 EUR
Maturity	3 months (92 days)
European-style barrier	108.65 USD–JPY
Theoretical value	71 555 EUR
Fixing source	European Central Bank

where K denotes the strike, S_T is the spot of the currency pair FOR–DOM at maturity T , ϕ takes the values $+1$ for a digital call and -1 for a digital put, and Q is the prespecified conversion rate from the domestic to the quanto currency. The valuation of the European-style quanto digitals follows the same principle as in the quanto vanilla option case. The value is

$$v = Qe^{-r_Q T} \mathcal{N}(\phi d_-) \quad (20)$$

We provide an example of a European-style digital put in USD/JPY quantoed into EUR in Table 1.

Hedging of Quanto Options

Hedging of quanto options can be done by running a multicurrency options book. All the usual Greeks can be hedged. **Delta hedging** is done by trading in the underlying spot market. An exception is the **correlation risk**, which can only be hedged with other derivatives depending on the same correlation. This is often difficult to do in practice. In FX, the correlation risk can be translated into vega positions as shown in [4, 5] or in **Foreign Exchange Basket Options**. We now illustrate this approach for quanto plain vanilla options.

Vega Positions of Quanto Plain Vanilla Options

Starting from equation (15), we obtain the sensitivities

$$\frac{\partial v}{\partial \sigma} = QS_0 e^{(\tilde{\mu} - r_Q)T} \left[n(d_+) \sqrt{T} - \phi \mathcal{N}(\phi d_+) \rho \tilde{\sigma} T \right],$$

$$\begin{aligned} \frac{\partial v}{\partial \tilde{\sigma}} &= -QS_0 e^{(\tilde{\mu} - r_Q)T} \phi \mathcal{N}(\phi d_+) \rho \sigma T, \\ \frac{\partial v}{\partial \rho} &= -QS_0 e^{(\tilde{\mu} - r_Q)T} \phi \mathcal{N}(\phi d_+) \sigma \tilde{\sigma} T, \\ \frac{\partial v}{\partial \sigma_3} &= \frac{\partial v}{\partial \rho} \frac{\partial \rho}{\partial \sigma_3} \\ &= \frac{\partial v}{\partial \rho} \frac{\sigma_3}{\sigma \tilde{\sigma}} \\ &= -QS_0 e^{(\tilde{\mu} - r_Q)T} \phi \mathcal{N}(\phi d_+) \sigma \tilde{\sigma} T \frac{\sigma_3}{\sigma \tilde{\sigma}} \\ &= -QS_0 e^{(\tilde{\mu} - r_Q)T} \phi \mathcal{N}(\phi d_+) \sigma_3 T \\ &= -QS_0 e^{(\tilde{\mu} - r_Q)T} \phi \mathcal{N}(\phi d_+) \sqrt{\sigma^2 + \tilde{\sigma}^2 + 2\rho\sigma\tilde{\sigma}} T \end{aligned} \quad (21)$$

Note that the computation is standard calculus and repeatedly uses the identity

$$S_0 e^{\tilde{\mu} T} n(\phi d_+) = Kn(\phi d_-) \quad (22)$$

The understanding of these greeks is that σ and $\tilde{\sigma}$ are both risky parameters, independent of each other. The third independent risk is either σ_3 or ρ , depending on what is more likely to be known.

This shows exactly how the three vega positions can be hedged with plain vanilla options in all the three legs, provided there is a liquid vanilla options market in all the three legs. In the example with XAU–USD–EUR, the currency pairs XAU–USD and EUR–USD are traded; however, there is no liquid vanilla market in XAU–EUR. Therefore, the correlation risk remains unhedgeable. Similar statements would apply for quantoed stocks or stock indices. However, in FX, there are situations with all the legs being hedgeable, for instance, EUR–USD–JPY.

The signs of the vega positions are not uniquely determined in all the legs. The FOR–DOM vega is smaller than the corresponding vanilla vega in the case of a call and positive correlation or put and negative correlation, and larger in case of a put and positive correlation or call and negative correlation. The DOM–QUANTO vega takes the sign of the correlation in case of a call and its opposite sign in case of a put. The FOR–QUANTO vega takes the opposite sign of the put–call indicator ϕ .

We provide an example of pricing and vega hedging scenario in Table 2, where we notice that the dominating vega risk comes from the FOR–DOM pair, whence most of the risk can be hedged.

Table 2 Example of a quanto plain vanilla

		Data set 1	Data set 2	Data set 3
FX pair	FOR–DOM	XAU–USD	XAU–USD	XAU–USD
Spot	FOR–DOM	800.00	800.00	800.00
Strike	FOR–DOM	810.00	810.00	810.00
Quanto	DOM–QUANTO	1.0000	1.0000	1.0000
Volatility	FOR–DOM	10.00%	10.00%	10.00%
Quanto volatility	DOM–QUANTO	12.00%	12.00%	12.00%
Correlation	FOR–DOM–DOM–QUANTO	25.00%	25.00%	–75.00%
Domestic interest rate	DOM	2.0000%	2.0000%	2.0000%
Foreign interest rate	FOR	0.5000%	0.5000%	0.5000%
Quanto currency rate	Q	4.0000%	4.0000%	4.0000%
Time in years	T	1	1	1
1 = call –1 = put	FOR	1	–1	1
Quanto vanilla option	Value	30.81329	31.28625	35.90062
Quanto vanilla option	Vega FOR–DOM	298.14188	321.49308	350.14600
Quanto vanilla option	Vega DOM–QUANTO	–10.07056	9.38877	33.38797
Quanto vanilla option	Vega FOR–QUANTO	–70.23447	65.47953	–35.61383
Quanto vanilla option	Correlation risk	–4.83387	4.50661	–5.34207
Quanto vanilla option	Vol FOR–QUANTO	17.4356%	17.4356%	8.0000%
Vanilla option	Value	32.6657	30.7635	32.6657
Vanilla option	Vega	316.6994	316.6994	316.6994

Applications

The standard applications are performance-linked deposits or performance notes as in [6]. Any time the performance of an underlying asset needs to be converted into the notional currency invested, and the exchange rate risk is with the seller, we need a quanto product. Naturally, an underlying like gold, which is quoted in USD, would be a default candidate for a quanto product, when the investment is in a currency other than USD.

Performance-linked Deposits

A performance-linked deposit is a deposit with a participation in an underlying market. The standard is that a GBP investor waives her coupon that the money market would pay and instead buys a EUR–GBP call with the same maturity date as the coupon, strike K and notional N in EUR. These parameters have to be chosen in such a way that the offer price of the EUR call equals the money market interest rate plus the sales margin. The strike

is often chosen to be the current spot. The notional is often a percentage p of the deposit amount A , such as 50 or 25%. The annual coupon paid to the investor is then a predefined minimum coupon plus the participation

$$p \cdot \frac{\max[S_T - S_0, 0]}{S_0} \quad (23)$$

which is the return of the exchange rate viewed as an asset, where the investor is protected against negative returns. So, obviously, the investor buys a EUR call GBP put with strike $K = S_0$ and notional $N = pA$ GBP or $N = pA/S_0$ EUR. Thus, if the EUR goes up by 10% against the GBP, the investor gets a coupon of $p \cdot 10\%$ per annum in addition to the minimum coupon.

Example 1 We consider the example shown in Table 3. In this case, if the EUR–GBP spot fixing is 0.7200, the additional coupon would be 0.8571% per annum. The breakeven point is at 0.7467, so this product is advisable for a very strong EUR bullish view. For a weakly bullish view, an alternative would

Table 3 Example of a performance-linked deposit, where the investor is paid 30% of the EUR–GBP return. Note that in GBP the day count convention in the money market is act^(a)/365 rather than act/360

Notional	5 000 000 GBP
Start date	3 June 2005
Maturity	2 September 2005 (91 days)
Number of days (act)	91
Money market reference rate	4.00% act/365
EUR–GBP spot reference	0.7000
Minimum rate	2.00% act/365
Additional coupon	$30\% \cdot \frac{100 \max[S_T - 0.7000, 0]}{0.7000}$ act/365
S_T	EUR–GBP fixing on 31 August 2005 (88 days)
Fixing source	ECB

^(a)(act = actual number of days)

be to buy an up-and-out call with barrier at 0.7400 and 75% participation, where we would find the best case to be 0.7399 with an additional coupon of 4.275% per annum, which would lead to a total coupon of 6.275% per annum.

Composition

- From the money market we get 49 863.01 GBP at the maturity date.
- The investor buys a EUR call GBP put with strike 0.7000 and with notional 1.5 million GBP.
- The offer price of the call is 26 220.73 GBP, assuming a volatility of 8.0% and a EUR rate of 2.50%.
- The deferred premium is 24 677.11 GBP.
- The investor receives a minimum payment of 24 931.51 GBP.
- Subtracting the deferred premium and the minimum payment from the money market leaves a sales margin of 254.40 GBP (which is extremely poor).
- Note that the option the investor is buying must be cash-settled.

Variations. There are many variations of the performance-linked notes. Of course, one can think of the European style knock-out calls or window-barrier calls. For a participation in a downward trend, the investor can buy puts. One of the frequent issues in

foreign exchange, however, is the deposit currency being different from the domestic currency of the exchange rate, which is quoted in FOR–DOM (foreign–domestic), meaning how many units of domestic currency are required to buy one unit of foreign currency. So, if we have a EUR investor who wishes to participate in a EUR–USD movement, we need to *quanto* the domestic payoff currency (USD) into the foreign currency (EUR). The payoff of the EUR call USD put

$$[S_T - K]^+ \quad (24)$$

is in domestic currency (USD). Of course, this payoff can be converted into the foreign currency (EUR) at maturity, but the question is, at what rate? If we convert at rate S_T , which is what we could do in the spot market at no cost, then the investor buys a vanilla EUR call. But here, the investor receives a coupon given by

$$p \cdot \frac{\max[S_T - S_0, 0]}{S_T} \quad (25)$$

If the investor wishes to have performance of equation (23) rather than equation (25), then the payoff at maturity is converted at a rate of 1.0000 into EUR, and this rate is set at the beginning of the trade. This is the *quanto factor*, and the vanilla is actually a *self-quanto* vanilla, that is, a EUR call USD put, cash settled in EUR, where the payoff in USD is converted into EUR at a rate of 1.0000. This self-quanto vanilla can be valued by inverting the exchange rate, that is, looking at USD–EUR. This way the valuation can incorporate the smile of EUR–USD.

Similar considerations need to be taken into account if the currency pair to participate in does not contain the deposit currency at all. A typical situation is a EUR investor, who wishes to participate in the gold price, which is measured in USD, so the investor needs to buy a XAU call USD put quantoed into EUR. So the investor is promised a coupon as in equation (23) for a XAU–USD underlying, where the coupon is paid in EUR; this implicitly means that we must use a quanto plain vanilla with a quanto factor of 1.0000.

References

- [1] Avellaneda, M., Buff, R., Friedman, C., Grandechamp, N., Kruk, L. & Newman, J. (2001). Weighted Monte Carlo: a new technique for calibrating asset-pricing models,

- International Journal of Theoretical and Applied Finance* **4**(1), 91–119.
- [2] Hakala, J. & Wystup, U. (2002). *Foreign Exchange Risk*, Risk Publications, London.
- [3] Shreve, S.E. (2004). *Stochastic Calculus for Finance I+II*, Springer.
- [4] Wystup, U. (2001). *How the Greeks would have hedged correlation risk of foreign exchange options*, Wilmott Research Report, August 2001.
- [5] Wystup, U. (2002). How the Greeks would have hedged correlation risk of foreign exchange options, in *Foreign Exchange Risk*, Risk Publications, London.
- [6] Wystup, U. (2006). *FX Options and Structured Products*, Wiley Finance Series.

Related Articles

Black–Scholes Formula; Foreign Exchange Markets; Foreign Exchange Options.

UWE WYSTUP

Vanna–Volga Pricing

The vanna–volga method, also called the *traders’ rule of thumb*, is an empirical procedure that can be used to infer an implied-volatility smile from three available quotes for a given maturity. It is based on the construction of locally replicating portfolios whose associated hedging costs are added to corresponding Black–Scholes prices to produce smile-consistent values. Besides being intuitive and easy to implement, this procedure has a clear financial interpretation, which further supports its use in practice. In fact, SuperDerivatives has implemented a type of this method in their pricing platform, as one can read in the patent that SuperDerivatives has filed.

The vanna–volga method is commonly used in foreign exchange options markets, where three main volatility quotes are typically available for a given market maturity: the delta-neutral straddle, referred to as *at-the-money (ATM)*; the risk reversal (RR) for 25 delta call and put; and the (vega-weighted) butterfly (BF) with 25 delta wings. The application of vanna–volga pricing allows us to derive implied volatilities for any option’s delta, in particular for those outside the basic range set by the 25 delta put and call quotes. The notion of risk reversals and butterflies is explained in the article on foreign exchange (FX) market terminology (see **Foreign Exchange Markets**).

In the financial literature, the vanna–volga approach was introduced by Lipton and McGhee in [2], who compare different approaches to the pricing of double-no-touch (DNT) options, and by Wystup in [5], who describes its application to the valuation of one-touch (OT) options. The vanna–volga procedure is reviewed in more detail and some important results concerning the tractability of the method and its robustness are derived by Castagna and Mercurio in [1].

The following is based on the section *Traders’ Rule of Thumb* by Wystup in [6].

The traders’ rule of thumb is a method of traders to determine the cost of risk managing the volatility risk of exotic options with vanilla options. This cost is then added to the theoretical value (*TV*) in the Black–Scholes model and is called the *overhedge*. We explain the rule and then consider an example of a one-touch option.

Delta and vega are the most relevant sensitivity parameters for FX options maturing within one year. A delta-neutral position can be achieved by trading the spot. Changes in the spot are explicitly allowed in the Black–Scholes model. Therefore, model and practical trading have very good control over spot change risk. The more sensitive part is the vega position. This is not taken care of in the Black–Scholes model. Market participants need to trade other options to obtain a vega-neutral position. However, even a vega-neutral position is subject to changes of spot and volatility. For this reason, the sensitivity parameters *vanna* (change of vega due to change of spot) and *volga* (change of vega due to change of volatility) are of special interest. Vanna is also called $d \text{ vega} / d \text{ spot}$, volga is also called $d \text{ vega} / d \text{ vol}$. The plots for vanna and volga for a vanilla option are displayed in Figures 1 and 2. In this section, we outline how the cost of such a vanna and volga exposure can be used to obtain prices for options that are closer to the market than their theoretical Black–Scholes value.

Cost of Vanna and Volga

We fix the rates r_d and r_f , the time to maturity T , and the spot x and define

$$\begin{aligned} \text{cost of vanna} &\triangleq \text{exotic vanna ratio} \\ &\quad \times \text{value of RR} \end{aligned} \quad (1)$$

$$\begin{aligned} \text{cost of volga} &\triangleq \text{exotic volga ratio} \\ &\quad \times \text{value of BF} \end{aligned} \quad (2)$$

$$\text{exotic vanna ratio} \triangleq B_{\sigma x} / \text{RR}_{\sigma x} \quad (3)$$

$$\text{exotic volga ratio} \triangleq B_{\sigma\sigma} / \text{BF}_{\sigma\sigma} \quad (4)$$

$$\text{value of RR} \triangleq [\text{RR}(\sigma_\Delta) - \text{RR}(\sigma_0)] \quad (5)$$

$$\text{value of BF} \triangleq [\text{BF}(\sigma_\Delta) - \text{BF}(\sigma_0)] \quad (6)$$

where σ_0 denotes the ATM (forward) volatility and σ_Δ denotes the wing volatility at the delta pillar Δ , and B denotes the value function of a given exotic option. The values of risk reversals and butterflies are defined by

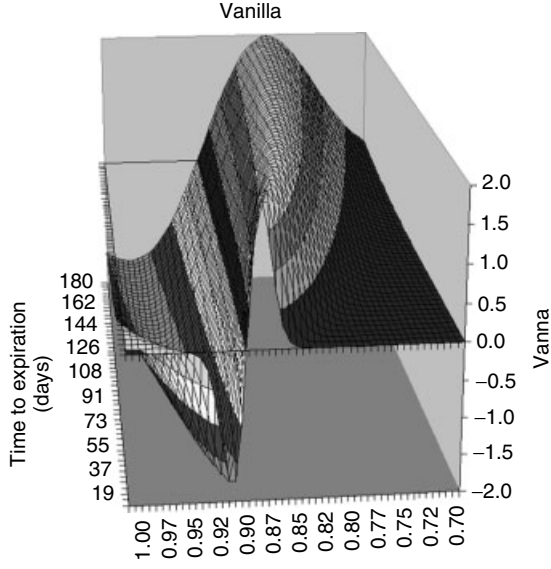


Figure 1 Vanna of a vanilla option as a function of spot and time to expiration, showing the skew symmetry about the at-the-money line

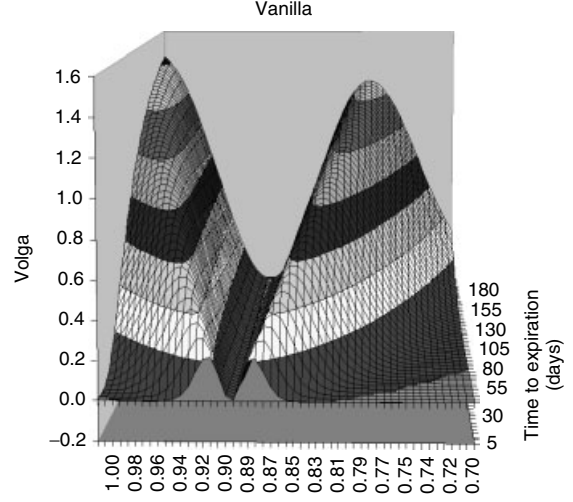


Figure 2 Volga of a vanilla option as a function of spot and time to expiration, showing the symmetry about the at-the-money line

$$RR(\sigma) \triangleq \text{call}(x, \Delta, \sigma, r_d, r_f, T) - \text{put}(x, \Delta, \sigma, r_d, r_f, T) \quad (7)$$

$$BF(\sigma) \triangleq \frac{\text{call}(x, \Delta, \sigma, r_d, r_f, T) + \text{put}(x, \Delta, \sigma, r_d, r_f, T)}{2} - \frac{\text{call}(x, \Delta_0, \sigma_0, r_d, r_f, T) + \text{put}(x, \Delta_0, \sigma_0, r_d, r_f, T)}{2} \quad (8)$$

where $\text{vanilla}(x, \Delta, \sigma, r_d, r_f, T)$ means $\text{vanilla}(x, K, \sigma, r_d, r_f, T)$ for a strike K chosen to imply $|\text{vanilla}_x(x, K, \sigma, r_d, r_f, T)| = \Delta$ and Δ_0 is the delta that produces the ATM strike. To summarize, we abbreviate

$$c(\sigma_\Delta^+) \triangleq \text{call}(x, \Delta^+, \sigma_\Delta^+, r_d, r_f, T) \quad (9)$$

$$p(\sigma_\Delta^-) \triangleq \text{put}(x, \Delta^-, \sigma_\Delta^-, r_d, r_f, T) \quad (10)$$

and obtain

cost of vanna

$$= \frac{B_{\sigma x}}{c_{\sigma x}(\sigma_\Delta^+) - p_{\sigma x}(\sigma_\Delta^-)} \times [c(\sigma_\Delta^+) - c(\sigma_0) - p(\sigma_\Delta^-) + p(\sigma_0)] \quad (11)$$

cost of volga

$$= \frac{2B_{\sigma\sigma}}{c_{\sigma\sigma}(\sigma_\Delta^+) + p_{\sigma\sigma}(\sigma_\Delta^-)} \times \left[\frac{c(\sigma_\Delta^+) - c(\sigma_0) + p(\sigma_\Delta^-) - p(\sigma_0)}{2} \right] \quad (12)$$

where we note that volga of the butterfly should actually be

$$\frac{1}{2} [c_{\sigma\sigma}(\sigma_\Delta^+) + p_{\sigma\sigma}(\sigma_\Delta^-) - c_{\sigma\sigma}(\sigma_0) - p_{\sigma\sigma}(\sigma_0)] \quad (13)$$

but the last two summands are close to zero. The *vanna–volga adjusted value* of the exotic is then

$$B(\sigma_0) + p \times [\text{cost of vanna} + \text{cost of volga}] \quad (14)$$

A division by the spot x converts everything into the usual quotation of the price in per cent of the underlying currency. The cost of vanna and volga is commonly adjusted by a number $p \in [0, 1]$, which is often taken to be the risk-neutral no-touch (NT) probability. The reason is that in the case of options that can knock out, the hedge is not needed anymore once the option has knocked out. The exact choice of p depends on the product to be priced; see Table 1. Taking $p = 1$ as the default value would lead to overestimated overhedges for DNT options as pointed out in [2].

The values of risk reversals and butterflies in equations (11) and (12) can be approximated by a first-order expansion as follows. For a risk reversal, we take the difference of the call with correct implied volatility and the call with ATM volatility minus the difference of the put with correct implied volatility and the put with ATM volatility. It is easy to see that this can be well-approximated by the vega of the ATM vanilla times the risk reversal in terms of volatility. Similarly, the cost of the butterfly can be approximated by the vega of the ATM volatility times the butterfly in terms of volatility. In formulae, this is

$$\begin{aligned} & c(\sigma_\Delta^+) - c(\sigma_0) - p(\sigma_\Delta^-) + p(\sigma_0) \\ & \approx c_\sigma(\sigma_0)(\sigma_\Delta^+ - \sigma_0) - p_\sigma(\sigma_0)(\sigma_\Delta^- - \sigma_0) \\ & = \sigma_0[p_\sigma(\sigma_0) - c_\sigma(\sigma_0)] + c_\sigma(\sigma_0)[\sigma_\Delta^+ - \sigma_\Delta^-] \\ & = c_\sigma(\sigma_0)RR \end{aligned} \quad (15)$$

and, similarly,

$$\begin{aligned} & \frac{c(\sigma_\Delta^+) - c(\sigma_0) + p(\sigma_\Delta^-) - p(\sigma_0)}{2} \\ & \approx c_\sigma(\sigma_0)BF \end{aligned} \quad (16)$$

Table 1 Adjustment factors for the overhedge for first-generation exotics

Option	p
KO	No-touch probability
RKO	No-touch probability
DKO	No-touch probability
OT	$0.9 \times \text{no-touch probability} - 0.5 \times \text{bid-offer-spread} \times (TV - 33\%)/66\%$
DNT	0.5

KO, knock out; RKO, reverse knockout; DKO, double knock-out; OT, one touch; DNT, double no touch

With these approximations, we obtain the formulae

$$\text{cost of vanna} \approx \frac{B_{\sigma x}}{c_{\sigma x}(\sigma_\Delta^+) - p_{\sigma x}(\sigma_\Delta^-)} c_\sigma(\sigma_0)RR \quad (17)$$

$$\text{cost of volga} \approx \frac{2B_{\sigma\sigma}}{c_{\sigma\sigma}(\sigma_\Delta^+) + p_{\sigma\sigma}(\sigma_\Delta^-)} c_\sigma(\sigma_0)BF \quad (18)$$

Observations

1. The price supplements are linear in butterflies and risk reversals. In particular, there is no cost of vanna supplement if the risk reversal is zero and no cost of volga supplement if the butterfly is zero.
2. The price supplements are linear in the ATM vanilla vega. This means supplements grow with growing volatility change risk of the hedge instruments.
3. The price supplements are linear in vanna and volga of the given exotic option.
4. We have not observed any relevant difference between the exact method and its first-order approximation. Since the computation time for the approximation is shorter, we recommend using the approximation.
5. It is not clear up front which target delta to use for the butterflies and risk reversals. We take a delta of 25% merely on the basis of its liquidity.
6. The prices for vanilla options are consistent with the input volatilities as shown in Figures 3, 4, and 5.
7. The method assumes a zero volga of risk reversals and a zero vanna of butterflies. This way the two sources of risk can be decomposed and hedged with risk reversals and butterflies. However, the assumption is actually not exact. For this reason, the method should be used with a lot of care. It causes traders and financial engineers to keep adding exceptions to the standard method.

Consistency Check

A minimum requirement for the vanna–volga pricing to be correct is the consistency of the method with vanilla options. We show in Figures 3, 4, and 5 that

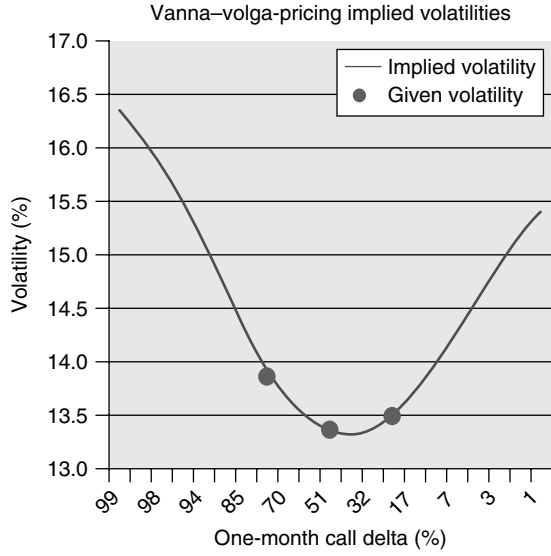


Figure 3 Consistency check of vanna-volga pricing. Vanilla option smile for a one-month maturity EUR/USD call, spot = 0.9060, $r_d = 5.07\%$, $r_f = 4.70\%$, $\sigma_0 = 13.35\%$, $\sigma_{\Delta}^+ = 13.475\%$, $\sigma_{\Delta}^- = 13.825\%$

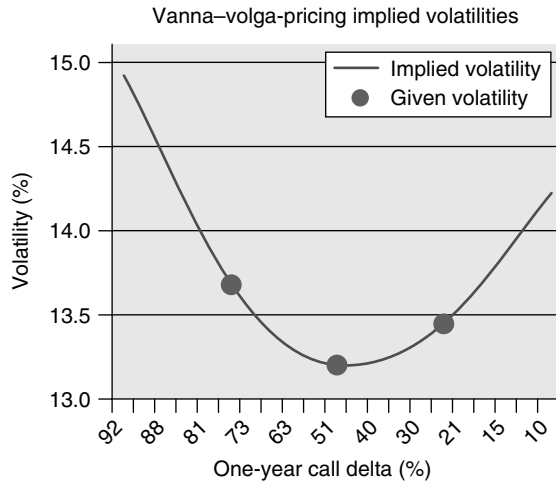


Figure 4 Consistency check of vanna-volga pricing. Vanilla option smile for a one-year maturity EUR/USD call, spot = 0.9060, $r_d = 5.07\%$, $r_f = 4.70\%$, $\sigma_0 = 13.20\%$, $\sigma_{\Delta}^+ = 13.425\%$, $\sigma_{\Delta}^- = 13.575\%$

the method does, in fact, yield a typical foreign exchange smile shape and produces the correct input volatilities ATM and at the delta pillars. We will now prove the consistency in the following way. Since the

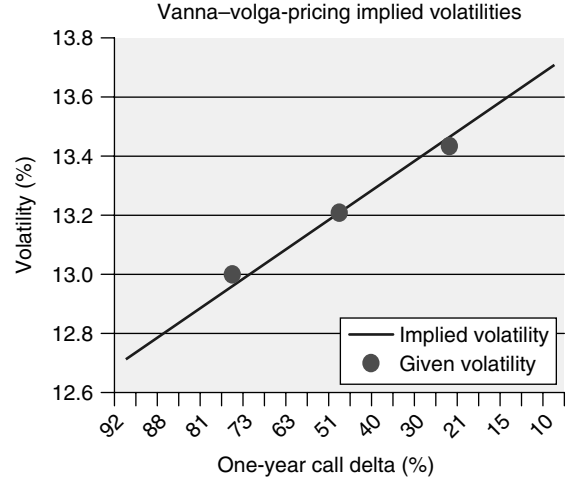


Figure 5 Consistency check of vanna-volga pricing. Vanilla option smile for a one-year maturity EUR/USD call, spot = 0.9060, $r_d = 5.07\%$, $r_f = 4.70\%$, $\sigma_0 = 13.20\%$, $\sigma_{\Delta}^+ = 13.425\%$, $\sigma_{\Delta}^- = 13.00\%$

input consists only of three volatilities (ATM and two delta pillars), it would be too much to expect that the method produces correct representation of the entire volatility matrix. We can only check if the values for ATM and target- Δ puts and calls are reproduced correctly. To verify this, we check if the values for an ATM call, a risk reversal, and a butterfly are priced correctly. Of course, we only expect approximately correct results. Note that the number p is taken to be 1, which agrees with the risk-neutral NT probability for vanilla options.

For an ATM call, vanna and volga are approximately zero, and hence there are no supplements due to vanna or volga cost.

For a target- Δ risk reversal,

$$c(\sigma_{\Delta}^+) - p(\sigma_{\Delta}^-) \quad (19)$$

we obtain

cost of vanna

$$\begin{aligned} &= \frac{c_{\sigma_x}(\sigma_{\Delta}^+) - p_{\sigma_x}(\sigma_{\Delta}^-)}{c_{\sigma_x}(\sigma_{\Delta}^+) - p_{\sigma_x}(\sigma_{\Delta}^-)} \\ &\quad \times [c(\sigma_{\Delta}^+) - c(\sigma_0) - p(\sigma_{\Delta}^-) + p(\sigma_0)] \\ &= c(\sigma_{\Delta}^+) - c(\sigma_0) - p(\sigma_{\Delta}^-) + p(\sigma_0) \end{aligned} \quad (20)$$

cost of volga

$$= \frac{2[c_{\sigma\sigma}(\sigma_{\Delta}^+) - p_{\sigma\sigma}(\sigma_{\Delta}^-)]}{c_{\sigma\sigma}(\sigma_{\Delta}^+) + p_{\sigma\sigma}(\sigma_{\Delta}^-)} \times \left[\frac{c(\sigma_{\Delta}^+) - c(\sigma_0) + p(\sigma_{\Delta}^-) - p(\sigma_0)}{2} \right] \quad (21)$$

and observe that the cost of vanna yields a perfect fit and the cost of volga is small, because in the first fraction we divide the difference of two quantities by the sum of the quantities, which are all of the same order.

For a target- Δ butterfly

$$\frac{c(\sigma_{\Delta}^+) + p(\sigma_{\Delta}^-)}{2} - \frac{c(\sigma_0) + p(\sigma_0)}{2} \quad (22)$$

we analogously obtain a perfect fit for the cost of volga and

cost of vanna

$$= \frac{c_{\sigma x}(\sigma_{\Delta}^+) - p_{\sigma x}(\sigma_0) - [c_{\sigma x}(\sigma_0) - p_{\sigma x}(\sigma_{\Delta}^-)]}{c_{\sigma x}(\sigma_{\Delta}^+) - p_{\sigma x}(\sigma_0) + [c_{\sigma x}(\sigma_0) - p_{\sigma x}(\sigma_{\Delta}^-)]} \times [c(\sigma_{\Delta}^+) - c(\sigma_0) - p(\sigma_{\Delta}^-) + p(\sigma_0)] \quad (23)$$

which is again small.

The consistency can actually fail for certain parameter scenarios. This is one of the reasons that the traders' rule of thumb has been criticized repeatedly by a number of traders and researchers.

We introduce the abbreviations for first generation exotics listed as below.

KO, knock-out; KI, knock-in; RKO, reverse knock-out; RKI, reverse knock-in; DKO, double knock-out; OT, one-touch; NT, no-touch; DOT, double one-touch; DNT, double no-touch.

Adjustment Factor

The factor p has to be chosen in a suitable fashion. Since there is no mathematical justification or indication, there is a lot of dispute in the market about this choice. Moreover, the choices may also vary over time. An example for *one of many possible choices* of p is presented in Table 1.

For options with strike K , barrier B and type $\phi = 1$ for a call and $\phi = -1$ for a put, we use the following pricing rules, which are based on no-arbitrage conditions:

- Knock in (KI) is priced *via* $KI = \text{vanilla} - KO$.
- Reverse knock in (RKI) is priced *via* $RKI = \text{vanilla} - RKO$.
- Reverse knockout (RKO) is priced *via* $RKO(\phi, K, B) = KO(-\phi, K, B) - KO(-\phi, B, B) + \phi(B - K)NT(B)$.
- Double one touch (DOT) is priced *via* DNT.
- NT is priced *via* OT.

Volatility for Risk Reversals, Butterflies and Theoretical Value

To determine the volatility and the vanna and volga for the risk reversal and butterfly, the convention is the same as for the building of the smile curve. Hence the 25% delta risk reversal retrieves the strike for 25% delta call and put with the spot delta and calculates the vanna and volga of these options using the corresponding volatilities from the smile.

The TV of the exotics is calculated using the ATM volatility, retrieving it with the same convention that was used to build the smile, to build the smile.

Pricing Barrier Options

Ideally, one would be in a situation to hedge all barrier contracts with a portfolio of vanilla options or simple barrier building blocks. In the Black–Scholes model, there are exact rules on how to statically hedge many barrier contracts. A state-of-the art reference is given in [3]. However, in practice, most of these hedges fail, because volatility is not constant.

For regular KO options, one can refine the method to incorporate more information about the global shape of the vega surface through time.

We chose M future points in time as $0 < a_1\% < a_2\% < \dots < a_M\%$ of the time to expiration. Using the same cost of vanna and volga, we calculate the overhedge for the regular KO with a reduced time to expiration. The factor for the cost is the probability not to touch the barrier within the remaining times to expiration $1 > 1 - a_1\% > 1 - a_2\% > \dots > 1 - a_M\%$ of the total time to expiration. Some desks believe that for ATM strikes, the long time to maturity should be weighted higher and for low-delta strikes the short time to maturity should be weighted higher. The weighting can be chosen (rather arbitrarily) as

$$w = \tanh[\gamma(|\delta - 50\%| - 25\%)] \quad (24)$$

with a suitable positive γ . For $M = 3$, the total overhedge is given by

$$OH = \frac{OH(1 - a_1\%) \times w + OH(1 - a_2\%) + OH(1 - a_3\%) \times (1 - w)}{3} \quad (25)$$

Which values to use for M , γ , and the a_i , whether to apply a weighting and what kind, varies for different trading desks.

An additional term can be used for single-barrier options to account for glitches in the stop loss of the barrier. The theoretical value of the barrier option is determined with a barrier that is moved by four basis points and 50% of that adjustment is added to the price if it is positive. If it is negative, it is omitted altogether. The theoretical foundation for such a method is explained in [4].

Pricing Double-barrier Options

Double-barrier options behave similar to vanilla options for a spot far away from the barrier and more like OT options for a spot close to the barrier. Therefore, it appears reasonable to use the traders' rule of thumb for the corresponding regular KO to determine the overhedge for a spot closer to the strike and for the corresponding OT option for a spot closer to the barrier. This adjustment is the intrinsic value of the RKO times the overhedge of the corresponding OT option. The border is the arithmetic mean between strike and the in-the-money barrier.

Pricing Double-no-touch Options

For DNT options with lower barrier L and higher barrier H at spot S , one can use the overhedge

$$OH = \max\{\text{vanna-volga-}OH; \delta(S - L) - TV - 0.5\%; \delta(H - S) - TV - 0.5\%\} \quad (26)$$

where δ denotes the delta of the DNT option.

Pricing European-style Options

Digital Options

Digital options are priced using the overhedge of the call/put spread with the corresponding volatilities.

European Barrier Options

European barrier options (EKO) are priced using the prices of European and digital options and the relationship

$$EKO(\phi, K, B) = \text{vanilla}(\phi, K) - \text{vanilla}(\phi, B) - \text{digital}(B)\phi(B - K) \quad (27)$$

No-touch Probability

The NT probability is obviously equal to the nondiscounted value of the corresponding NT option paying at maturity (under the risk-neutral measure). Note that the price of the OT option is calculated using an iteration for the touch probability. This means that the price of the OT option used to compute the NT probability is itself based on the traders' rule of thumb. This is an iterative process that requires an abortion criterion. One can use a standard approach that ends either after 100 iterations or as soon as the difference of two successive iteration results is less than 10^{-6} . However, the method is so crude that it actually does not make much sense to use such precision at just this point. Therefore, to speed up the computation, we suggest that this procedure is omitted and no iterations are taken, which means to use the non-discounted TV of the no-touch option as a proxy for the NT probability.

The Cost of Trading and Its Implication on the Market Price of One-touch Options

Now let us take a look at an example of the traders' rule of thumb in its simple version. We consider OT options, which hardly ever trade at TV . The tradable price is the sum of the TV and the overhedge. Typical examples are shown in Figure 6, one for an upper touch level in EUR/USD, and one for a lower touch level.

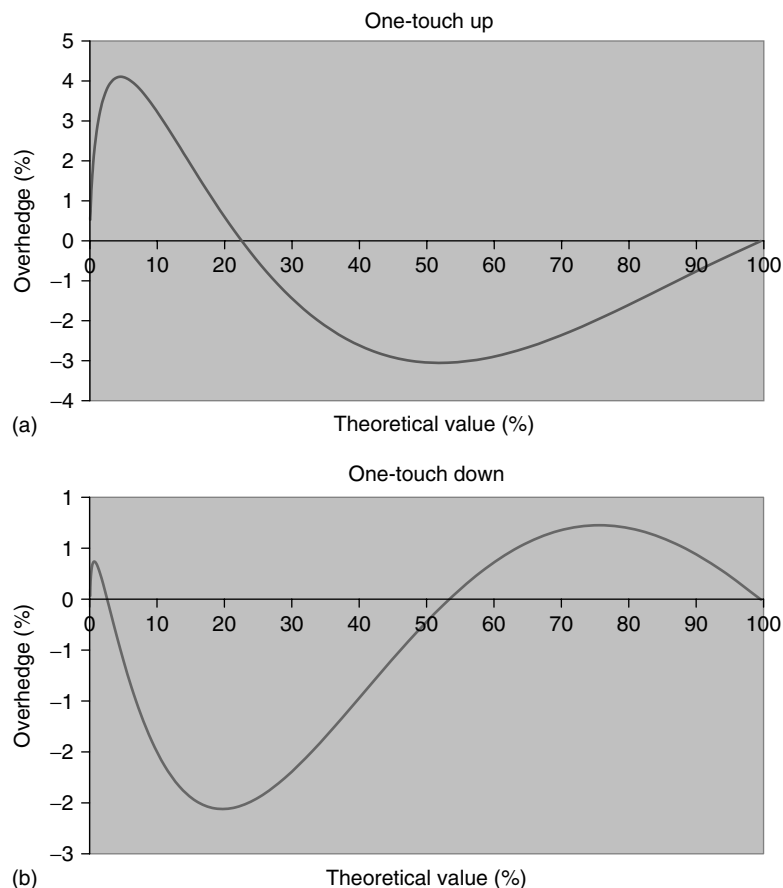


Figure 6 Overhedge of a one-touch option in EUR/USD for (a) an upper touch level and (b) a lower touch level, based on the traders' rule of thumb

Clearly, there is no overhedge for OT options with a TV of 0% or 100%, but it is worth noting that low-TV OT options can be twice as expensive as their TV, sometimes even more. The overhedge arises from the cost of risk managing the OT option. In the Black–Scholes model, the only source of risk is the underlying exchange rate, whereas the volatility and interest rates are assumed constant. However, volatility and rates are themselves changing, whence the trader of options is exposed to instable vega and rho (change of the value with respect to volatility and rates). For short-dated options, the interest rate risk is negligible compared to the volatility risk as shown in Figure 7. Hence the overhedge of an OT option is a reflection of a trader's cost occurring because of the risk management of his vega exposure.

Example

We consider a one-year OT option in USD/JPY with payoff in USD. As market parameters, we assume a spot of 117.00 JPY per USD, JPY interest rate 0.10%, USD interest rate 2.10%, volatility 8.80%, 25-delta risk reversal -0.45% ,^a and 25-delta butterfly 0.37% .^b

The touch level is 127.00, and the TV is at 28.8%. If we now only hedge the vega exposure, then we need to consider two main risk factors, namely,

1. the change of vega as the spot changes, often called vanna;
2. the change of vega as the volatility changes, often called volga or volgamma or vomma.

To hedge this exposure, we treat the two effects separately. The vanna of the OT option is 0.16%,

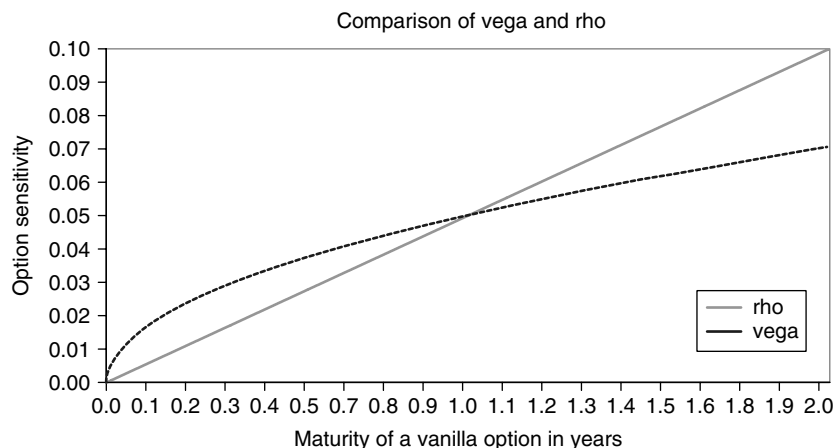


Figure 7 Comparison of interest rate and volatility risk for a vanilla option. The volatility risk behaves like a square-root function, whereas the interest rate risk is close to linear. Therefore, short-dated FX options have higher volatility risk than interest rate risk

and the vanna of the risk reversal is 0.04%. So we need to buy 4 ($= 0.16/0.04$) risk reversals, and, for each of them, we need to pay 0.14% of the USD amount, which causes an overhedge of -0.6% . The volga of the OT is -0.53% , and the volga of the butterfly is 0.03% . So we need to sell 18 ($= -0.53/0.03$) butterflies, each of which pays us 0.23% of the USD amount, which causes an overhedge of -4.1% . Therefore, the overhedge is -4.7% . However, we will get to the touch level with a risk-neutral probability of 28.8%, in which case we would have to pay to unwind the hedge. Therefore, the total overhedge is $-71.2\% \times 4.7\% = -3.4\%$. This leads to a midmarket price of 25.4%. Bid and offer could be 24.25%–36.75%. There are different beliefs among market participants about the unwinding cost. Other observed prices for OT options can be due to different existing vega profiles of the trader's portfolio, a marketing campaign, a hidden additional sales margin, or even the overall view of the trader in charge.

Further Applications

The method illustrated above shows how important the current smile of the vanilla options market is for the pricing of simple exotics. Similar types of approaches are commonly used to price other exotic options. For long-dated options, the interest rate risk will take over the lead in comparison to short-dated options where the volatility risk is dominant.

End Notes

^aThis means that a 25-delta USD call is 0.45% cheaper than a 25-delta USD put in terms of implied volatility.

^bThis means that a 25-delta USD call and 25-delta USD put is, on average, 0.37% more expensive than an ATM option in terms of volatility.

References

- [1] Castagna, A. & Mercurio, F. (2007). The Vanna-Volga method for implied volatilities, *Risk* Jan. 106–111.
- [2] Lipton, A. & McGhee, W. (2002). Universal Barriers, *Risk*, **15**(5), 81–85.
- [3] Poulsen, R. (2006). Barrier options and their static hedges: simple derivations and extensions, *Quantitative Finance*, Vol. **6**(4), 327–335.
- [4] Schmock, U., Shreve, S.E. & Wystup, U. (2002). Dealing with dangerous digitals, in *Foreign Exchange Risk*, Risk Publications, London, <http://www.mathfinance.com/FXRiskBook/>
- [5] Wystup, U. (2003). The market price of one-touch options in foreign exchange markets, *Derivatives Week*, London, **XII**(13), 8–9.
- [6] Wystup, U. (2006). *FX Options and Structured Products*, Wiley Finance Series.

Related Articles

Barrier Options; Foreign Exchange Markets.

UWE WYSTUP

Foreign Exchange Smiles

Smile Regularities for Foreign Exchange Options

One trend in the empirical investigation of implied volatilities has been to concentrate on understanding the behavior of implied volatilities across strike prices and time to expiration [see 10]. This line of research assumes implicitly that these divergences provide information about the dynamics of the options markets. Another approach [3, 5, 6, 14] suggests that the divergences of implied volatilities across strike prices may provide information about the expected dispersion process for underlying asset prices. These papers assume that asset return volatility is a (locally) deterministic function of the asset price and time and that this information can be used to enhance the traditional Black–Scholes–Merton (BSM) option-pricing approach (*see also Dupire Equation; Local Volatility Model*). All these papers examine implied volatility patterns at a single point in time and assume that option prices provide an indication of the deterministic volatility function. However, Dumas *et al.* [4] (1998) tested for the existence of a deterministic implied volatility function and rejected the hypothesis that the inclusion of such a model in option pricing was an improvement in terms of predictive or hedging performance compared with BSM. Their research examined whether at a single point in time, implied volatility surfaces provide predictions of implied volatilities at some future date (one week hence).

Tompkins [15] looked at this problem in a slightly different way. The approach of Dumas *et al.* [4] assumes that the deterministic volatility function provides both a prediction of the future levels of implied volatility and the relative shapes of implied volatilities across strike prices and time. If the future levels of implied volatilities cannot be predicted, this does not mean that the *relative* shapes of implied volatilities cannot be predicted. Tompkins [15] examined the relative implied volatility bias rather than the absolute implied volatility bias. When the volatilities of each strike price were standardized by dividing the level of the at-the-money (ATM) volatility, regularities in the volatility function were found. He further found that these standardized smile

patterns were dependent upon the term to expiration of the option. For a large sample of option expiration cycles, the smile patterns were almost identical for all options with the same time to expiration.

For currency options, Tompkins [15] examined options on futures for US dollar/Deutsche mark, US dollar/British pound, US dollar/Japanese yen, and US dollar/Swiss franc for a time period from 1985 to 2000. To determine relative shapes, the implied volatilities for each currency pair were standardized by

$$VSI = \left(\frac{\sigma_x}{\sigma_{ATM}} \right) \cdot 100 \quad (1)$$

where VSI is the volatility smile index, σ_x is the volatility of an option with strike price x , and σ_{ATM} is the volatility of the ATM option. The ATM volatility was determined using a simple linear interpolation for the two implied volatilities of the strike prices that bracketed the underlying asset price. This relative volatility measure will facilitate comparisons of biases (in percentage terms) within and between markets.

The strike prices were standardized to allow intra- and intermarket comparisons to be drawn. The standardized strike prices can be expressed as

$$\frac{\ln(X_\tau/F_\tau)}{\sigma\sqrt{\tau/365}} \quad (2)$$

where X is the strike price of the option, F is the underlying futures price and the square root of time factor reflects the percentage in a year of the remaining time until the expiration of the option. The sigma (σ) is the level of the ATM volatility.

As the analysis was restricted to the actively traded quarterly expiration schedule of March, June, September, and December maturities, implied volatility surfaces with a maximum term to expiration of approximately 90 days were obtained. Data were further pruned by restricting the analysis to 18 time points from (the date nearest to) 90 calendar days to expiration to (the date nearest) 5 calendar days to expiration in 5-day increments. Finally, the analysis of the implied volatilities was limited to those strike prices in the range ± 3.5 standard deviations away from the underlying futures price. Figure 1 displays the aggregated patterns for the 15-year period.

A logical starting point for an appropriate functional form to fit an implied volatility surface is the approach suggested by Dumas *et al.* [4] (1996), who

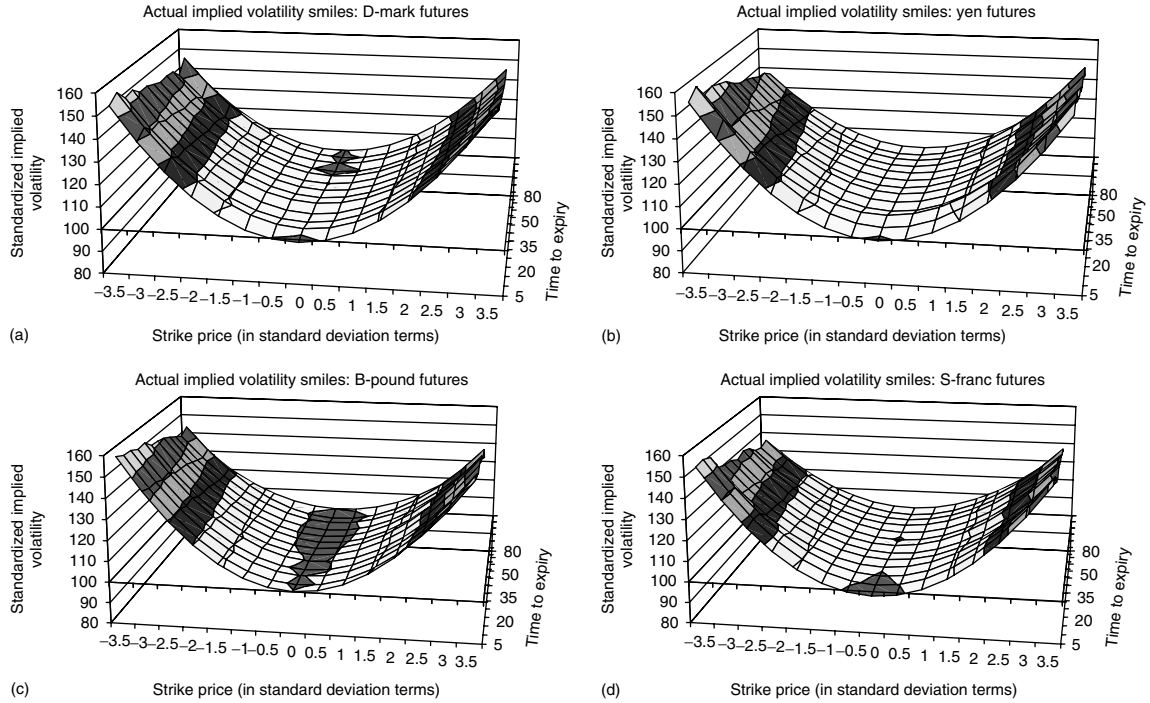


Figure 1 Actual implied volatility surfaces of option prices for four foreign exchange futures standardized to the level of the ATM volatility (1985–2000)

tested a number of arbitrary models based upon a polynomial expansion across strike price (x) and time (t). Tompkins [15] extended the polynomial expansion to degree three and included additional factors, which might also influence the behaviors of volatility surfaces.

For all four foreign exchange options markets, a parsimonious model explains the vast majority of the variance in the standardized implied volatility surfaces. The analysis allowed strike price effects to be separated into a first-order effect (the skew), a second-order effect (the smile), and higher order effects. For the skew effect, the results suggested that an asymmetrical smile pattern is a function of the level of the foreign exchange rate. The evidence suggests that when futures prices are low (high), the implied volatility pattern becomes more negatively (positively) skewed.

For the second-order “curved” pattern, all four markets display a convex pattern that becomes more extreme as the options expiration date is approached. Furthermore, a significant negative relationship is found between the degree of curvature and the level

of the ATM implied volatility. Curved patterns are independent of the level of the exchange rate. Finally, Tompkins [15] reports a significant third-order strike price effect for all four foreign exchange option markets. Tompkins [15] shows that the high degree of explanatory power is invariant to the time period of analysis and that the model provides accurate smile predictions outside of the estimation sample period. Under these assumptions, we conclude that regularities in implied volatility surfaces exist and are similar for the four currency markets. Furthermore, the regularities are time period invariant. These general results provide means to test alternative models, which could potentially explain why implied volatility surfaces exist. This is discussed in the following section.

Empirical Regularities for Currency Option Smiles

From [15], the following general conclusions can be drawn for the behaviors of implied volatility surfaces

for options on foreign exchange:

1. Implied volatility patterns are symmetrical on average for options on currencies.
2. For three of the four markets, the skew effect is related to the level of the underlying futures price. The only exception is for the British pound/US dollar. The level of the futures price impacts the skewness in an inverse manner to the pure skewness effect. This suggests that for low futures prices a negative skew occurs and at higher futures prices the skew flattens and can become positive.
3. The skew effect for currency options is relatively invariant to the time to expiration of the options. It is solely due to extreme levels of the underlying exchange rate or to some market shock.
4. For two of the four markets, the level of the skew effect is inversely related to the level of the ATM implied volatility. For the Deutsche mark and Swiss franc, the higher (lower) the level of the ATM implied volatility, the more negative (positive) the level of the skew.
5. Shocks change the degree and sign of the skew effect. For the Deutsche mark and Swiss franc, the concerted intervention in the currency markets by the Group of Seven (G7) caused a negative skew to occur. The 1987 stock crash had minimal impact on the currency markets, with only a slightly negative skew impact for the Deutsche mark. For the second shock, the only currency option skew affected was the Japanese yen. This occurred in January 1988 and appears to have been associated with international capital flows out of the US dollar into yen.
6. All implied volatility patterns display some degree of curvature and the degree of curvature is inversely related to the option's term to expiration. The longer the term to expiration, the less extreme the degree of curvature in the smile.
7. Shocks change the degree of curvature of the implied volatility pattern. However, the effect is not systematic and often shocks reduce the degree of curvature. For the G7 intervention in 1985, there was a reduction in the degree of smile curvature for both the Deutsche mark and Swiss franc, while for the Japanese yen, this event caused greater curvature for the smiles. For the second shock, both the British pound and Japanese yen displayed greater smile curvature thereafter.
8. For all four markets, the degree of curvature of the implied volatility pattern is inversely related to the level of the ATM implied volatility. Thus, the higher the level of ATM implied volatility, the less pronounced the degree of curvature in the smile.
9. For three of the four currency markets, the degree of curvature is independent of the level of the underlying futures price. The only exception is for the Japanese yen, where the higher the level of the exchange rate, the lesser the curvature (however, this impact is small).
10. For all four markets, the degree of curvature of the implied volatility pattern is asymmetrical. For the Deutsche mark, Japanese yen, and Swiss franc, the degree of asymmetry is negative. This suggests that the curvature is more extreme for options with strike prices below the current level of the underlying futures. For the British pound, the relationship is positive, indicating that the curvature is more extreme for options with strike prices above the current level of the underlying futures.

Using these 10 stylized facts as clues, we now examine alternative explanations for the existence of implied volatility smiles. It is crucial that any coherent explanation must conform to all of these facts simultaneously. If selected models are internally inconsistent with these facts, it is grounds for rejection.

A nontrivial problem is that the statistical testing of any option-pricing model has to be a joint hypothesis that the option-pricing model is correct and that the markets are efficient. Given that smiles do exist, we can reject the hypothesis that actual option values conform to the Black [2] model. However, we are uncertain as to why this occurs. Consider two possible reasons for the existence of smiles: the underlying asset may follow an alternative price process or the Black [2] model is correct but market imperfections exist. The next sections discuss both possibilities to better understand the regularities in implied volatility surfaces presented in [15].

Models with Alternative Price and Volatility Processes

Consider first that some alternative price (and volatility) process is at work instead of geometric Brownian motion with constant variance. Following the general approach of Jarrow and Rudd [11], we consider alternative true terminal distributions for the underlying asset. Consider the following models that include stochastic volatility ($\hat{\sigma}$) and alternative price processes. For the sake of convenience, the volatility processes will be evaluated in terms of a stochastic variance process ($\sqrt{V}$). Given that our previous results examined options on futures, the notation indicates that the underlying asset is a futures price (F). The first model, which will be considered, is a stochastic volatility model: the square root process model proposed by Heston [7] (*see also Stochastic Volatility Models: Foreign Exchange; Heston Model*). This choice is due to the ability of this model to allow correlated underlying and volatility processes. This will be defined as

Model 1

$$dF(t) = \mu F(t) dt + \hat{\sigma} F(t) dZ_1(t) \quad (3)$$

with the variance process defined by

$$dV(t) = \kappa(\theta - V(t)) dt + \xi \sqrt{V(t)} dZ_2(t) \quad (4)$$

where Z_1 and Z_2 are standard Wiener processes with correlation ρ . The term κ indicates the rate of mean reversion of the variance, θ is the long-term variance, and ξ indicates the volatility of the variance. The terms V and $\sqrt{V}$ represent the variance and the volatility of the process, respectively.

The second model that we consider is the jump-diffusion model proposed by Merton [13] (*see also Jump-diffusion Models*). Using his notation, this can be expressed as

Model 2

$$dF(t) = F(\alpha - \lambda\kappa) dt + F\sigma(t) dZ(t) + dq(t) \quad (5)$$

Using his notation, α is the instantaneous expected return on the futures contract, $\sigma(t)$ is the instantaneous volatility of the futures contract, conditional on no arrivals of important new information (no

jumps), dZ is a standard Wiener process, $q(t)$ is the independent Poisson process, which captures the jumps. The term λ is the mean number of arrivals per unit time and κ represents the jumps size (which can also be a random variable).

Bates [1], Ho *et al.* [8], and Jiang [12] assumed that the volatility process is subordinated in a non-normal price process; this provides the inspiration for the third model (see [1] for tests of these models). In this spirit, the third proposed model is a variant of the Heston [7] model, proposed by Tompkins [16, 17], which includes jumps (as captured by a normal inverse Gaussian (NIG) process) in the underlying price process.

Model 3

$$dF(t) = \mu F(t-) dt + \hat{\sigma}(t) F(t-) dN(t) \quad (6)$$

with the variance process defined by

$$dV(t) = k(\theta - V(t)) dt + \xi \sqrt{V(t)} dZ(t) \quad (7)$$

where $N(t)$ is a purely discontinuous martingale corresponding to log returns driven by an NIG Lévy process (*see Normal Inverse Gaussian Model*). This model will be referred to as *normal inverse Gaussian stochastic volatility (NIGSV)* for the sake of convenience.

Smile Patterns Associated with the Proposed Models

Tompkins [17] discussed how parameters for each of these models could be estimated (under the physical measure) and the change of measure to allow risk neutral pricing. Of more interest to this article is the resulting smile behavior of each model. This can be seen in Figure 2 (restricted solely to the Deutsche mark/US dollar). Figure 2(a) shows the empirical smile patterns for Deutsche mark/US dollar from 1985 to 2000. Figure 2(b) shows the smile surface associated with the Heston [7] model. Figure 2(c) shows the smile surface associated with the jump-diffusion model of Merton [13]. Figure 2(d) represents the combination of stochastic volatility and jump processes (NIGSV model).

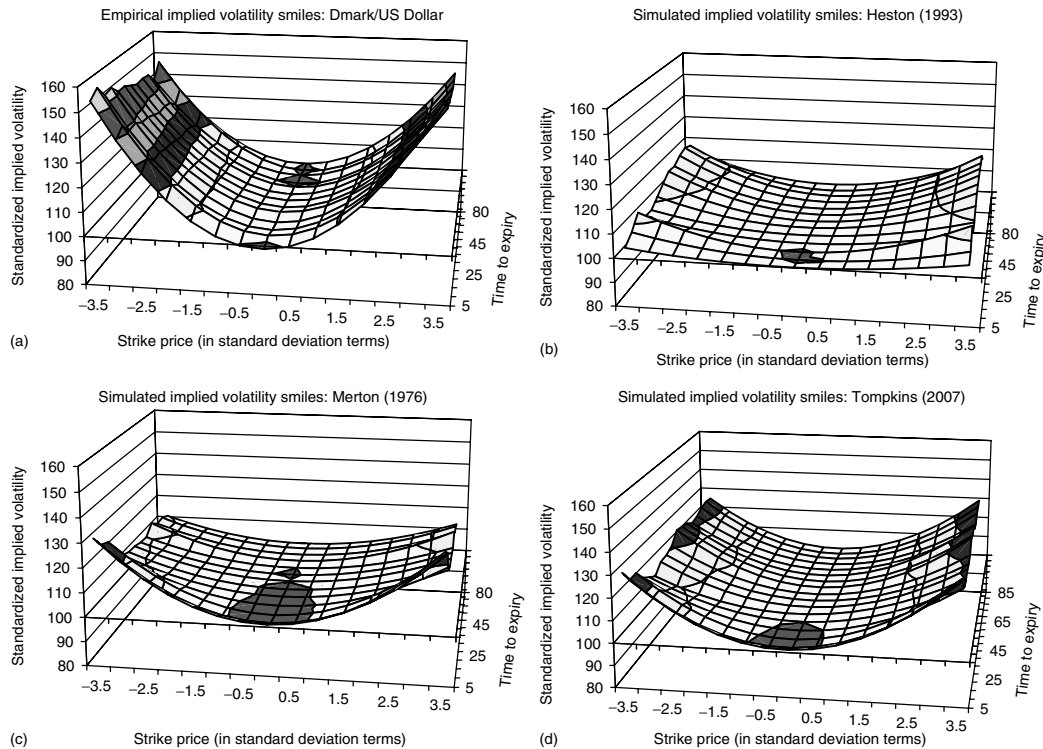


Figure 2 Simulated implied volatility smiles for options on Deutsche mark/US dollar

Smile Patterns Associated with Stochastic Volatility

As one can see in Figure 2(b), the Heston [7] model does generate a symmetrically curved smile function consistent with point #1, but the smiles are flat as the option expiration approaches and become more curved, the longer the term to expiration (which is inconsistent with point #6). This is exactly the opposite of what is observed for currency smiles empirically. The Heston [7] model can generate a skewed implied volatility pattern from a nonzero correlation between the volatility and underlying processes (see equations 3 and 4). However, the longer the term to expiration, the more extreme the skew pattern would be. This is inconsistent with point #3, that skewed patterns for currency options are time invariant and are only associated with the levels of the ATM implied volatility or the underlying currency exchange rate. However, this model is consistent with fact #5 that shocks could change the degree of skewness. The model could still be valid under a regime

of stochastic correlations. However, it seems inconsistent from an economic standpoint; if shocks change the degree of asymmetry in the expected terminal distribution of the underlying asset, it is not clear why in half of the instances the degree of curvature (fact #8) is reduced. This model is also inconsistent with fact #8, that the higher the level of expected variance (ATM volatility), the flatter the degree of curvature. Given that this model would produce effects that are contradictory to both first and second strike price effects observed empirically, we must reject it. An alternative explanation is that the jump-diffusion model of Merton [13] may be more appropriate.

Smile Patterns Associated with Stochastic Volatility

According to Hull [9], this model could produce a curved implied volatility surface and this curve would be consistent with fact #6, that curves exist and become more extreme the shorter the time to expiration of the option. This can be seen in Figure 2,

where the degree of curvature is most extreme closest to expiration. However, as the Poisson process in equation (5) is independent and identically distributed (i.i.d.), this will converge over time to a normal distribution and thus, the implied volatility surface would flatten, which is what occurs in Figure 2. It could also hold under a regime associated with fact #7, that shocks do change the degree of curvature. It could be that the inflow of new information changes the expectations of market agents regarding the degree and magnitude of future jumps. However, the model, as it stands, would not be able to explain the first-order strike price effects. One alternative would be to allow the shocks to be asymmetric. This would allow a skewed implied volatility pattern to exist. However, if the jumps follow some i.i.d. process, the central limit theorem would imply that the degree of skewness would be highest when the options are closest to expiration and would flatten as the term to expiration is lengthened. This is at variance with fact #3 that for currency options the skew effects are time invariant. Therefore, we can also reject a jump-diffusion model as being inconsistent with the empirical record.

Smile Patterns Associated with the NIGSV Model

This model assumes a symmetrical jump-diffusion process with a subordinated stochastic volatility process with nonzero correlations between the two processes. The simulated implied volatility smiles appear in Figure 2(d) and seem to resemble most the actual smiles for Deutsche mark/US dollars options in Figure 2(a). As can be seen, there is curvature in the smile patterns for both short term and longer term options. The shorter term curvature is associated with the jump process, while the longer term curvature is associated with stochastic volatility. This is consistent with both fact #1 and fact #6, that the average smile pattern is symmetrical and the degree of curvature is inversely related to time. Dynamics of the skew relationship can be explained with variations of the correlation between the two processes. Finally, the asymmetry of the smile shapes can be explained by the jump process. While this model appears to display many of the dynamics of empirical smiles, the degree of curvature is not as extreme as

is observed for the actual smiles. The reason for this is that the parameters for the model were estimated using the underlying Deutsche mark/US dollar currency futures (see [17] for details). While a feasible measure change was used to price options (that omitted arbitrage), it is unlikely that this measure change is unique as nontraded sources of risk have been introduced into the state space. These include jumps and stochastic volatility. Given this, we should expect that option prices will also contain some risk premium above and beyond the values associated with the underlying asset.

Conclusions and Implications

In this article, we have examined currency option smiles. Previous research by Tompkins [15] suggests that when implied volatility patterns are standardized, regularities are observed both across markets and across time. He concludes that this may suggest that market participants have developed some consistent algorithm to vary option prices in a consistent manner away from Black [2] values.

To better understand the nature of this algorithm, 10 stylized results are identified from his results for the four currency option markets. With these 10 results we test whether alternative models, which have been proposed to explain the existence of implied volatility surfaces, can generate the same dynamics as these empirical results. Initially, models were examined that suggest an alternative price process may better define the underlying price and volatility processes. We reject both the Heston [7] and the Merton [13] models as appropriate models, as they cannot produce all the empirical dynamics for actual smiles. The only model that could explain all the dynamics is a model that combines stochastic volatility and nonnormal innovations for currency returns. When appropriate parameters are input into this model and a feasible change of measure is made, option prices can be determined. The smiles associated with this model match the dynamics observed for actual currency option smiles. However, the model smiles do not display the same extreme degree of curvature as the empirical smiles. Following Tompkins [17], this suggests that a substantial risk premium exists for currency options and that the hypothesis that the existence of implied volatility surfaces are due solely to an alternative price process is rejected.

Alternatively, market imperfections may be the reason for the existence of implied volatility surfaces. Given that existing research has previously rejected this, we tend to concur that market imperfections alone are also probably not sufficient to explain the existence of implied volatility smiles. However, it is possible that both alternative price processes and market imperfections jointly contribute to the existence of implied volatility smiles.

References

- [1] Bates, D.S. (1996). jumps and stochastic volatility: exchange rate process implicit in Deutsche Mark options, *Review of Financial Studies* **9**, 69–107.
- [2] Black, F. (1976). The pricing of commodity contracts, *Journal of Financial Economics* **3**, 167–179.
- [3] Derman, E. & Kani, I. (1994). Riding on the smile, *Risk* **7**, 32–39.
- [4] Dumas, B., Fleming, J. & Whaley, R.E. (1998). Implied volatility functions: empirical tests, *The Journal of Finance* **53**, 2059–2106.
- [5] Dupire, B. (1992). *Arbitrage Pricing with Stochastic Volatility*, Working Paper, Société Générale Options Division.
- [6] Dupire, B. (1994). Pricing with a smile, *Risk* **7**, 18–20.
- [7] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**, 327–343.
- [8] Ho, M.S., Perraudin, W.R.M. & Sørensen, B.E. (1996). A continuous-time arbitrage-pricing model with stochastic volatility and jumps, *Journal of Business & Economic Statistics* **14**, 31–43.
- [9] Hull, J. (1997). *Options, Futures and other Derivative Securities*, 3rd Edition, Prentice Hall, Upper Saddle River.
- [10] Jackwerth, J.C. & Rubinstein, M. (1996). Recovering probability distributions from option prices, *The Journal of Finance* **51**, 1611–1631.
- [11] Jarrow, R. & Rudd, A. (1982). Approximate option valuation for arbitrary stochastic processes, *Journal of Financial Economics* **10**, 347–369.
- [12] Jiang, G. (1999). Stochastic volatility and jump-diffusion—implications on option pricing, *International Journal of Theoretical and Applied Finance* **2**(4), 409–440.
- [13] Merton, R. (1976). Option pricing when underlying stock returns are discontinuous, *Journal of Financial Economics* **3**, 125–144.
- [14] Rubinstein, M. (1994). Implied binomial trees, *The Journal of Finance* **49**, 771–818.
- [15] Tompkins, R.G. (2001). Implied volatility surfaces: uncovering regularities for options of financial futures, *The European Journal of Finance* **7**, 198–230.
- [16] Tompkins, R.G. (2003). Options on bond futures: isolating the risk premium, *Journal of Futures Markets* **23**(2), 169–215.
- [17] Tompkins, R.G. (2006). Why smiles exist in foreign exchange options: isolating components of the risk neutral process, *The European Journal of Finance* **12**, 583–604.

Further Reading

- Balyeat, R.B. (2002). The economic significance of risk premiums in the S&P 500 options market, *Journal of Futures Markets* **22**, 1145–1178.
- Garman, M. & Kohlhagen, S. (1983). Foreign currency option values, *Journal of International Money and Finance* **2**, 231–237.
- Henker, T. & Kazemi, H.B. (1998). The impact of deviations from random walk, in *Security Prices on Option Prices*, Working Paper, University of Massachusetts, Amherst.

Related Articles

Foreign Exchange Smile Interpolation; Implied Volatility Surface; Stochastic Volatility Models; Foreign Exchange.

ROBERT G. TOMPKINS

Foreign Exchange Smile Interpolation

This article provides a short introduction into the handling of FX-implied volatility market data—especially their inter- and extrapolation across delta space and time. We discuss a low-dimensional Gaussian kernel approach as the method of choice showing several advantages over usual smile interpolation methods such as cubical splines.

FX-implied Volatility

Implied volatilities for FX vanilla options are normally quoted against Black–Scholes deltas

$$\Delta_{BS} = e^{-r_f T} N \left(\frac{\ln(S/K) + (r_d - r_f + (\sigma^2(\Delta)/2)) T}{\sigma(\Delta)\sqrt{T}} \right) \quad (1)$$

Note that these deltas are dependent on $\sigma(\Delta)$, that is, the market-given volatility should be quoted. Thus, when retrieving a volatility for a given strike, an iterative processes is needed. However, under normal circumstances, the mapping from a delta-volatility to a strike-volatility coordinate system works *via* a quickly converging fixed point iteration.

Proposition 1 (Delta–Strike fixed point iteration). *Let*

$\Delta_n : A \mapsto A, A \subset (0, 1)$ *be a mapping, defined by*

$$\begin{aligned} \sigma_0 &= \sigma_{ATM} \\ \Delta_0 &= \Delta(K_{Call}, \sigma_{ATM}) \end{aligned}$$

$$\begin{aligned} \Delta_{n+1} &= e^{-r_f(T-t)} N(d_1(\Delta_n)) \\ &= e^{-r_f(T-t)} N \left(\frac{\ln(S/K) + (r_d - r_f + \sigma^2(\Delta_n)/2)(T-t)}{\sigma(\Delta_n)\sqrt{T-t}} \right) \end{aligned} \quad (2)$$

(3)

For sufficiently large $\sigma(\Delta_n)$ and a smooth, differentiable volatility smile, the sequence converges for $n \rightarrow \infty$ against the unique fixed point $\Delta^ \in A$ with $\sigma^* = \sigma(\Delta^*)$, corresponding to strike K .*

The usual FX smiles normally satisfy the above mentioned regularity conditions. More details concerning this proposition can be found in [5]. However, note that already smoothness is demanded here, which directly leads to the issue of an appropriate smile interpolation.

Interpolation

Before the discussion of specific interpolation methods, let us take a step backward and remember Rebonato’s well-known statement of implied volatility as the wrong number in the wrong formula to obtain the right price [3]. Therefore, the explanatory power of implied volatilities for the dynamics of a stochastic process remains limited. Implied volatilities give a lattice on which marginal distributions can be constructed. However, even using many data points to generate marginal distributions, forward distributions and extremal distributions, which determine the prices of some products such as compound and barrier products, cannot be uniquely defined by implied volatilities (see [4] for a discussion of this).

The attempt to capture FX smile features can lead to two different general approaches.

Parametrization

One possibility to express smile or skew patterns is just to capture it as the calibration parameter set of an arbitrary stochastic volatility or jump diffusion model that generates the observed market implied volatilities. However, as spreads are rather narrow in liquid FX options markets, it is preferred to exactly fit the given input volatilities. This automatically leads to an interpolation approach.

Pure Interpolation

As an introduction, we would like to pose four requirements for an acceptable volatility surface interpolation:

1. Smoothness in the sense of continuous differentiability. Especially with respect to the possible

application of Dupire-style local volatility models, it is crucial to construct an interpolation that is at least $\mathcal{C}^2$ in strike and at least $\mathcal{C}^1$ in time direction. This becomes obvious when considering the expression for the local volatility in this context:

$$\begin{aligned}\sigma_{loc}^2(S, K) &= \frac{\partial C(K, T)/\partial T + r_f C(K, T) + K(r_d - r_f)\partial C(K, T)/\partial K}{(1/2)K^2(\partial^2 C(K, T)/K^2)} \\ &= \frac{\frac{\sigma_i}{T} + 2(r_d - r_f)K(\partial\sigma_i/\partial K) + 2(\partial\sigma_i/\partial T)}{K^2 \left(\frac{1}{\sigma_i} \left(1/(K\sqrt{T} + d_+ \partial\sigma_i/\partial K) \right)^2 + \partial^2\sigma_i/\partial K^2 - d_+ \sqrt{T} (\partial\sigma_i/\partial K)^2 \right)}\end{aligned}\quad (4)$$

where $C(K, T)$ denotes the Black–Scholes prices of a call option with strike K , σ_i its corresponding implied volatility, and

$$d_+ = \frac{\ln(S/K) + (r_d - r_f + \sigma^2(\Delta)/2)T}{\sigma(\Delta)\sqrt{T}} \quad (5)$$

Note in addition that local volatilities can directly be extracted from delta-based FX volatility surfaces, that is, the Dupire formula can alternatively be expressed in terms of delta. See [2] for details.

2. Absence of oscillations, which is guaranteed if the sign of the curvature of the surface does not change over different strike or delta levels.
3. Absence of arbitrage possibilities on single smiles of the surface as well as absence of calendar arbitrage
4. A reasonable extrapolation available for the interpolation method.

A widely used classical interpolation method is cubical splines. They attempt to fit surfaces by fitting piecewise cubical polynomials to given data points. They are specified by matching their second derivatives at each intersection. Although this ensures the required smoothness by construction, it does not prevent oscillations, which directly leads to the danger of arbitrage possibilities or it does not define how to extrapolate the smile. We, therefore, introduce the concept of a slice kernel volatility surface as an alternative:

Definition 1 (Slice Kernel). Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be n given points and $g : \mathbb{R} \mapsto \mathbb{R}$ a smooth function which fulfills

$$g(x_n) = y_n, \quad n = 1, \dots, n \quad (6)$$

A smooth interpolation is then given by

$$g(x) := \frac{1}{\Phi_\lambda(x)} \sum_{i=1}^N \alpha_i K_\lambda(\|x - x_i\|) \quad (7)$$

where

$$\Phi_\lambda(x) := \sum_{i=1}^N K_\lambda(\|x - x_i\|) \quad (8)$$

and

$$K_\lambda(u) := \exp \left\{ -\frac{u^2}{2\lambda^2} \right\} \quad (9)$$

The described kernel is also called *Gaussian Kernel*. The interpolation reduces to determining the α_i , which is straightforward *via* solving a linear equation system. Note that λ remains as a free smoothing parameter, which also affects the condition of the equation system. At the same time, it can be used to fine-tune the extrapolation behavior of the kernel.

Generally, the slice kernel produces reasonable output smiles based on a maximum of seven delta-volatility points. Then it fulfills all the above-mentioned requirements. It is $\mathcal{C}^\infty$, does not create oscillations, passes typical no-arbitrage conditions as they are, for example, posed by Gatheral [1], and finally has an inherent extrapolation method.

In time direction, one might connect different slice kernels by linear interpolation of the variances for same deltas. This also normally ensures the absence of calendar arbitrage, for which a necessary condition

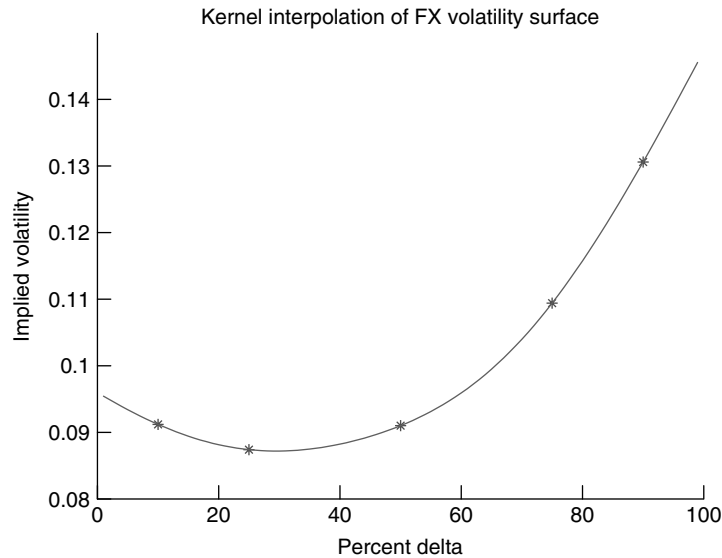


Figure 1 Kernel interpolation of an FX volatility surface

is a nondecreasing variance for constant moneyness F/K (see also [1] for a discussion of this).

Figure 1 displays the shape of a slice kernel applied to a typical FX volatility surface constructed from 10 and 25 delta volatilities, and the ATM volatility (in this example $\lambda = 0.25$ was chosen).

References

- [1] Gatheral, J. (2004). *A Parsimonious Arbitrage-free Implied Volatility Parameterization with Application to the Valuation of Volatility Derivatives*, Workshop Presentation, Madrid.
- [2] Hakala, J. & Wystup, U. (2002). Local volatility surfaces—tackling the smile, *Foreign Exchange Risk*, Risk Books.

- [3] Rebonato, R. (1999). *Volatility and Correlation*, John Wiley & Sons.
- [4] Tistaert, J., Schoutens, W. & Simons, E. (2004). A perfect calibration now what? *Wilmott Magazine* (March), 66–78.
- [5] Wystup, U. (2006). *FX Options and Structured Products*, John Wiley & Sons.

Related Articles

Foreign Exchange Markets; Foreign Exchange Options: Delta- and At-the-money Conventions.

UWE WYSTUP

Margrabe Formula

An exchange option gives its owner the right, but not the obligation, to exchange b units of one asset for a units of another asset at a specific point in time, that is, it is a claim that pays off $(aS_1(T) - bS_2(T))^+$ at time T .

Outperformance option or Margrabe option are alternative names for the same payoff.

Let us assume that the interest rate is constant (r) and that the underlying assets follow correlated ($dW_1 dW_2 = \rho dt$) geometric Brownian motions under the risk-neutral measure,

$$dS_i = \mu_i S_i dt + \sigma_i S_i dW_i \quad \text{for } i = 1, 2 \quad (1)$$

Note that allowing μ_i 's that are different from r enables us to use resulting valuation formula for the exchange option directly in cases with nontrivial carrying costs on the underlying. This could be for futures (where the drift rate is 0), currencies (where the drift rate is the difference between domestic and foreign interest rates, *see Foreign Exchange Options*), stocks with dividends (where the drift rate is r less the dividend yield), or nontraded quantities with convenience yields.

The value of the exchange option at time t is

$$\pi^{\text{EO}}(t) = \text{EO}(T - t, aS_1(t), bS_2(t)) \quad (2)$$

where the function EO is given by

$$\text{EO}(\tau, S_1, S_2) = S_1 e^{(\mu_1 - r)\tau} \mathcal{N}(d_+) - S_2 e^{(\mu_2 - r)\tau} \mathcal{N}(d_-) \quad (3)$$

with

$$d_{\pm} = \frac{\ln(S_1/S_2) + (\mu_1 - \mu_2 \pm \sigma^2/2)\tau}{\sigma\sqrt{\tau}} \quad (4)$$

where $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho}$, $\mathcal{N}$ denotes the standard normal distribution function and $\tau = T - t$. The formula was derived independently by Margrabe [12] and Fisher [6], but despite the two papers being published side by side in the *Journal of Finance*, the formula commonly bears only the former author's name. The result is most easily proven by using a

change of numeraire (*see Change of Numeraire*), writing

$$\pi^{\text{EO}}(t) = S_2(t) \mathbf{E}_t^{Q_{S_2}}((aS_1(T)/S_2(T) - b)^+) \quad (5)$$

noting that S_1/S_2 follows a geometric Brownian motion, and reusing the Black–Scholes calculation for the mean of a truncated lognormal variable.

If the underlying asset prices are multiplied by a positive factor, then the exchange option's value changes by that same factor. This means that we can use Euler's homogeneous function theorem to read off the partial derivatives of the option value with respect to the underlying assets (the deltas) directly from the Margrabe formula (see [15] for more such tricks), specifically

$$\frac{d\text{EO}}{dS_1} = e^{(\mu_1 - r)\tau} \mathcal{N}(d_+) \quad (6)$$

and similarly for S_2 . If the S assets are traded, then a portfolio with these holdings (scaled by a and b) that is made self-financing with the risk-free asset replicates the exchange option, and the Margrabe formula gives *the only* no-arbitrage price.

If the underlying assets do not pay dividends during the life of the exchange option (so that the risk-neutral drift rates are $\mu_1 = \mu_2 = r$), then early exercise is never optimal, and the Margrabe formula holds for American options too. With nontrivial carrying costs, this is not true, but as noted by [2], a change of numeraire reduces the dimensionality of the problem so that standard one-dimensional methods for American option pricing can be used.

The Margrabe formula is still valid with stochastic interest rates, provided the factors that drive interest rates are independent of those driving the S assets.

Exchange options are most common in over-the-counter foreign exchange markets, but exchange features are embedded in many other financial contexts; mergers and acquisitions (see [12]) and indexed executive stock options (see [9]) to give just two examples.

Variations and Extensions

Some variations of exchange options can be valued in closed form. In [10], a formula for a so-called traffic light option that pays

$$(S_1(T) - K_1)^+(S_2(T) - K_2)^+ \quad (7)$$

is derived, and [4] gives a formula for the value of a compound exchange option, that is, a contract that pays

$$(\pi^{\text{EO}}(T_C) - S_2(T_C))^+ \text{ at time } T_C < T \quad (8)$$

Both formulas involve the bivariate normal distribution function, and in the case of the compound exchange option a nonlinear but well-behaved equation that must be solved numerically.

For knock-in and knockout exchange options whose barriers are expressed in terms of the ratio of the two underlying assets, [7] show that the reflection-principle-based closed-form solutions (see [14]) from the Black-Scholes model carry over; this means that barrier option values can be expressed solely through the EO-function evaluated at appropriate points.

However, there are not always easy answers; in the simple case of a spread option

$$(S_1(T) - S_2(T) - K)^+ \quad (9)$$

there is no commonly accepted closed-form solution. The reason for this is that a sum of lognormal variables is not lognormal. More generally, many financial valuation problems can be cast as follows: calculate the expected value of

$$\left(\sum_{i=1}^n \alpha_{i,n} X_{i,n} - K \right)^+ \quad (10)$$

where the $X_{i,n}$'s are lognormally distributed. One can use generic techniques such as direct integration, numerical solution of partial differential equations, or Monte Carlo simulation, but there is an extensive literature on other approximation methods. These include

- moment approximation, where the moments of $\sum_{i=1}^n \alpha_{i,n} X_{i,n}$ are calculated, the variable then treated as lognormal, and the option priced by a Black-Scholes-like formula; an application to Asian options is given in [11].
- integration by Fourier transform techniques, which extends beyond lognormal models and works well if n is not too large (say 2–4); an application to spread options is given in [1].
- limiting results for $n \rightarrow \infty$ as obtained in [5] and [13]; the relation to the reciprocal gamma

distribution has been used for Asian and basket options.

- changing to Gaussian processes as suggested in [3]; this may be suitable for commodity markets where spread contracts are popular, and it allows for the inclusion of mean reversion.
- if the $\alpha_{i,n} X_{i,n}$'s depend monotonically on a common random variable, then Jamshidian's approach from [8] can be used to decompose an option on a portfolio into a portfolio of simpler options. This is used to value options on coupon-bearing bonds in one-factor interest-rate models.

References

- [1] Alexander, C. & Scourse, A. (2004). Bivariate normal mixture spread option valuation, *Quantitative Finance* **4**, 637–648.
- [2] Bjerksund, P. & Stensland, G. (1993). American exchange options and a put-call transformation: a note, *Journal of Business, Finance and Accounting* **20**, 761–764.
- [3] Carmona, R. & Durrleman, V. (2003). Pricing and hedging spread options, *SIAM Review* **45**, 627–685.
- [4] Carr, P. (1988). The valuation of sequential exchange opportunities, *Journal of Finance* **43**, 1235–1256.
- [5] Dufresne, D. (2004). The log-normal approximation in financial and other computations, *Advances in Applied Probability* **36**, 747–773.
- [6] Fischer, S. (1978). Call option pricing when the exercise price is uncertain, and the valuation of index bonds, *Journal of Finance*, **33**, 169–176.
- [7] Haug, E.G. & Haug, J. (2002). Knock-in/out Margrabe, *Wilmott Magazine* **1**, 38–41.
- [8] Jamshidian, F. (1989). An exact bond option formula, *Journal of Finance* **44**, 205–209.
- [9] Johnson, S.A. & Tian, Y.S. (2001). Indexed executive stock options, *Journal of Financial Economics* **57**, 35–64.
- [10] Jørgensen, P.L. (2007). Traffic light options, *Journal of Banking and Finance* **31**, 3698–3719.
- [11] Levy, E. (1992). Pricing European average rate currency options, *Journal of International Money and Finance* **11**(5), 474–491.
- [12] Margrabe, W. (1978). The value of an option to exchange one asset for another, *Journal of Finance* **33**, 177–186.
- [13] Milevsky, M.A. & Posner, S.E. (1998). Asian options, the sum of lognormals, and the reciprocal gamma distribution, *Journal of Financial and Quantitative Analysis* **33**, 409–422.

- [14] Poulsen, R. (2006). Barrier options and their static hedges: simple derivations and extensions, *Quantitative Finance* **6**, 327–335.
- [15] Reiss, O. & Wystup, U. (2001). Efficient computation of options price sensitivities using homogeneity and other tricks, *Journal of Derivatives* **9**, 41–53.

Related Articles

Black–Scholes Formula; Change of Numeraire; Exchange Options; Foreign Exchange Options.

ROLF POULSEN

Foreign Exchange Options: Delta- and At-the-money Conventions

In financial markets, the value of a plain-vanilla European option is generally quoted in terms of its implied volatility, that is, the volatility that, when plugged into the Black–Scholes formula, gives the correct market price. By observation of market prices the implied volatility, however, turns out to be a function of the option’s strike, thus giving rise to the so-called *volatility smile*.

In foreign exchange (FX) markets, it is common practice to quote volatilities for FX call and put options in terms of their delta sensitivities rather than in terms of their strikes or their moneyness. Volatilities and deltas are quoted by means of a table, the *volatility smile table*, consisting of rows for each FX option expiry date and columns for a number of delta values, as well as a column for the at-the-money (ATM) volatilities.

The definition and usage of a volatility smile table is complicated by the fact that FX markets have established various delta and ATM conventions. In this article, we summarize these conventions and highlight their intuition. For each delta convention, we give formulas and methods for the conversion of deltas to strikes and *vice versa*. We describe how to retrieve volatilities from the table for an arbitrary FX option that is to be priced in accordance with the information contained therein. We point out some mathematical problems and pitfalls when trying to do so and give criteria under which these problems surface.

Definitions

FX Rate

Before discussing the various delta conventions, we summarize some basic terms and definitions that we use in this article.

- *FX spot rate* $S(t)$: The FX spot rate $S(t)$ is the *current* exchange rate at the present time t

(“today”) between the domestic and the foreign currency. It is specified as the number of units of domestic currency that an investor gets in exchange for one unit of foreign currency,

$$S(t) := \frac{\text{number of units of domestic currency}}{\text{one unit of foreign currency}} \quad (1)$$

- *FX forward rate* $F(t, T)$: The FX forward rate $F(t, T)$ is the exchange rate between the domestic and the foreign currency at some *future* point of time T as observed at the present time t ($t < T$). It is again specified as the number of units of domestic currency that an investor gets in exchange for one unit of foreign currency at time T .

Using arbitrage arguments, spot and forward FX rates are related by (see, for instance, [3]):

$$F(t, T) = S(t) \cdot \frac{D_{\text{for}}(t, T)}{D_{\text{dom}}(t, T)} \quad (2)$$

where $D_{\text{for}} := D_{\text{for}}(t, T)$ is the foreign discount factor for time T (observed at time t) and $D_{\text{dom}} := D_{\text{dom}}(t, T)$ is the domestic discount factor for time T (observed at time t).

Note that the terminology in FX transactions is always confusing. In this article, we refer to the “domestic” currency in the sense of a base currency in relation to which “foreign” amounts of money are measured (see also [4]). By definition (1), an amount x in foreign currency, for example, is equivalent to $x \cdot S(t)$ units of domestic currency at time t .

In the markets, FX rates are usually quoted in a standard manner. For example, the USD–JPY exchange rate is usually quoted as the number of Japanese yen an investor receives in exchange for 1 USD. For a Japanese investor, the exchange rate would fit the earlier definition, while a US investor would either need to look at the reverse exchange rate $1/S(t)$ or think of Japanese yen as the “domestic” currency.

Value of FX Forward Contracts

When two parties agree on an FX forward contract at time s , they agree on the exchange of an amount of money in foreign currency at an agreed exchange rate

K against an amount of money in domestic currency at time $T > s$. When choosing $K = F(s, T)$, the FX forward contract has no value to either of the parties at time s .

As, in general, the forward exchange rate changes over time, at some time t ($s < t < T$), the FX forward contract will have a nonzero value (in domestic currency) given by

$$\begin{aligned} v_f(t, T) &= D_{\text{dom}}(F(t, T) - K) \\ &= S(t)D_{\text{for}} - K D_{\text{dom}} \end{aligned} \quad (3)$$

Value of FX Options

Upon deal inception, the holder of an FX option obtains the right to exchange a specified amount of money in domestic currency against a specified amount of money in foreign currency at an agreed exchange rate K . Assuming nonstochastic interest rates and the standard lognormal dynamics for the spot exchange rate, at time t , the domestic currency values of plain-vanilla European call and put FX options with strike K and expiry date T are given by their respective Black–Scholes formulas:

$$\begin{aligned} \text{Call option: } v_c(t, T) &= D_{\text{dom}} F(t, T) \mathcal{N}(d_+) \\ &\quad - D_{\text{dom}} K \mathcal{N}(d_-) \end{aligned} \quad (4)$$

$$\begin{aligned} \text{Put option: } v_p(t, T) &= (-1) \cdot [D_{\text{dom}} F(t, T) \mathcal{N}(-d_+) \\ &\quad - D_{\text{dom}} K \mathcal{N}(-d_-)] \end{aligned} \quad (5)$$

$$\begin{aligned} &= D_{\text{dom}} F(t, T) (\mathcal{N}(d_+) - 1) \\ &\quad - D_{\text{dom}} K (\mathcal{N}(d_-) - 1) \end{aligned} \quad (6)$$

where

$$d_{\pm} = \frac{\ln(F(t, T)/K) \pm 1/2\sigma^2\tau}{\sigma\sqrt{\tau}},$$

K : strike of the FX option,

σ : Black–Scholes volatility,

$\tau = T - t$: time to expiry of FX option, and

$\mathcal{N}(x)$: cumulative normal distribution function (7)

Note that $v_c(t, T)$ and $v_p(t, T)$ as given earlier are measured in *domestic* currency. The option position, however, may also be held in foreign currency.

We call the currency in which an option's value is measured as its *premium currency*.

Also note that the present value of call and put options are related by the *put–call parity*:

$$v_c - v_p = v_f = D_{\text{dom}}(F(t, T) - K) \quad (8)$$

Definition of Delta Types

This section summarizes the delta conventions used in FX markets and gives some of their properties. We outline the correspondence of each delta sensitivity with a particular delta hedge strategy the holder of an FX option chooses. FX options are peculiar in that the underlying coincides with the exchange rate. While, in general, it makes no sense to measure the value of an option in units of its underlying (e.g., in the number of shares of a company), the FX option position can be held either in domestic or in foreign currency. This gives rise to the premium-adjusted deltas.

For the ease of notation, we drop the time dependency of $S(t)$ and $F(t, T)$ in the following and denote the spot exchange rate as of time t by S and the forward exchange rate for time T as observed in t by F .

Unadjusted Deltas

Spot Delta.

Definition 1 For FX options, the *spot delta* is defined as the derivative of the option price $v_{c/p}$ with respect to the FX spot rate S :

$$\Delta_S^{c/p} := \frac{\partial v_{c/p}}{\partial S} \quad (9)$$

Interpretation Spot delta is the “usual” delta sensitivity that follows from the Black–Scholes equation. It can be derived by considering an FX option position that is held and hedged in *domestic* currency (see, for instance, [2, 3]).

Note that Δ_S is an amount in units of foreign currency. This makes sense from the hedging perspective: an amount of money in foreign currency is needed to make up for changes in the domestic currency value of an FX option (held in domestic currency), due to changes of the exchange rate. If,

for example, an investor is long an FX call option, a decrease in the exchange rate will lead to a decrease in his option position. By having shorted Δ_S units of foreign currency, the investor will make a hedge profit in domestic currency balancing his losses in the option position to first order.

Properties

$$\text{Call option: } \Delta_S^c = D_{\text{for}} \mathcal{N}(d_+) \quad (10)$$

$$\begin{aligned} \text{Put option: } \Delta_S^p &= D_{\text{for}} \{ \mathcal{N}(d_+) - 1 \} \\ &= -D_{\text{for}} \mathcal{N}(-d_+) \end{aligned} \quad (11)$$

$$\text{Put-call delta parity: } \Delta_S^c - \Delta_S^p = D_{\text{for}} \quad (12)$$

Forward Delta

Definition 2 The forward delta Δ_F (also called driftless delta [4]) of an FX option is defined as the ratio of the option's spot delta and the delta of a long forward contract on the FX rate (where the forward price of the FX forward contract equals the strike of the FX option):

$$\Delta_F^{c/p} := \frac{\partial v_{c/p} / \partial S}{\partial v_f / \partial S} = \frac{\Delta_S^{c/p}}{\Delta_S^f} \quad (13)$$

Interpretation The forward delta is not simply the derivative of the option price formula with respect to the forward FX rate F . The rationale for the above-mentioned definition follows from the construction of a hedge portfolio using FX forward contracts as hedge instruments for the FX option position (both held in domestic currency). The forward delta gives the number of forward contracts that an investor needs to enter into to completely delta hedge his/her FX option position; Δ_F , therefore, is simply a number without units.

Properties

$$\text{Call option: } \Delta_F^c = \frac{D_{\text{for}} \cdot \mathcal{N}(d_+)}{D_{\text{for}}} = \mathcal{N}(d_+) \quad (14)$$

$$\begin{aligned} \text{Put option: } \Delta_F^p &= \frac{D_{\text{for}} \cdot (\mathcal{N}(d_+) - 1)}{D_{\text{for}}} \\ &= \mathcal{N}(d_+) - 1 = -\mathcal{N}(-d_+) \end{aligned} \quad (15)$$

Put-call delta parity:

$$\Delta_F^c - \Delta_F^p = 1 \quad (16)$$

Premium-adjusted Deltas

Spot Delta Premium Adjusted.

Definition 3 The premium-adjusted spot delta is defined as

$$\Delta_{S,pa}^{c/p} := S \cdot \frac{\partial}{\partial S} \left(\frac{v_{c/p}}{S} \right) = \Delta_S^{c/p} - \frac{v_{c/p}}{S} \quad (17)$$

Interpretation The definition of the premium-adjusted spot delta follows from an FX option position that is held in foreign currency, while being hedged in domestic currency.

While v is the option's value in domestic currency, $v/S(t)$ is the option's value converted to foreign currency (i.e., its premium currency) at time t . The term $\partial(v/S(t))/\partial S \cdot dS$, thus, gives the change of the option value (measured in foreign currency) with the underlying exchange rate. To complete a delta hedge in domestic currency, the derivative needs to be multiplied by $S(t)$ from where the defining equation (17) for the premium-adjusted spot delta follows.

Note that the delta sensitivity is equal to spot delta, adjusted for the premium effect $v/S(t)$. This is easily interpreted as a delta that is corrected by a premium amount already paid in foreign currency. Also note that $\Delta_{S,pa}$ itself is denominated in units of foreign currency.

Properties

Call option:

$$\Delta_{S,pa}^c = D_{\text{for}} \frac{K}{F} \mathcal{N}(d_-) \quad (18)$$

Put option:

$$\begin{aligned} \Delta_{S,pa}^p &= D_{\text{for}} \frac{K}{F} \{ \mathcal{N}(d_-) - 1 \} \\ &= -D_{\text{for}} \frac{K}{F} \{ \mathcal{N}(-d_-) \} \end{aligned} \quad (19)$$

Put-call delta parity:

$$\begin{aligned} \Delta_{S,pa}^c + \Delta_{S,pa}^p &= D_{\text{for}} \frac{K}{F} (\mathcal{N}(d_-) - \mathcal{N}(-d_-)) \\ &= 2\Delta_{S,pa}^c - D_{\text{for}} \frac{K}{F} \end{aligned} \quad (20)$$

$$\Delta_{S,pa}^c - \Delta_{S,pa}^p = D_{\text{for}} \frac{K}{F} \quad (21)$$

The defining equations for premium-adjusted deltas have interesting consequences: while put deltas are unbounded and strictly monotonous functions of K , call deltas are bounded (i.e., $\Delta_S^c \in [0; \Delta_{\max}^c]$ with $\Delta_{\max}^c < 1$) and are *not* monotonous functions of K . Thus, the relationship between call deltas and strikes K is not one to one.

Forward Delta Premium Adjusted.

Definition 4 *The premium-adjusted forward delta is defined in analogy to the unadjusted delta:*

$$\Delta_{F,pa}^{c/p} = \frac{S \cdot \partial / \partial S (v_{c/p} / S)}{\partial v_f / \partial S} = \frac{\Delta_{S,pa}^{c/p}}{\partial v_f / \partial S} \quad (22)$$

Interpretation The intuition behind the definition of the premium-adjusted forward delta follows from an FX option position that is held in *foreign* currency and hedged by forward FX contracts in *domestic* currency. The premium-adjusted forward delta gives the number of forward contracts that are needed for the delta hedge in *domestic* currency of an FX option held in *foreign* currency. The derivation of the defining equation (22) is similar to the one for spot delta premium adjusted (cf. the section Spot Delta Premium Adjusted). Note that $\Delta_{F,pa}$ is a pure number without units.

Properties

Call option:

$$\Delta_{F,pa}^c = \frac{K}{F} \mathcal{N}(d_-) \quad (23)$$

Put option:

$$\Delta_{F,pa}^p = \frac{K}{F} \cdot \{\mathcal{N}(d_-) - 1\} = -\frac{K}{F} \mathcal{N}(-d_-) \quad (24)$$

Put-call delta parity:

$$\begin{aligned} \Delta_{F,pa}^c + \Delta_{F,pa}^p &= \frac{K}{F} (\mathcal{N}(d_-) - \mathcal{N}(-d_-)) \\ &= 2\Delta_{F,pa}^c - \frac{K}{F} \end{aligned} \quad (25)$$

$$\Delta_{F,pa}^c - \Delta_{F,pa}^p = \frac{K}{F} \quad (26)$$

Also note the important remarks in the previous section on the domain, the range of values, and the relationship between call delta and option strike. They apply likewise to premium-adjusted forward deltas.

Definition of At-the-money Types

In this section, we summarize the various ATM definitions, comment on their financial interpretation, and give the relations between all relevant quantities in Table 1.

Table 1 Strike values and delta values at the ATM point for the different FX delta conventions^(a)

	Δ -neutral	$K_{\text{atm}} = F$	Vega/gamma = max	$\Delta = 50\%$
ATM strike values				
Spot delta	$F \, \text{e}^{1/2\sigma^2\tau}$	F	$F \, \text{e}^{1/2\sigma^2\tau}$	$F \, \text{e}^{1/2\sigma^2\tau}$
Fwd delta	$F \, \text{e}^{1/2\sigma^2\tau}$	F	$F \, \text{e}^{1/2\sigma^2\tau}$	
Spot delta p.a.	$F \, \text{e}^{-1/2\sigma^2\tau}$	F	$F \, \text{e}^{1/2\sigma^2\tau}$	
Fwd delta p.a.	$F \, \text{e}^{-1/2\sigma^2\tau}$	F	$F \, \text{e}^{1/2\sigma^2\tau}$	
ATM delta values				
Spot delta	$D_{\text{for}} \, \mathcal{N}(0)$	$D_{\text{for}} \, \mathcal{N}\left(\frac{1}{2}\sigma\sqrt{\tau}\right)$	$D_{\text{for}} \, \mathcal{N}(0)$	$\mathcal{N}(0)$
Fwd delta	$\mathcal{N}(0)$	$\mathcal{N}\left(\frac{1}{2}\sigma\sqrt{\tau}\right)$	$\mathcal{N}(0)$	
Spot delta p.a.	$D_{\text{for}} \, \text{e}^{-1/2\sigma^2\tau} \mathcal{N}(0)$	$D_{\text{for}} \, \mathcal{N}\left(-\frac{1}{2}\sigma\sqrt{\tau}\right)$	$D_{\text{for}} \, \text{e}^{+1/2\sigma^2\tau} \mathcal{N}\left(-\frac{1}{2}\sigma\sqrt{\tau}\right)$	
Fwd delta p.a.	$\text{e}^{-1/2\sigma^2\tau} \mathcal{N}(0)$	$\mathcal{N}\left(-\frac{1}{2}\sigma\sqrt{\tau}\right)$	$\text{e}^{+1/2\sigma^2\tau} \mathcal{N}\left(-\frac{1}{2}\sigma\sqrt{\tau}\right)$	

^(a)Note that $\mathcal{N}(0) = 1/2$. The delta values are given for call options, the corresponding values for put options can be obtained, by replacing $\mathcal{N}(x)$ with $(\mathcal{N}(x) - 1)$

ATM Definition “Delta Neutral”

Definition 5 The ATM point is defined as the strike K_{atm} , for which the delta of a call and a put option add up to zero:

$$\Delta_x^c(K_{\text{atm}}, \sigma_{\text{atm}}) + \Delta_x^p(K_{\text{atm}}, \sigma_{\text{atm}}) = 0 \quad (27)$$

Here, x represents any of the delta conventions defined in the section Definition of Delta Types.

Interpretation The definition follows directly from a “straddle” position where a long call and a long put option with the same strike are combined. If the strike is chosen appropriately, the change in value of the call and the put option compensate (to first order) when the underlying FX rate changes. The straddle position’s value, thus, is insensitive (“delta neutral”) to changes in the underlying FX rate. The reason for this choice is that traders can use straddles to hedge the vega of their position without upsetting the delta.

Properties The ATM definition mentioned earlier for delta neutral FX options is equivalent to $\mathcal{N}(d_+) = 1/2$ in case of the unadjusted delta conventions and $\mathcal{N}(d_-) = 1/2$ in case of the premium-adjusted delta conventions. From this, the relationships of Table 1 follow in a straightforward manner.

ATM Definition via Forward

Definition 6 The ATM point is defined as the strike equaling the forward exchange rate:

$$K_{\text{atm}} := F \quad (28)$$

Interpretation This definition reflects the view that (given the information at deal inception) an option is ATM when its strike is chosen equal to the expected exchange rate at option expiry. If the spot exchange rate, indeed, approached F as $t \rightarrow T$ (as would be the case in a fully deterministic world by arbitrage arguments, cf. equation (2)), then the ATM strike would mark the dividing point between options that expire in-the-money (ITM) and out-of-the-money (OTM). From the put–call parity (8), we see that this is also the strike at which put and call options have the same value. Thus, this ATM definition is also called *value parity* [4].

Properties The relationships of Table 1 can again be derived in a straightforward manner from the definitions.

ATM Definition “vega = max”

Definition 7 The ATM point is defined as the strike K_{atm} for which vega of the FX option is at its maximum. Vega is the sensitivity of the FX option, with respect to the implied volatility of the underlying exchange rate. It is given by (cf. [4])

$$\text{vega}_{c/p} = \frac{\partial v_{c/p}}{\partial \sigma} = S D_{\text{for}} \sqrt{\tau} n(d_+) \quad (29)$$

where $n(x)$ is the normal density distribution.

The ATM strike can be derived from $\partial \text{vega} / \partial K = 0$ as

$$K_{\text{atm}} = F e^{1/2\sigma^2\tau} \quad (30)$$

Properties Table 1 again summarizes the relevant quantities for this ATM definition. Note that in case of unadjusted deltas, this ATM definition is equivalent to the delta neutral ATM definition. This is, however, not the case for adjusted deltas.

ATM Definition “ $\gamma = \max$ ”

Definition 8 The ATM point is defined as the strike K_{atm} for which the gamma sensitivity of the FX option is at its maximum.

We restrict the discussion to the case of gamma spot,

$$\gamma_S^{c/p} := \frac{\partial \Delta_S^{c/p}}{\partial S} = D_{\text{for}} \frac{n(d_+)}{S \sigma \sqrt{\tau}} \quad (31)$$

From $\partial \gamma / \partial K = 0$, the ATM strike can be derived as

$$K_{\text{atm}} = F e^{1/2\sigma^2\tau} \quad (32)$$

thus revealing the equivalence to the ATM definition “vega = max”.^a

ATM Definition “50%”

Definition 9 According to this convention, the ATM point is defined by

$$\Delta^c = 0.5 \quad \text{and} \quad |\Delta^p| = 0.5 \quad (33)$$

This condition can only be true for the forward delta convention and, thus, does not apply to any of the other delta conventions.

Properties of ATM Definitions

Table 1 summarizes the properties of the ATM point for all possible combinations of ATM definitions and delta conventions.

In this context, it is interesting to note that, beside their financial interpretation, mathematically, the various definitions of the ATM point lead to three characteristic relationships between the strike K and the forward exchange rate F :

$$K = F \Rightarrow d_{\pm} = \pm \sigma \sqrt{\tau} \quad (34)$$

$$K = F e^{1/2\sigma^2\tau} \Rightarrow d_+ = 0, \quad d_- = -\sigma \sqrt{\tau} \quad (35)$$

$$K = F e^{-1/2\sigma^2\tau} \Rightarrow d_+ = \sigma \sqrt{\tau}, \quad d_- = 0 \quad (36)$$

Converting Deltas to Strikes and Vice Versa

Quoting volatilities as a function of the options' deltas rather than as a function of the options' strikes brings about a problem when it comes to pricing FX options. Consider the case that we want to price a vanilla European option for a given strike K . To price this option, we have to find the correct volatility. As the volatility is given in terms of delta and delta itself is a function of volatility *and* strike, we have to solve an implicit problem, which, in general, has to be done numerically.

The following sections outline the algorithms that can be used to that end for the various delta conventions and directions of conversion. As the spot and forward deltas differ only by constant discount factors, we restrict the presentation to the forward versions of the adjusted and unadjusted deltas.

Forward Delta

For unadjusted deltas, there are simple one-to-one relationships between put and call deltas, the put-call delta parities (12) and (16). PUT deltas can therefore easily be translated into the corresponding CALL deltas and it is sufficient to perform all the calculations for call deltas here.

Conversion of Forward Delta to Strike. If volatilities are given as a function of forward delta, the strike corresponding to a given forward delta Δ_F^c can be calculated analytically. Let $\sigma(\Delta_F^c)$ denote the volatility associated with Δ_F^c ; $\sigma(\Delta_F^c)$ may either be quoted or interpolated from the volatility smile table.

With Δ_F^c and $\sigma(\Delta_F^c)$ given, we can directly solve equation (14),

$$\Delta_F^c = \mathcal{N}(d_+) = \mathcal{N}\left(\frac{\ln(F/K) + 1/2\sigma(\Delta_F^c)^2 T}{\sigma(\Delta_F^c)\sqrt{T}}\right) \quad (37)$$

for the strike K . We get

$$K = F \exp\left\{-\sigma(\Delta_F^c)\sqrt{T}\mathcal{N}^{-1}(\Delta_F^c) + \frac{1}{2}\sigma(\Delta_F^c)^2 T\right\} \quad (38)$$

Conversion of Strike to Forward Delta. The reverse conversion from strikes to forward deltas is more difficult and can only be achieved numerically. The following algorithm can be shown to converge [1] and has empirically proven to be very efficient.

In the first step, calculate a zero-order guess Δ_0 by using the ATM volatility σ_{atm} in equation (14):

$$\Delta_0 = \Delta_F^c(K, \sigma_{\text{atm}}) = \mathcal{N}\left(\frac{\ln(F/K) + 1/2\sigma_{\text{atm}}^2 T}{\sigma_{\text{atm}}\sqrt{T}}\right) \quad (39)$$

In the second step, use this zero-order guess for delta to derive a first-order guess σ_1 for the volatility by **interpolating** the curve $\sigma(\Delta)$:

$$\sigma_1 = \sigma(\Delta_0) \quad (40)$$

Finally, calculate the corresponding first-order guess for Δ_F^c in the third step:

$$\Delta_1 = \Delta_F^c(K, \sigma_1) = \mathcal{N}\left(\frac{\ln(F/K) + 1/2\sigma_1^2 T}{\sigma_1\sqrt{T}}\right) \quad (41)$$

and repeat steps two and three until the changes in Δ from one iteration to the next are below the specified accuracy.

Forward Delta Premium Adjusted

For the sake of clarity, we again restrict the discussion below to the case of call deltas. The algorithms for put deltas work analogously.

Conversion of Forward Delta Premium Adjusted to Strike. The conversion from forward delta premium adjusted to an absolute strike is more complicated than in the case of an unadjusted delta and cannot be formulated in a closed-form expression. The reason is that in the equation for the premium-adjusted call delta (23),

$$\begin{aligned}\Delta_{F,pa}^c &= \frac{K}{F} \mathcal{N}(d_-) \\ &= \frac{K}{F} \mathcal{N}\left(\frac{\ln(F/K) - 1/2\sigma(\Delta_{F,pa}^c)^2 T}{\sigma(\Delta_{F,pa}^c) \sqrt{T}}\right) \quad (42)\end{aligned}$$

the strike K appears inside and outside the cumulative normal distribution function so that one cannot solve directly for K . Even though both $\Delta_{F,pa}^c$ and $\sigma(\Delta_{F,pa}^c)$ are given, the problem has to be solved numerically.

So when converting a given call delta $\Delta_{F,pa}^c$, a root finder has to be used to solve for the correspondent strike K . This could, for example, be a simple bisection method where, for call deltas, the ATM-strike K_{atm} is the lower bound and some high (i.e., quasi infinite) value such as $100 K_{\text{atm}}$ can be used as upper bound. Of course, a more elaborate root finder to solve this problem could (and should!) be used, but a discussion of the various methods lies beyond the scope of this article.

Note, however, that all these methods require that Δ is a strictly monotonous function of K . We will see in section Ambiguities in the Conversion from Δ to Strike for Premium-adjusted Deltas that this is not always the case.

Conversion of Strike to Forward Delta Premium Adjusted. The conversion of an absolute strike to forward delta premium adjusted can be done analogously to the conversion into an unadjusted forward delta as described in section Conversion of Strike to Forward Delta.

First, use the previous guess Δ_{i-1} to obtain an improved guess σ_i for the volatility by interpolating

in the $\sigma(\Delta)$ curve.

$$\sigma_i = \sigma(\Delta_{i-1}) \quad (43)$$

In the second step, calculate the corresponding guess Δ_i by

$$\Delta_i = \Delta_{F,pa}^c(K, \sigma_i) = \frac{K}{F} \mathcal{N}\left(\frac{\ln(F/K) - 1/2\sigma_i^2 T}{\sigma_i \sqrt{T}}\right) \quad (44)$$

Iterate these two steps until the change in Δ from one step to the next is below the specified accuracy. A good initial guess for the volatility is, of course, again the ATM-volatility σ_{atm} .

Using the Volatility Smile Table

In FX markets, linear combinations of plain-vanilla FX options, such as “strangles”, “risk reversals”, and “butterflies”, are liquidly traded. These instruments are composed of ATM and OTM plain-vanilla put and call options at specific values of delta (typically, 0.25 or 0.1). When aggregating this market information, one obtains a scheme called the **volatility smile table** consisting of rows for each FX option expiry date, a column for ATM volatilities, and two distinct sets of columns for volatilities of OTM put and call options. We call these sets the *put and call* sides of the table.

Thus, OTM options can be priced by retrieving volatilities from the respective side of the volatility smile table, for example, OTM calls are priced using volatilities from the call side. By virtue of the put–call parity, ITM options can be priced using volatilities from the opposite side of the table, that is, ITM calls are priced using volatilities from the put side.

To exemplify this, consider the case of an option with arbitrary time to maturity and strike. The typical procedure to retrieve this option’s volatility from the smile table would be the following.

1. Determine the volatilities of the ATM point and the call and put sides at the option’s expiry. For this, the volatilities of each delta column will, in general, have to be interpolated in time.
2. Decide which side of the table to use depending on the option’s strike K : options with $K >$

K_{atm} are either OTM calls or ITM puts and are therefore priced using volatilities from the call side. Accordingly, options with $K < K_{\text{atm}}$ are priced using volatilities from the put side (cf. equation (4)).

3. Convert the option's strike to delta. This depends on the side of the table chosen in the previous step: convert to a call delta if $K > K_{\text{atm}}$, and to a put delta if $K < K_{\text{atm}}$. See the section Converting Deltas to Strikes and *vice versa* for the details of these conversions.
4. Retrieve the volatility from the table by interpolating volatilities in delta.

Alternatively, one could also translate the full smile volatility table from deltas to strikes. The conversion of deltas to strikes would then be necessary for the grid points of the table only, thus, making steps 2 and 3 of the earlier listed procedure obsolete. The interpolation in step 4 would be done in strikes. However, keep in mind that the strike grid points would vary from row to row.

It is important to note that the earlier procedure is based on the assumption that delta is a strictly monotonous function of the option's strike K : only in this case the option's delta and strike are equivalent measures of the option's moneyness, that is, only in this case we are guaranteed the equivalence

$$K > K_{\text{atm}} \iff \Delta^c < \Delta_{\text{atm}}^c \quad (45)$$

In the following section, we will show that the assumption of monotonicity is not always true and derive the conditions under which it is violated.

Problems and Pitfalls

Interpolation in Time Dimension when Delta Conventions Change

In FX markets, it is common to switch delta conventions with an option's time to maturity. For example, volatilities for options with less than two years to maturity are often quoted in terms of spot deltas, whereas volatilities for longer expiries are usually quoted in terms of one of the forward delta conventions.

When conventions change from one expiry T_k to the next one, it is *a priori* unclear how an interpolation in time on a delta-volatility table should

be performed. Which of the two conventions shall be used for options with expiries t within the range $T_k < t < T_{k+1}$?

Some possibilities are as follows: convert the grid points at T_k into the convention used for T_{k+1} and do interpolation in the long-term convention or, conversely, convert the grid points at T_{k+1} into the convention used for T_k and interpolate in the short-term convention. Another possibility would be to translate the delta grid into a strike grid and do the interpolation on strikes.

None of these approaches is *a priori* superior to the others. In real life, however, a choice has to be made. Even though the differences may be small, one should be aware that the choice is arbitrary.

ATM Delta Falling in the OTM Range

For long times to maturity, the delta of an ATM FX option may become smaller than the delta of the closest OTM option. In a sense, the ATM delta "crosses" the nearest OTM delta. In this case, it is unclear how the interpolation of volatilities (in delta) should be done or whether the ATM point or the crossed delta point should possibly be ignored.

The conditions under which this problem occurs vary with the delta and ATM types. In the following, we outline the derivation of these conditions for an exemplarily chosen combination (premium-adjusted spot deltas and forward ATM definition), summarize the results for other combinations, and finally discuss a typical numerical example for long-dated FX options.

Exemplary Derivation. We start from equation (17) for the ATM delta using $K_{\text{atm}} = F$:

$$\Delta_{S, pa}^c(F, \sigma_{\text{atm}}) = D_{\text{for}} \mathcal{N}\left(-\frac{\sigma_{\text{atm}}}{2}\sqrt{T}\right) \quad (46)$$

Obviously, the ATM delta will decrease with decreasing values of D_{for} and *increasing* values of $\sigma_{\text{atm}}\sqrt{T}$. Small values of D_{for} and/or large values of $\sigma_{\text{atm}}\sqrt{T}$ will therefore lead to ATM deltas that are smaller than the nearest OTM point Δ_1^c (which is usually $\Delta_1^c = 0.25$).

From the expression mentioned earlier, it follows immediately that the ATM delta is larger than the first

Table 2 Restrictions on discount factors and ATM volatilities for various combinations of delta and ATM types^(a)

	ATM-type forward	ATM-type delta neutral
Spot delta	$D_{\text{for}} > -2\Delta_1^p$ $\sigma_{\text{atm}} \sqrt{T} < 2\mathcal{N}^{-1} \left(1 + \frac{\Delta_1^p}{D_{\text{for}}} \right)$	$D_{\text{for}} > 2\Delta_1^c$
Forward delta	$\sigma_{\text{atm}} \sqrt{T} < 2\mathcal{N}^{-1} (1 + \Delta_1^p)$	No constraint
Spot delta p.a.	$D_{\text{for}} > 2\Delta_1^c$ $\sigma_{\text{atm}} \sqrt{T} < 2\mathcal{N}^{-1} \left(1 - \frac{\Delta_1^c}{D_{\text{for}}} \right)$	$D_{\text{for}} > 2\Delta_1^c$ $\sigma_{\text{atm}} \sqrt{T} < \sqrt{2 \ln \left(\frac{D_{\text{for}}}{2\Delta_1^c} \right)}$
Forward delta p.a.	$\sigma_{\text{atm}} \sqrt{T} < 2\mathcal{N}^{-1} (1 - \Delta_1^c)$	$\sigma_{\text{atm}} \sqrt{T} < \sqrt{2 \ln \left(\frac{1}{2\Delta_1^c} \right)}$

^(a)If these conditions are violated, the ATM point will cross the nearest OTM point. Note that the put delta Δ_1^p is negative

OTM point Δ_1^c only if

$$\sigma_{\text{atm}} \sqrt{T} < -2\mathcal{N}^{-1} \left(\frac{\Delta_1^c}{D_{\text{for}}} \right) = 2\mathcal{N}^{-1} \left(1 - \frac{\Delta_1^c}{D_{\text{for}}} \right) \quad (47)$$

This inequality has meaningful solutions only if the right-hand side is positive, that is, only if $1 - \Delta_1^c/D_{\text{for}} > 1/2$. We therefore have the additional constraint:

$$D_{\text{for}} > 2\Delta_1^c \quad (48)$$

The derivations for all other possible combinations are analogous. Their results are summarized in Table 2.

Numerical Example. To gain some intuition for these constraints, let us consider a numerical example for a relevant case. A typical combination of delta and ATM conventions for long-dated FX options is the ATM-type delta neutral for premium-adjusted forward deltas. Usually, the first OTM point is quoted at $\Delta_1^c = 0.25$. Inserting this value into the respective formula (see Table 2) yields the condition

$$\sigma_{\text{atm}} \sqrt{T} < \sqrt{2 \ln \left(\frac{1}{2\Delta_1^c} \right)} = \sqrt{2 \ln(2)} \approx 1.1774 \quad (49)$$

For a 30-year option, this condition restricts the ATM volatility to a maximum of 21.5%, a value that—*a priori*—does not seem unattainable.

Ambiguities in the Conversion from Δ to Strike for Premium-adjusted Deltas

Premium-adjusted deltas can cause further complications. Recall from the section Forward Delta Premium Adjusted

$$\Delta_F^c = \frac{K}{F} \mathcal{N}(d_-)$$

$$\text{where } d_- = \frac{\ln(F/K) - 1/2\sigma^2\tau}{\sigma\sqrt{\tau}} \quad (50)$$

and note that Δ_F^c is **not** a monotonous function of the strike K and therefore not invertible for all strikes as illustrated by Figure 1(a). Thus, while Δ_F^c can be calculated for any strike K , the reverse is not true: for any $\Delta_F^c < \Delta_{\text{max}}^c$, there are two strikes $K_1 \neq K_2$ with $\Delta_F^c(K_1) = \Delta_F^c(K_2)$.

In case $K_{\text{max}} < K_{\text{atm}}$, there is no problem. $\Delta_{F,pa}^c(K)$ is a monotone function of K for all $K > K_{\text{atm}}$ and is therefore directly related to the option's moneyness: the smaller a call option's delta value, the higher its strike and the deeper it is OTM. If desired, the volatility smile table defined in terms of delta can be translated uniquely into a smile table in strikes. Besides, when retrieving volatilities from the call side of the volatility

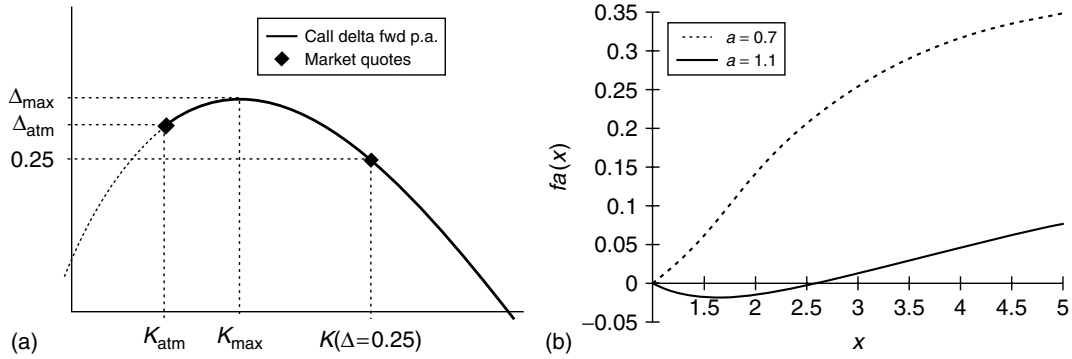


Figure 1 (a) Premium adjusted (forward) call delta as a function of strike. (b) The function $f_\alpha(x)$ as defined in equation (56) for $\alpha = 0.7$ (broken line) and $\alpha = 1.1$ (solid line)

smile table, this can be done by direct interpolation of the entries, possibly including the ATM point.

In case $K_{\max} > K_{\text{atm}}$, however, the situation is more complicated. $\Delta_{F,pa}^c(K)$ is no longer a monotonous function of K and the conversion of deltas to strikes is no longer unique for all OTM options. Thus, when translating a smile table in deltas into a smile table in strikes, particular care has to be taken. In addition, when retrieving the volatility for options with strikes $K \in (K_{\text{atm}}; K_{\max})$ from the volatility smile table, one has to *extrapolate* volatilities in delta beyond the ATM point which seems odd.

Note that the counterintuitive *extrapolation* in delta beyond the ATM point does not occur in case the volatility smile table in deltas is translated to a table in strikes on the grid points first (see the section Using the Smile Volatility Table) and the interpolation is done in strikes. This is possible as long as there is no crossing of the ATM point and the closest delta grid point as discussed in the section ATM Delta Falling in the OTM Range.

In the following discussion, we show that the case $K_{\max} > K_{\text{atm}}$ can, indeed, occur. We restrict ourselves to premium adjusted *forward* call deltas and ATM-type delta neutral. The conditions for the ATM-type forward can be derived analogously and we summarize the results at the end of this section. Similar arguments hold for premium-adjusted spot deltas, however, it is reasonable to assume that the problems outlined in the section ATM Delta Falling in the OTM Range surface beforehand.

Starting from the expressions for the ATM strike and the call delta, see Table 1,

$$K_{\text{atm}} = F e^{-1/2\sigma_{\text{atm}}^2 T} \quad (51)$$

$$\Delta_{\text{atm}}^c = \Delta_{F,pa}^c(K_{\text{atm}}, \sigma_{\text{atm}}^2) = \frac{1}{2} e^{-1/2\sigma_{\text{atm}}^2 T} \quad (52)$$

we need to find conditions under which there is a second solution with strike $K > K_{\text{atm}}$ that solves

$$\Delta_{\text{atm}}^c = \Delta_{F,pa}^c(K, \sigma(\Delta_{F,pa}^c)) \quad (53)$$

This may seem a difficult problem at first glance since the delta on the right-hand side depends on the volatility, which itself is given in terms of delta. It is, however, simplified considerably by the following argument: while we do not know the strike for the point that solves equation (53), we do know the delta there: it is the ATM delta. As the volatility is given in terms of delta, we also know the volatility on the right-hand side of equation (53): it must be the ATM-volatility σ_{atm} .

Therefore, the problem can be reformulated as follows: when does

$$\Delta_{\text{atm}}^c = \Delta_{F,pa}^c(K, \sigma_{\text{atm}}) \quad (54)$$

have a second solution with $K > K_{\text{atm}}$ (besides the trivial one at $K = K_{\text{atm}}$)?

Inserting the expressions for the ATM delta and the premium-adjusted forward call delta into equation

(54) we get

$$\frac{1}{2} e^{-1/2 \sigma_{\text{atm}}^2 T} = \frac{K}{F} \mathcal{N} \left(\frac{\ln(F/K) - 1/2 \sigma_{\text{atm}}^2 T}{\sigma_{\text{atm}} \sqrt{T}} \right) \quad (55)$$

With the definitions

$$\begin{aligned} \alpha &:= \sigma_{\text{atm}} \sqrt{T}, \\ x &:= \frac{K}{F} e^{1/2 \sigma_{\text{atm}}^2 T} = \frac{K}{F} e^{\alpha^2/2} \end{aligned}$$

the problem is equivalent to finding the roots of

$$f_{\alpha}(x) := e^{-\alpha^2/2} \left\{ \frac{1}{2} - x \mathcal{N} \left(-\frac{\ln(x)}{\alpha} \right) \right\} \quad (56)$$

The condition $K > K_{\text{atm}}$ in the previous equation corresponds to $x > 1$. Plotting $f_{\alpha}(x)$ for various values of α , see Figure 1(b), we make the following observation: for small values of α the function increases monotonically, taking on only positive values for $x > 1$. For larger values of α , the function decreases at first, thus, taking on negative values in a certain range, and then increases again, eventually reaching a second zero.

Therefore, the question reduces further to this: under which conditions does the function $f_{\alpha}(x)$ have a negative slope at $x = 1$? The first derivative of f_{α} can easily be calculated,

$$f'_{\alpha}(x) = -e^{-\alpha^2/2} \left\{ \mathcal{N} \left(-\frac{\ln(x)}{\alpha} \right) - \frac{1}{\alpha} \mathcal{N}' \left(-\frac{\ln(x)}{\alpha} \right) \right\} \quad (57)$$

and the slope of f_{α} at $x = 1$ is now readily obtained as

$$\begin{aligned} f'_{\alpha}(x=1) &= -e^{-\alpha^2/2} \left\{ \mathcal{N}(0) - \frac{1}{\alpha} \mathcal{N}'(0) \right\} \\ &= -e^{-\alpha^2/2} \left\{ \frac{1}{2} - \frac{1}{\alpha} \frac{1}{\sqrt{2\pi}} \right\} \end{aligned} \quad (58)$$

The constraint that it has to be positive yields the condition

$$\alpha = \sigma_{\text{atm}} \sqrt{T} < \sqrt{\frac{2}{\pi}} \quad (59)$$

which restricts the ATM volatility for a 30-year option to only 14.6%.

In other words, for a 30-year option, with ATM volatilities larger than 14.6%, the conversion between $\Delta_{F, pa}$ and strikes becomes ambiguous.

A similar condition can be derived for the forward ATM type. The major difference is that the function $f_{\alpha}(x)$ takes on a different form that ultimately yields the following expression for the turnover point α at which the conversion becomes ambiguous:

$$\mathcal{N} \left(-\frac{\alpha}{2} \right) = \frac{1}{\alpha} \mathcal{N}' \left(-\frac{\alpha}{2} \right) \quad (60)$$

This equation can be solved numerically and yields the constraint

$$\alpha = \sigma_{\text{atm}} \sqrt{T} < 1.224 \quad (61)$$

so that for a 30-year option, ATM volatilities above 22.3% will lead to ambiguities.

End Notes

^a. It should be noted that equation (32) was obtained under the assumption that the slope of the volatility σ as a function of the strike K in equation (31) is 0 ATM. This is necessary as otherwise this ATM definition would become impracticable. In general, however, the volatility smile implies a nonzero slope and the maximum value of γ will not be found at the strike given by equation (32).

References

- [1] Borowski, B. (2005). *Hedgingverfahren für Foreign Exchange Barrieroptionen*, Diploma Thesis, Technical University of Munich.
- [2] Carr, P. & Bandyopadhyay, A. (2000). *How to derive the Black-Scholes Equation Correctly?* <http://faculty.chicagogsb.edu/akash.bandyopadhyay/research/> (accessed Mar 2000).
- [3] Hull, J.C. (1997). *Options, Futures and Other Derivative Securities*, 3rd Edition, Prentice Hall, NJ.
- [4] Wystup, U. (2006). *FX Options and Structured Products*, Wiley.

Related Articles

Foreign Exchange Markets; Foreign Exchange Symmetries; Foreign Exchange Smile Interpolation.

CLAUS CHRISTIAN BEIER & CHRISTOPH
RENNER

Stochastic Volatility Models: Foreign Exchange

In finance, the term *volatility* has a dilative meaning. There exists a definition in the statistical sense, which states that volatility is the standard deviation per unit time (usually per year) of the (logarithmic) asset returns. However, empirical evidence about derivatives markets shows that refinements of this definition are necessary.

First, one can observe a dependency of Black–Scholes-implied volatility on at least the strike price and time to maturity implicit in option prices. This dependency defines the **implied volatility surface**. One possibility to incorporate this dependency into a model is by using a deterministic function for the instantaneous volatility, that is, the volatility governing the changes in spot returns in infinitesimal time steps. This function of spot price and time is called *local volatility* (see **Local Volatility Model**).

In addition, empirical evidence indicates that the local volatility surface is not constant over time, but is subject to changes. This is not surprising since the expectations of the market participants with respect to the future instantaneous volatility might change over time. Furthermore, for some derivative products, the dynamics of the volatility surface is crucial for a reliable valuation. A prominent example is the product class of cliquet options, which are basically a collection of forward start options with increasing forward start time. The payoff depends on the absolute or the relative performance of the underlying during the lifetime of the options. It is intuitive that an option with a forward start time of one year, for instance, depends substantially on the one-year forward volatility surface. Since this volatility is uncertain, it seems advisable to model this risk by an additional stochastic factor.

Stochastic Volatility FX Models

Now, we review some common procedures for incorporating a stochastic behavior of volatility into foreign exchange (FX) rate models. The first approach

is to model the FX rate X_t with a volatility process v_t by a system of stochastic differential equations, like

$$\begin{aligned} dX_t &= \mu_t X_t dt + \sigma_t X_t dW_t^X \\ dv_t &= \eta(v_t) dt + \zeta(v_t) dW_t^v \end{aligned} \quad (1)$$

where $v_t = f(\sigma_t)$ for a function f and the increments of the Brownian motions W^X and W^v are possibly correlated. Table 1 gives an overview of some common stochastic volatility models.

A general class of stochastic volatility models is formed by the affine jump diffusion models. They have been studied by Duffie *et al.* [5]. The Heston model is a special case of this kind of model.

The second approach is inspired by the observation that higher volatility comes along with an increased trading activity and *vice versa*. This is realized by a time change in the FX rate process. For instance, consider a standard Brownian motion $\{W_t\}_{t \geq 0}$ with variance t for W_t , that is, the value of the process after t units of physical time. Now, if the economic time elapses twice as fast as the physical time due to market activity, the process could be expressed by the deterministically time-changed process $\{W_{2t}\}_{t \geq 0}$. Hence, the variance for the values of the process after t units of physical time would then be $2t$. This idea can be generalized by representing the economic time as a stochastic process Y_t , which is named *stochastic clock*. For every realization of the process $\{Y_t\}_{t \geq 0}$, economic time must be a monotone function of physical time, that is, the process is a subordinator. Recently proposed models use a normal inverse gamma or variance gamma Lévy process for representing the exchange rate process and an integrated Cox–Ingersoll–Ross or Gamma–Ornstein–Uhlenbeck process for stochastic time. All these models have the common feature that the characteristic function of the logarithm of the time-changed exchange rate, $\ln X_{Y_t}$, can be expressed in closed form.

Besides the models mentioned here, there exist other modeling approaches. For a general overview of stochastic volatility models, see [6, 11] and for especially in the FX context see [12].

Heston’s Stochastic Volatility Model

In the following, we discuss the stochastic volatility model of Heston and its option valuation applied to

Table 1 Some common stochastic volatility models

$f(\sigma_t)$	$\eta(v_t)$	$\zeta(v_t)$	Reference
σ_t^2	$\kappa(\theta - v_t)$	ξv_t	GARCH similar diffusion model [11]
σ_t^2	$\kappa(\theta - v_t)$	$\xi \sqrt{v_t}$	Heston model [9]
σ_t^2	$\kappa(\theta v_t - v_t^2)$	$\xi v_t^{3/2}$	3/2 model [11]
$\ln \sigma_t^2$	$\kappa(\theta - v_t)$	ξ	Log-volatility Ornstein–Uhlenbeck [11]

an FX setting. The model is characterized by the stochastic differential equations:

$$\begin{aligned} dX_t &= (r_d - r_f)X_t dt + \sqrt{v_t}X_t dW_t^X \\ dv_t &= \kappa(\theta - v_t)dt + \sigma \sqrt{v_t}dW_t^v \end{aligned} \quad (2)$$

with $\text{Cov}[dW_t^X, dW_t^v] = \rho dt$. Here, the FX rate process $\{X_t\}_{t \geq 0}$ is modeled by a process, similar to the geometric Brownian motion, but with a non-constant instantaneous variance v_t . The variance process $\{v_t\}_{t \geq 0}$ is driven by a mean-reverting stochastic square-root process. The increments of the two Wiener processes $\{W_t^X\}_{t \geq 0}$ and $\{W_t^v\}_{t \geq 0}$ are assumed to be correlated with rate ρ . In an FX setting, the risk-neutral drift term of the underlying process is the difference between the domestic and the foreign interest rates $r_d - r_f$. The quantities $\kappa \geq 0$ and $\theta \geq 0$ denote the rate of mean reversion and the long-term variance. The parameter σ is often called *vol* of *vol*, but it should be called *volatility of instantaneous variance*.

The term $\sqrt{v_t}$ in equation (2) ensures a nonnegative volatility in the FX rate process. It is known that the distribution of values of $\{v_t\}_{t \geq 0}$ is given by a noncentral chi-squared distribution. Hence, the probability that the variance takes a negative value is equal to zero. Thus, if the process touches the zero bound, the stochastic part of the volatility process turns zero and the deterministic part will ensure a nonnegative volatility because of the positivity of κ and θ .

The Heston model is often not capable of fitting complicated structures of implied volatility surfaces. In particular, this is true if the term structure exhibits a nonmonotone form or the sign of the skew changes with increasing maturity. For a discussion of the implied volatility surface generated by this model, see [7]. One approach to tackle this limitation is to extend the original Heston model by time-dependent parameters [3, 14].

Valuation of Options in the Heston Model

For the valuation of options in the Heston model, we consider the value function of a general contingent claim $V(t, v, X)$. As shown in [8], applying Itô's lemma, the self-financing condition, and the possibility to trade in the underlying exchange rate, money market, and another option, which is dependent on time, volatility, and X , we arrive at Garman's partial differential equation:

$$\begin{aligned} \frac{\partial V}{\partial t} + \kappa(\theta - v) \frac{\partial V}{\partial v} + (r_d - r_f)X \frac{\partial V}{\partial X} + \frac{1}{2}\sigma^2 v \frac{\partial^2 V}{\partial v^2} \\ + \frac{1}{2}vX^2 \frac{\partial^2 V}{\partial X^2} + \rho\sigma vX \frac{\partial^2 V}{\partial v \partial X} - r_d V = 0 \end{aligned} \quad (3)$$

A solution to the above equation can be obtained by specifying appropriate boundary and exercise conditions, which depend on the contract specifications. In the case of European vanilla options, Heston [9] provided a closed-form solution, namely,

$$\text{Vanilla} = \phi \left[e^{-r_f \tau} X_t P_1 - K e^{-r_d \tau} P_2 \right] \quad (4)$$

where $\tau = T - t$ is the time to maturity, $\phi = \pm 1$ is the call–put indicator and K is the strike price. The quantities P_1 and P_2 define the probability that the exchange rate X at maturity is greater than K under the spot and the risk-neutral measure, respectively. The spot delta of the European vanilla option is equal to $\phi e^{-r_f \tau} P_1$.

Assuming that the distribution of $\ln X_T$ at time t under the two different measures is determined uniquely by its characteristic function φ_j , for $j = 1, 2$, it is shown, in [15], that P_1 and P_2 can be expressed in terms of the inverse Fourier transformation

$$P_j = \frac{1}{2} + \frac{1}{\pi} \phi \int_0^\infty \Re \left[\frac{\exp(-iu \ln K) \varphi_j(u)}{iu} \right] du \quad (5)$$

The integration in equation (5) can be done using numerical integration methods such as Gauss–Laguerre integration or fast Fourier transform approximation. In [10], it is shown that the computational time of the fast Fourier transform approach to compute vanilla option prices is higher

compared to a numerical integration method with certain caching techniques.

The characteristic function is exponentially affine and available in closed form as

$$\varphi_2(u) = \exp(B(u) + A(u)v_t + iu \ln X_t) \quad (6)$$

The functions A and B arise as the solution of the so-called Riccati differential equations as shown in [8]. They are defined as follows:

$$A(u) = -iu(1 - iu) \frac{(1 - e^{-d(u)\tau})}{\gamma(u)} \quad (7)$$

$$B(u) = iu(r_d - r_f)\tau + \frac{\kappa\theta}{\sigma^2}(\kappa - iu\rho\sigma - d(u))\tau + \frac{2\kappa\theta}{\sigma^2} \ln \frac{2d(u)}{\gamma(u)} \quad (8)$$

with $d(u) = \sqrt{(\rho\sigma iu - \kappa)^2 + iu\sigma^2(1 - iu)}$ and $\gamma(u) = d(u)(1 + e^{-d(u)\tau}) + (\kappa - iu\rho\sigma)(1 - e^{-d(u)\tau})$. The characteristic function φ_1 has the same form as the function φ_2 , but with u replaced by $u - i$ and multiplied by a factor $\exp(-(r_d - r_f)\tau - X_0)$. This is due to the change from the spot to the risk-neutral measure in the derivation of φ_1 .

There exist several different representations of the characteristic function φ . In some formulations of φ , the characteristic function can become discontinuous if the multivalued complex logarithm contained in the integrand is restricted to the calculation of its principal branch, as is the case in many implementations. Wrong results of the value P_j may occur unless a rotation count algorithm is employed. For other representations of φ , stability for all choices of model parameters can be proved. Details can be found in [1].

Besides vanilla options, closed-form solutions for exotic options have been found for the volatility option, the correlation option, the exchange option, the forward start option, the American option, the discrete barrier option, and others. Numerical pricing of exotic options in the Heston model can be carried out by using conventional numerical methods such as Monte Carlo simulation [2, 13], finite differences [8], or an exact simulation method [4].

Calibration of Heston's Model

We realize the estimation by fitting the Heston model parameters to the smile of the current vanilla option market. Thereby, the choice of the loss function to minimize the differences between the model and market Black–Scholes-implied volatilities is crucial. Here, we decide to do a least-squared error fit over absolute values of volatilities, rather than minimizing over relative volatilities or option values.

For a fixed time to maturity τ , given market-implied volatilities $\sigma_1^{\text{market}}, \dots, \sigma_n^{\text{market}}$ and corresponding spot delta (premium unadjusted) values $\Delta_1, \dots, \Delta_n$, the calibration is set up as follows:

1. Before starting the optimization, we determine the strikes K_i corresponding to σ_i^{market} with

$$K_i = X_0 \exp \left\{ -\phi N^{-1}(\phi e^{r_f \tau} \Delta_i) \sigma_i^{\text{market}} \sqrt{\tau} + \left(r_d - r_f + \frac{1}{2}(\sigma_i^{\text{market}})^2 \right) \tau \right\} \quad 1 \leq i \leq n \quad (9)$$

which requires the inversion of the cumulative normal distribution function N .

2. The aim is to minimize the objective function M defined below. We repeat the steps (a)–(c) until a certain accuracy in the optimization routine is achieved.
 - a. We use the analytic formula in equation (4) to calculate the vanilla option values in the Heston model for the strikes $K_1, \dots, K_n$:

$$H_i(\kappa, \sigma, \theta, \rho, v_0) = \text{Vanilla}(\kappa, \sigma, \theta, \rho, v_0, \text{market data}, K_i, \phi) \quad (10)$$

- b. For $i = 1, \dots, n$, we compute all option values H_i in terms of Black–Scholes-implied volatilities $\sigma_i^{\text{model}}(\kappa, \sigma, \theta, \rho, v_0)$ by applying a root search.

c. The objective function is given as

$$\begin{aligned} M(\kappa, \sigma, \theta, \rho, v_0) \\ = \sum_{i=1}^n w_i [\sigma_i^{\text{market}} - \sigma_i^{\text{model}}(\kappa, \sigma, \theta, \rho, v_0)]^2 \\ + \text{penalty} \end{aligned} \quad (11)$$

The implementation of a penalty function *penalty* and some weights w_i may give the calibration routine some additional stability. There exist various choices for the penalty function. For example, in [14], it is suggested to penalize the retraction from the initial set of model parameters, but we may also use the penalty to introduce further constraints such as the condition $2\kappa\theta - \sigma^2 > 0$ to ensure that in subsequent simulations the volatility process cannot sojourn in zero. In addition, we could use the weights w_i to favor at-the-money (ATM) or out-of-the-money fits.

For the minimization, a great variety of either local or global optimizers in multidimensions could be used. Algorithms of Levenberg–Marquardt type are frequently used, because they utilize the least-squares form of the objective function. Since the objective function is usually not convex, there may exist many local extrema and the use of a computationally more expensive global (stochastic) algorithm, such as simulated annealing or differential evolution, in the calibration routine may be considered. From a practical point of view, taking the value of a short-dated implied volatility as an initial value for v_0 is a good start for the calibration. In light of parameter stability, the result of the previous (the day before) calibration could be used as an initial guess for the remaining parameters. Furthermore, to enhance the speed of calibration, it is suggested in [8] to fix the model parameter κ and run the calibration in only four dimensions, since the influence of the mean reversion is often compensated by a stronger volatility of variance σ . To ensure that the correlation parameter ρ attains values in $[-1, 1]$, we reparametrize with the function $2 \arctan(\rho)/\pi$.

Hedging

If volatility is introduced as a stochastic factor but cannot be traded, the market is incomplete. By

the introduction of a tradable market instrument $U(t, v, X)$, which depends on the volatility, the market can be completed and volatility risk can be hedged dynamically. In the Heston model, to make a portfolio containing the contingent claim $V(t, v, X)$ instantaneously risk free, the hedge portfolio has to consist of Δ_X units of the foreign currency and Δ_U units of the contingent claim $U(t, v, X)$, with

$$\Delta_X = \frac{\partial V}{\partial X} - \Delta_U \frac{\partial U}{\partial X} \quad \text{and} \quad \Delta_U = \frac{\partial V / \partial v}{\partial U / \partial v} \quad (12)$$

Common FX market instruments applicable for market completion and hedging are, on the one hand, ATM forward plain vanilla options for different maturities. On the other hand, for most of the FX markets risk reversals (RR) and butterflies (BF) are traded for certain maturities and strikes. These instruments are defined as^a

$$\text{RR}(T, \Delta) = \text{Call}(T, K_\Delta) - \text{Put}(T, K_{-\Delta}) \quad (13)$$

$$\begin{aligned} \text{BF}(T, \Delta) = \frac{1}{2} (\text{Call}(T, K_\Delta) + \text{Put}(T, K_{-\Delta})) \\ - \text{Call}(T, K_{\text{ATM}}) \end{aligned} \quad (14)$$

where K_Δ is the strike as given in equation (9), such that the corresponding plain vanilla option has a Black–Scholes delta of Δ . K_{ATM} denotes the ATM strike, which is often taken to be the strike generating a zero delta for a straddle in FX markets. Risk reversals and butterflies are quoted in Black–Scholes-implied volatilities instead of prices, that is, if v_Δ denotes the implied volatility of a call with strike K_Δ (analogously, $v_{-\Delta}$ for puts and v_{ATM}), the FX smile quotes are

$$v_{\text{RR}} = v_\Delta - v_{-\Delta} \quad \text{and} \quad v_{\text{BF}} = \frac{1}{2}(v_\Delta + v_{-\Delta}) - v_{\text{ATM}} \quad (15)$$

Since individual ATM options are liquidly traded, and therefore v_{ATM} is known, the volatilities v_Δ and $v_{-\Delta}$ can be calculated from v_{RR} and v_{BF} .

The quantities v_{RR} and v_{BF} relate to the skew and smile, respectively, of the implied volatility surface. This is schematically illustrated for $\Delta = 0.25$ in Figure 1.

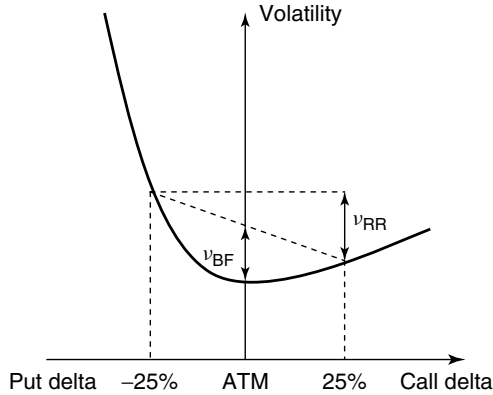


Figure 1 The meaning of v_{RR} and v_{BF} in the context of the implied volatility curve for a fixed maturity

Example

In this section, we give an example of the difference between Heston and Black–Scholes option prices. Thereby, we consider discrete down-and-out

put options written on the USD/JPY exchange rate. The option holder receives a vanilla put option payoff at maturity in 18 months as long as the FX rate does not fall below a given barrier B at the barrier fixing times in 6, 12, and 18 months. Otherwise, the option expires worthless. We compare ATM option prices for barriers of 10%, 50%, 60%, 70%, 80%, and 90% of the spot price.

For the valuation in the Heston model, we calibrate to European plain vanilla options with maturities of 1, 2, 3, 6, 9, 12, and 24 months and strikes with respect to 10% Δ put, 25% Δ put, ATM, 10% Δ call, and 25% Δ call for May 21, 2009 (Figure 2). The weights are set to 1 for ATM strikes, to 0.75 for 25% Δ strikes, and to 0.25 for 10% Δ strikes, since usually options with strikes far from the ATM forward are less liquidly traded. Figure 2(a) demonstrates that the market-implied volatilities (dots) are adequately matched by the calibrated model volatilities (circles connected by lines). Figure 2(b) shows the term structure of market-implied volatilities (dots) and calibrated implied volatilities (circles) for strikes with respect to 25% Δ call, ATM, and 25% Δ put (from

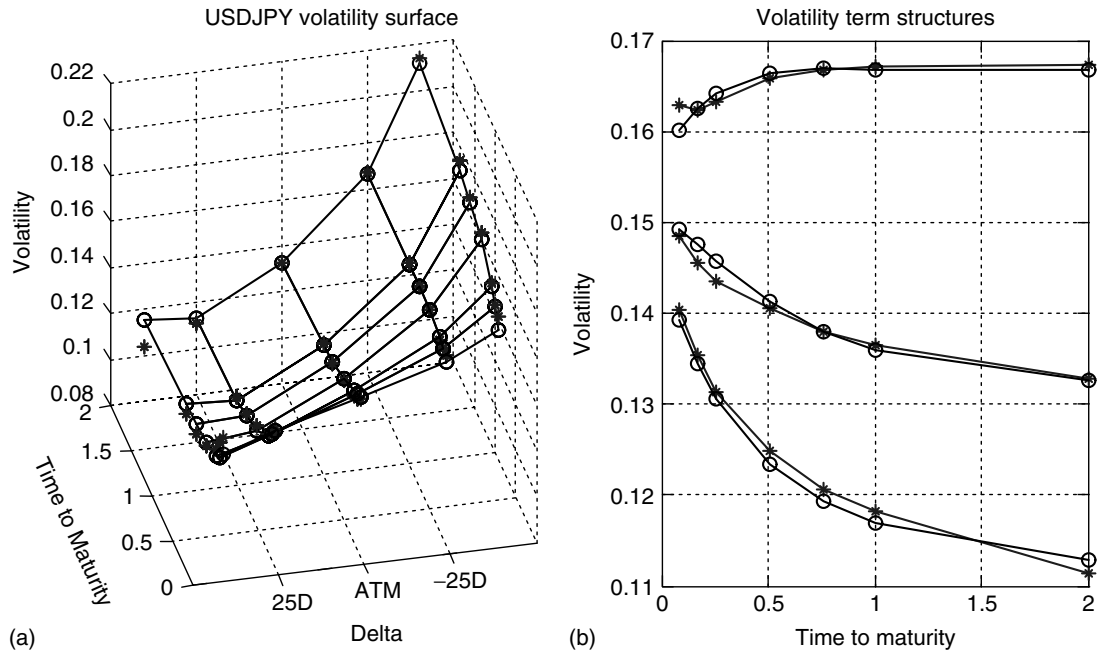


Figure 2 Implied volatilities of the Heston model fitted to market volatilities for USD/JPY with maturities of 1, 2, 3, 6, 9, 12, and 24 months and strikes for 10% Δ and 25% Δ put, ATM, 10% Δ , and 25% Δ call. The dots show the market volatilities and the circles the calibrated volatilities. (a) The whole volatility surface. (b) The implied volatility term structure for strikes 25% Δ call, ATM, and 25% Δ put (from bottom to top)

Table 2 Down-and-out put values with at-the-money strike and discrete monitoring at 6, 12, and 18 months

Barrier (% of spot)	90	80	70	60	50	10
Black–Scholes	0.8752	3.7068	5.7670	6.2663	6.3000	6.3012
Heston	0.6309	1.8723	3.1830	4.3376	5.2202	6.3023

The value of the corresponding plain vanilla put in the Heston model is given by 6.3060

bottom to top). The resulting model parameters are given by $\kappa = 0.2170$, $\theta = 0.0444$, $\sigma = 0.3410$, $\rho = -0.5927$, and $v_0 = 0.0228$.

As mentioned in the section Valuation of Options in the Heston Model, there exists a value function in (semi-) closed form for discrete barrier options in the Heston model. The distribution of the random variables $\ln X_{t_i}$ at the barrier fixing times t_i can be determined uniquely by the derivation of their joint characteristic function and with the application of Shephard's theorem given in [15] the required knock-out probabilities can be computed.

For the valuation in the Black–Scholes model, we use the interpolated ATM forward-implied volatility for plain vanilla options with a maturity of 18 months, which is $\sigma_{BS} = 0.1270$ in our example.

Finally, the FX spot trades at $X_0 = 94.43$, and the domestic and foreign interest rates for 18 months are given by $r_d = 0.0065$ and $r_f = 0.0139$. The resulting prices for the described options are shown in Table 2.

Comparing the prices, we can observe two effects. First, the Heston prices are lower than the corresponding Black–Scholes prices. This behavior might be in major part due to the fact that the Black–Scholes valuation uses a flat volatility, whereas the Heston model incorporates the whole volatility smile. Since the volatilities below the ATM strikes increase substantially (Figure 2), the knock-out probabilities also increase and the option prices drop. Second, the prices of Heston and Black–Scholes converge with decreasing barrier level. This appears reasonable, since the likelihood of a knock-out decreases more and more and the valuation finally results in put prices, which should be equal for both models in the case of a good calibration fit.

End Notes

^aThere exist different definitions for risk reversals and butterflyes in the literature with respect to sign and coefficients.

References

- [1] Albrecher, H., Mayer, P., Schoutens, W. & Tistaert, J. (2007). *The Little Heston Trap*, Wilmott No. 1, pp. 83–92.
- [2] Andersen, L. (2007). *Efficient Simulation of the Heston Stochastic Volatility Model*. Working paper. Available at SSRN: <http://ssrn.com/abstract=946405>
- [3] Benhamou, E., Gobet, E. & Miri, M. (2009). *Time Dependent Heston Model*. Working paper. Available at SSRN: <http://ssrn.com/abstract=1367955>.
- [4] Broadie, M. & Kaya, O. (2006). Exact simulation of stochastic volatility and other affine jump diffusion models, *Operations Research* **54**(2), 217–231.
- [5] Duffie, D., Singleton, K. & Pan, J. (2000). Transform analysis and asset pricing for affine jump-diffusions, *Econometrica* **68**, 1343–1376.
- [6] Fouque, J.P., Papanicolaou, G. & Sircar, K.R. (2000). *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press.
- [7] Gatheral, J. (2006). *The Volatility Surface*, Wiley.
- [8] Hakala, J. & Wystup, U. (2002). *Foreign Exchange Risk*, Risk Publications.
- [9] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**, 327–343.
- [10] Kilin, F. (2007). *Accelerating the Calibration of Stochastic Volatility Models*. Working Paper. Available at SSRN: <http://ssrn.com/abstract=965248>
- [11] Lewis, A.L. (2000). *Option Valuation under Stochastic Volatility*, Finance Press.
- [12] Lipton, A. (2001). *Mathematical Methods for Foreign Exchange*, World Scientific.
- [13] Lord, R., Koekkoek, R. & van Dijk, D. (2006). *A Comparison of Biased Simulation Schemes for Stochastic Volatility Models*. & Working Paper. Available at SSRN: <http://ssrn.com/abstract=903116>
- [14] Nögel, U. & Mikhailov, S. (2003). Heston's Stochastic Volatility Model. Implementation, Calibration and some Extensions, *Wilmott Juli*, 74–49.
- [15] Shephard, N.G. (1991). From characteristic function to distribution function, *Econometric Theory* **7**(4), 519–529. Cambridge University Press.

Related Articles

Model Calibration; Simulation of Square-root Processes; Stochastic Volatility Models.

Foreign Exchange Options; Foreign Exchange Smiles; Heston Model; Implied Volatility Surface;

SUSANNE A. GRIEBSCH & KAY F. PILZ

Foreign Exchange Basket Options

Quite often, corporate and institutional currency managers are faced with an exposure in more than one currency. Generally, these exposures would be hedged using individual strategies for each currency. These strategies are composed of spot transactions, forwards, and, in many cases, options on a single currency. Nevertheless, there are instruments that include several currencies, and these can be used to build a multicurrency strategy that is almost always cheaper than the portfolio of the individual strategies. As a leading example, we explain basket options in detail.^a

Basket Options

Basket options are derivatives based on a common base currency, say EUR and several other risky currencies. The option is actually written on the basket of risky currencies. Basket options are European options paying the difference between the basket value and the strike, if positive, for a basket call, or the difference between strike and basket value, if positive, for a basket put, respectively, at maturity. The risky currencies have different weights in the basket to reflect the details of the exposure.

For example, a basket call on two currencies USD and JPY pays off

$$\max \left(a_1 \frac{S_1(T)}{S_1(0)} + a_2 \frac{S_2(T)}{S_2(0)} - K; 0 \right) \quad (1)$$

at maturity T , where $S_1(t)$ denotes the exchange rate of EUR/USD and $S_2(t)$ denotes the exchange rate of EUR/JPY at time t , a_i the corresponding weights, and K the strike.

A basket option protects against a drop in both currencies at the same time. Individual options on each currency cover some cases, which are not protected by a basket option (shaded triangular areas in Figure 1) and that is why they cost more than a basket.

The ellipsoids connect the points that are reached with the same probability assuming that the forward prices are at the center.

Pricing Basket Options

Basket options should be priced in a consistent way with plain vanilla options. Hence the basic model assumption is a lognormal process for the individual correlated basket components. A decomposition into uncorrelated components of the exchange rate processes

$$dS_i = \mu_i S_i dt + S_i \sum_{j=1}^N \Omega_{ij} dW_j \quad (2)$$

is the basis for pricing. Here μ_i denotes the difference between the foreign and the domestic interest rate of the i th currency pair and dW_j the j th component of independent Brownian increments. The covariance matrix is given by $C_{ij} = (\Omega \Omega^T)_{ij} = \rho_{ij} \sigma_i \sigma_j$. Here σ_i denotes the volatility of the i th currency pair and ρ_{ij} the correlation coefficients.

Exact Method

Starting with the uncorrelated components, the pricing problem is reduced to the N -dimensional integration of the payoff. This method is accurate but rather slow for more than two or three basket components.

A Simple Approximation

A simple approximation method assumes that the basket spot itself is a lognormal process with drift μ and volatility σ driven by a Wiener Process $W(t)$:

$$dS(t) = S(t)[\mu dt + \sigma dW(t)] \quad (3)$$

with solution

$$S(T) = S(t)e^{\sigma W(T-t) + (\mu - 1/2\sigma^2)(T-t)} \quad (4)$$

given we know the spot $S(t)$ at time t . It is a fact that the sum of lognormal processes is not lognormal itself, but as a crude approximation, it is certainly a quick method that is easy to implement. To price the basket call, the drift and the volatility of the basket spot need to be determined. This is done by matching the first and second moment of the basket spot with the first and second moment of the lognormal model for the basket spot. The moments of lognormal spot are

$$\begin{aligned} E(S(T)) &= S(t)e^{\mu(T-t)} \\ E(S(T)^2) &= S(t)^2 e^{(2\mu + \sigma^2)(T-t)} \end{aligned} \quad (5)$$

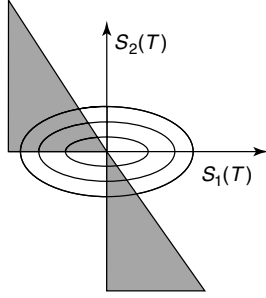


Figure 1 Basket-payoff and contour lines for probabilities

We solve these equations for the drift and volatility:

$$\begin{aligned}\mu &= \frac{1}{T-t} \ln \left(\frac{E(S(T))}{S(t)} \right) \\ \sigma &= \sqrt{\frac{1}{T-t} \ln \left(\frac{E(S(T)^2)}{E(S(T))^2} \right)}\end{aligned}\quad (6)$$

In these formulae we now use the moments for the basket spot:

$$\begin{aligned}E(S(T)) &= \sum_{i=1}^N \alpha_i S_i(t) e^{\mu_i(T-t)} \\ E(S(T)^2) &= \sum_{i,j=1}^N \alpha_i \alpha_j S_i(t) S_j(t) \\ &\quad \times e^{(\mu_i + \mu_j + \sum_{k=1}^N \Omega_{ki} \Omega_{jk})(T-t)}\end{aligned}\quad (7)$$

The pricing formula is the well-known Black–Scholes–Merton formula for plain vanilla call options:

$$\begin{aligned}v(0) &= e^{-r_d T} (F \mathbf{N}(d_+) - K \mathbf{N}(d_-)) \\ F &= S(0) e^{\mu T} \\ d_{\pm} &= \frac{1}{\sigma \sqrt{T}} \left(\ln \left(\frac{F}{K} \right) \pm \frac{1}{2} \sigma^2 T \right)\end{aligned}\quad (8)$$

Here $\mathbf{N}$ denotes the cumulative normal distribution function and r_d the domestic interest rate.

A More Accurate and Equally Fast Approximation

The previous approach can be taken one step further by introducing one more term in the Itô–Taylor

expansion of the basket spot, which results in

$$\begin{aligned}v(0) &= e^{-r_d T} (F \mathbf{N}(d_1) - K \mathbf{N}(d_2)) \\ F &= \frac{S(0)}{\sqrt{\eta}} e^{(\mu - \lambda/2 + (\lambda \sigma^2/2(\eta)))T} \\ d_2 &= \frac{\sigma - \sqrt{\sigma^2 + \lambda \left((1 + (\lambda/\eta)) \sigma^2 T - 2 \ln(F \sqrt{\eta}/K) \right)}}{\sqrt{T} \lambda} \\ d_1 &= \sqrt{\eta} d_2 + \frac{\sigma \sqrt{T}}{\sqrt{\eta}}\end{aligned}\quad (9)$$

where $\eta = 1 - \lambda T$

The new parameter λ is determined by matching the third moment of the basket spot and the model spot. For details, see [1].

Most remarkably, this major improvement in the accuracy only requires a marginal additional computation effort.

Correlation Risk

Correlation coefficients between market instruments are usually not obtained easily. Either historical data analysis or implied calibrations need to be done. However, in the foreign exchange (FX) market, the cross instrument is traded as well. For instance, in the example above, USD/JPY spot and options are traded, and the correlation can be determined from this contract. In fact, denoting the volatilities as in the tetrahedron (Figure 2), we obtain formulae for the correlation coefficients in terms of known market implied volatilities:

$$\begin{aligned}\rho_{12} &= \frac{\sigma_3^2 - \sigma_1^2 - \sigma_2^2}{2\sigma_1\sigma_2} \\ \rho_{34} &= \frac{\sigma_1^2 + \sigma_6^2 - \sigma_2^2 - \sigma_5^2}{2\sigma_3\sigma_4}\end{aligned}\quad (10)$$

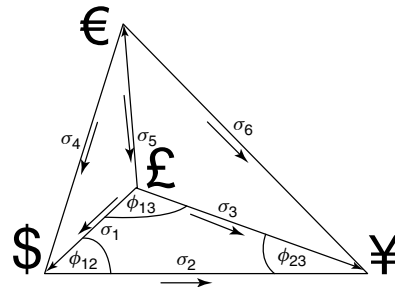


Figure 2 Currency tetrahedron including cross contracts

Table 1

GBP/USD	8.9%
USD/JPY	10.1%
GBP/JPY	9.8%
EUR/USD	10.5%
EUR/GBP	7.5%
EUR/JPY	10.0%

FX implied volatilities for three-month at-the-money vanilla options as of November 23, 2001. *Source:* Reuters

This method also allows hedging correlation risk by trading FX implied volatility. For details see [1].

Practical Example

To find out how much one can save using a basket option, we take EUR as a base currency and consider a basket of three currencies USD, GBP, and JPY. For the volatilities, we take the values in Table 1.

The resulting correlation coefficients are given in Table 2.

The amount of option premium one can save using a basket call rather than three individual call options is illustrated in Table 3.

The amount of premium saved essentially depends on the correlation of the currency pairs (Figure 3). In Figure 3, we take the parameters of the previous scenario, but restrict ourselves to the currencies USD and JPY.

Upper Bound by Vanilla Options

It is actually clear that the price of the two vanilla options in the previous example is an upper bound of the basket option price. It seems intuitively clear that for a correlation of 100% the price is the same.

Table 3

Base currency	EUR	Interest rate	4.0%
Nominal in EUR	39 007	Strike K	1
Currencies	USD	JPY	GBP
Nominals	29%	30%	41%
1/spot	1.1429	0.00919	1.6091
Spot	0.8750	108.81	0.6215
Strikes (in EUR)	1.1432	0.00927	1.5985
Volatilities	10.5%	10.0%	7.5%
Interest rates	4.0%	0.5%	7.0%
BS-values (in EUR)	235	227	233
Basket value	563		
Sum of individuals	695		

Comparison of a basket call with three currencies for a maturity of three months *versus* the cost of three individual call options

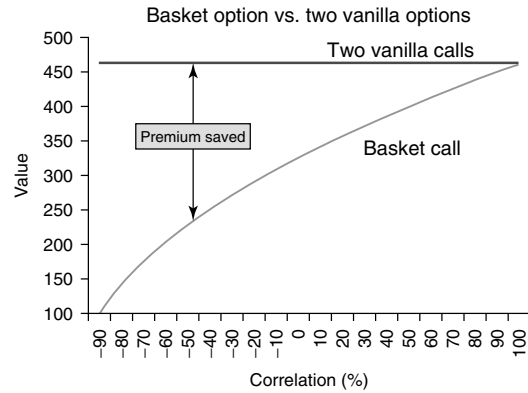


Figure 3 Premium of basket option versus premium of option strategy depending on the correlation

Surprisingly, this is just the case if a specific relation between the strike of the individual options and their volatilities is satisfied. The basket strike has to satisfy

$$K = a_1 \frac{K_1}{S_1(0)} + a_2 \frac{K_2}{S_2(0)} \quad (11)$$

Table 2

	GBP/USD (%)	USD/JPY (%)	GBP/JPY (%)	EUR/USD (%)	EUR/GBP (%)	EUR/JPY (%)
GBP/USD	100	-47	42	71	-19	27
USD/JPY	-47	100	60	-53	-18	45
GBP/JPY	42	60	100	10	-36	71
EUR/USD	71	-53	10	100	55	52
EUR/GBP	-19	-18	-36	55	100	40
EUR/JPY	27	45	71	52	40	100

FX implied three-month correlation coefficients as of Nov 23, 2001

which leads to the natural choice

$$K_i = K \frac{S_i(0)}{a_1 + a_2} \quad (12)$$

Each strike K_i satisfies the above constraint by choosing

$$K_i = S_i(0)e^{(\mu_i + 1/2\sigma_i^2)T + \chi\sigma_i\sqrt{T}} \quad (13)$$

for some arbitrary, but common χ for all basket components.

Smile Adjustment

For the pricing method, described there is no smile considered. Given the volatility smile for vanilla options, $\sigma_i(K, T)$, with the same maturity as the basket option, the implied density P for each currency pair in the basket can be derived from vanilla prices V .

$$P(K, T) = e^{rT} \partial_{KK} V(K, \sigma_i(K, T)) \quad (14)$$

A mapping $\varphi(w)$ can be derived that maps the Gaussian random numbers to smile-adjusted random numbers for each currency pair. The implicit construction solves the problem for the probability of the mapped Brownian to be the same as the smile-implied probability.

Using Monte Carlo simulation to price vanilla options using the mapping, it can be shown that in the limit, the derived prices are perfectly in line with the smile. The formula for the Monte Carlo simulation for a realization of a Brownian w is given by

$$S_i(0, w) = S_i(0)e^{(\mu_i + 1/2\sigma_i^2)T + \varphi(w)\sigma_i\sqrt{T}} \quad (15)$$

To price the basket option using the smile in Monte Carlo, a sequence of independent random numbers is used. These random numbers are correlated using the square-root matrix Ω as above and these are fed into the individual mappings, hence generating the simulated spot at the basket maturity. Evaluating the payoff and averaging will generate a smile-adjusted price (see Table 4). Black–Scholes prices and smile-adjusted prices are shown next to each other for a direct comparison.

Conclusion

Many corporate portfolios are exposed to multicurrency risk. One way to turn this fact into an advantage

Table 4

Base currency	EUR		
Nominal in EUR	39 007		
Currencies	USD	JPY	GBP
Nominals	29%	30%	41%
RR 25d	−0.25%	−4.30%	1.10%
Fly 25d	0.30%	0.17%	0.25%
BS-values (in EUR)	235	227	233
Smile values (in EUR)	233	168	278
Basket smile value	554		
Sum of individuals (smile)	680		
Basket value	563		
Sum of individuals	695		

is to use multicurrency hedge instruments. We have shown that basket options are convenient instruments protecting against exchange rates of most of the basket components changing in the *same* direction. A rather unlikely market move of half of the currencies' exchange rates in *opposite* directions is not protected by basket options, but when taking this residual risk into account, the hedging cost is reduced substantially. The smile impact on the basket value can be calculated rather easily without referring to a specific model, because the product is path-independent.

End Notes

^aThis article is an extension of “Hakala, J. & Wystup, U. (2002) *Making the most out of Multiple Currency Exposure: Protection with Basket Options, The Euromoney Foreign Exchange and Treasury Management Handbook 2002*, Adrian Hornbrook.”

References

- [1] Hakala, J. & Wystup, U. (2001). *Foreign Exchange Risk*, Risk Publications, London.

Further Reading

Wystup, U. (2006). *FX Options and Structured Products*, Wiley.

Related Articles

Basket Options.

JÜRGEN HAKALA & UWE WYSTUP

Call Options

Call options appeared as rights to buy an underlying traded asset for a prespecified price, named the *option strike* or the *exercise price*, at a prespecified future date named the *option expiry* or the *maturity*. Put options are analogous rights to sell an underlying asset. For strike K and maturity T with the underlying asset trading at maturity for S , the call expires unexercised if S is below K while the put expires unexercised if S is above K . On exercise, the value of the call option is $S - K$ while that of the put option is $K - S$. Hence, one may write the payoffs at maturity to the call and put options as $(S - K)^+$ and $(K - S)^+$, respectively. More generally, one may define a call or put payoff for any underlying random variable, which need not be a traded asset, for which the realized value at maturity is known to be A , as $(A - K)^+$ and $(K - A)^+$, respectively.

When call and put options trade before the maturity, on an underlying uncertainty resolved at maturity for various strikes K , with prices determined in markets at time $t < T$ as $c(t, K, T)$, $p(t, K, T)$, respectively, we have an options market for the underlying risk. Such markets provide a rich source of opportunities for holding the underlying asset or risk while simultaneously providing information on the prices of these risks. With regard to the opportunities, they make it possible to hold any function $f(A)$ of the underlying risk *via* a portfolio of put and call options. This fact is easily demonstrated as follows [2].

Let $f(A)$ be the function we wish to hold. We note that

$$\begin{aligned} f(A) &= f(a) + \mathbf{1}_{A>a} \int_a^A f'(u) du - \mathbf{1}_{a>A} \int_A^a f'(u) du \\ &= f(a) + \mathbf{1}_{A>a} \int_a^A \left[f'(a) + \int_a^u f''(v) dv \right] du \\ &\quad - \mathbf{1}_{a>A} \int_A^a \left[f'(a) - \int_u^a f''(v) dv \right] du \\ &= f(a) + f'(a)(A - a) + \mathbf{1}_{A>a} \\ &\quad \times \int_a^A dv f''(v)(A - v) \\ &\quad + \mathbf{1}_{a>A} \int_A^a dv f''(v)(v - A) \end{aligned}$$

$$\begin{aligned} &= f(a) + f'(a)(A - a) + \mathbf{1}_{A>a} \\ &\quad \times \int_a^\infty dv f''(v)(A - v)^+ \\ &\quad + \mathbf{1}_{a>A} \int_0^a dv f''(v)(v - A)^+ \end{aligned} \quad (1)$$

On the right hand side, we have a position in a bond with face value given by the constant term, a position in the underlying risk of $f'(a)$ and a position in puts struck below a and calls struck above a at strike v of $f''(v)$.

With regard to the information content of the market prices, we consider Breeden and Litzenberger [1], who showed how one may extract the pricing density at time $t < T$, $p(t, A)$ for the underlying risk from market option prices. By definition, we have

$$c(t, K, T) = e^{-r(T-t)} \int_K^\infty (A - K) p(t, A) dA \quad (2)$$

where r is the interest rate prevailing at time t for the maturity $(T - t)$. We may differentiate twice with respect to the strike to get

$$p(t, K) = e^{r(T-t)} \frac{\partial^2 c(t, K, T)}{\partial K^2} \quad (3)$$

In the case when the underlying risk is an asset price with a specific dynamics with exposure to a Brownian motion with a space–time deterministic volatility (*see Local Volatility Model*) as postulated by Dupire [6] plus a compensated jump martingale with a space–time deterministic arrival rate of jumps and a fixed dependence of the arrival rate on the jump size, one may extract information on the dynamics from market prices. Here, we follow Carr *et al.* [4]. Let $(S(t), t > 0)$ denote the path of the stock price, where r is the interest rate, η the dividend yield, $\sigma(S, t)$ the deterministic space–time volatility function, $(W(t), t > 0)$ a Brownian motion, $m(dx, ds)$ the integer-valued counting measure associated with the jumps in the logarithm of the stock price, $a(S, t)$ the deterministic space–time jump arrival rate, and $k(x)$ the Lévy density across jump sizes x . The dynamics for the stock price may be written as

$$\begin{aligned} S(t) &= S(0) + \int_0^t S(u_-)(r - \eta) du \\ &\quad + \int_0^t S(u_-)\sigma(S(u_-), u) dW(u) \end{aligned}$$

$$\begin{aligned}
& + \int_0^t \int_{-\infty}^{\infty} S(u_-) (e^x - 1) (m(dx, du) \\
& - a(S(u_-), u)k(x) dx du) \quad (4)
\end{aligned}$$

We now apply a generalization of Itô's lemma to convex functions known as the *Meyer–Tanaka formula* (see, e.g., [5, 7, 8] for the specific formulation below) to the call option payoff at maturity to obtain

$$\begin{aligned}
(S(T) - K)^+ &= (S(0) - K)^+ + \int_0^T \mathbf{1}_{S(u_-) > K} dS(u) \\
&+ \frac{1}{2} \int_0^T \delta_K(S(u_-)) \sigma^2 \\
&\times (S(u), u) S^2(u) du \\
&+ \sum_{u \leq T} \mathbf{1}_{S(u_-) > K} (K - S(u))^+ \\
&+ \mathbf{1}_{S(u_-) < K} (S(u) - K)^+ \quad (5)
\end{aligned}$$

The second integral denotes the value at K of the continuous local time L_T^a ; $a \in \mathbb{R}$, which is globally defined for every bounded Borel function f , as $\int_{-\infty}^{\infty} f(a) L_T^a da = \int_0^T f(S(u_-)) d\langle S^c \rangle_u$, where $d\langle S^c \rangle_u = \sigma^2(S(u), u) S^2(u) du$, and is applied here formally to the Dirac measure $f(a) = \delta_K(a)$. The last term, which is the discontinuous component of local time at level K , is made up of just the crossovers, whereby one receives $S(u) - K$ on crossing the strike into the money, whereas one receives $(K - S(u))$ on crossing the strike out of the money.

Computing expectations on both sides of equation (5) and introducing $q(\Sigma, u)$, the transition density that the stock price is Σ at time u given that at time 0 it is at $S(0)$, we may write the call price function at time zero as

$$\begin{aligned}
e^{rT} C(K, T) &= (S(0) - K)^+ \\
&+ \int_0^T \int_K^{\infty} dY q(Y, u) Y (r - \eta) du \\
&+ \frac{1}{2} \int_0^T q(K, u) \sigma^2(K, u) K^2 du \\
&+ \int_0^T \int_K^{\infty} dY q(Y, u) a(Y, u) \\
&\times \int_{\ln(\frac{K}{Y})}^{\ln(\frac{K}{Y})} (K - Y e^x) k(x) dx du
\end{aligned}$$

$$\begin{aligned}
&+ \int_0^T \int_0^K dY q(Y, u) a(Y, u) \\
&\times \int_{\ln(\frac{K}{Y})}^{\infty} (Y e^x - K) k(x) dx du \quad (6)
\end{aligned}$$

Now differentiating equation (6) with respect to T , we get

$$\begin{aligned}
r e^{rT} C + e^{rT} C_T &= (r - \eta) \int_K^{\infty} q(Y, T) Y dY \\
&+ \frac{\sigma^2(K, T) K^2}{2} q(K, T) \\
&+ \int_K^{\infty} dY Y q(Y, T) a(Y, T) \int_{-\infty}^{\ln(\frac{K}{Y})} \left(e^{\ln(\frac{K}{Y})} - e^x \right) \\
&\times k(x) dx + \int_0^K dY Y q(Y, T) a(Y, T) \\
&\times \int_{\ln(\frac{K}{Y})}^{\infty} \left(e^x - e^{\ln(\frac{K}{Y})} \right) k(x) dx \quad (7)
\end{aligned}$$

We now isolate C_T on the left, using some elementary properties of the relationship between call prices and the pricing density. In particular, we note

$$e^{-rT} \int_0^{\infty} Y q(Y, T) dY = C - K C_K \quad (8)$$

$$e^{-rT} q(K, T) = C_{KK} \quad (9)$$

and obtain

$$\begin{aligned}
C_T &= -\eta C - (r - \eta) K C_K \\
&+ \frac{\sigma^2(K, T) K^2}{2} C_{KK} \\
&+ \int_K^{\infty} dY Y C_{YY} a(Y, T) \\
&\times \int_{-\infty}^{\ln(\frac{K}{Y})} \left(e^{\ln(\frac{K}{Y})} - e^x \right) k(x) dx \\
&+ \int_0^K dY Y C_{YY} a(Y, T) \\
&\times \int_{\ln(\frac{K}{Y})}^{\infty} \left(e^x - e^{\ln(\frac{K}{Y})} \right) k(x) dx \quad (10)
\end{aligned}$$

We may define the function

$$\begin{aligned}\psi(x) &= \mathbf{1}_{x < 0} \int_{-\infty}^x (e^x - e^s)k(s) ds \\ &\quad + \int_x^{\infty} (e^s - e^x)k(s) ds\end{aligned}\quad (11)$$

and then write

$$\begin{aligned}C_T &= -\eta C - (r - \eta)KC_K \\ &\quad + \frac{\sigma^2(K, T)K^2}{2}C_{KK} \\ &\quad + \int_0^{\infty} C_{YY}Ya(Y, T)\psi\left(\ln\left(\frac{K}{Y}\right)\right)dY\end{aligned}\quad (12)$$

When there are no jumps in the process for X and $\psi \equiv 0$, equation (12) is identical to the formula in [6] for local volatility (see **Dupire Equation**). In the opposite case, when there is no continuous martingale component, we have the following result:

$$\begin{aligned}C_T + \eta C + (r - \eta)KC_K \\ = \int_0^{\infty} C_{YY}Ya(Y, T)\psi\left(\ln\left(\frac{K}{Y}\right)\right)dY\end{aligned}\quad (13)$$

It is now useful to rewrite equation (13) in terms of $k = \ln(K)$, $y = \ln(Y)$, and $c(k, T) = C(e^k, T)$. With this substitution, we may rewrite equation (12) as

$$\begin{aligned}c_T + \eta c + \left(r - \eta + \frac{\sigma^2(e^k, T)}{2}\right)c_k \\ - \frac{\sigma^2(e^k, T)}{2}c_{kk} = \int_{-\infty}^{\infty} b(y, T)\psi_e(k - y)dy\end{aligned}\quad (14)$$

where $b(y, T) = e^{2y}C_{YY}a(e^y, T)$. The forward speed function, $a(Y, T)$, may be identified as

$$a(Y, T) = \frac{b(\ln(Y), T)}{Y^2 C_{YY}}\quad (15)$$

For specific Lévy measures, the convolution equation (14) may be solved in closed form to yield explicit solutions for the Markov process from data on option prices (see **Fourier Methods in**

Options Pricing; Partial Integro-differential Equations (PIDEs)).

Apart from spanning all functions of the underlying risk and providing us with information on the possible dynamic movements of the stock price consistent with market option prices, we have the question of understanding the absence of arbitrage between option prices. This question was addressed in [3], where it was shown that if call spread, butterfly spread, and calendar spread arbitrages are excluded then the option quotes are free of static arbitrage. It is, therefore, important to document the three arbitrages that need to be checked. For a call spread, we have the inequality for two strikes $K_1 < K_2$ for a fixed maturity T :

$$\frac{c(K_1, T) - c(K_2, T)}{K_2 - K_1} < 1$$

For a butterfly spread, we have three strikes $K_1 < K_2 < K_3$ and a fixed maturity T for which we must have

$$\begin{aligned}c(K_1, T) - \left(\frac{K_3 - K_1}{K_3 - K_2}\right)c(K_2, T) \\ + \frac{K_2 - K_1}{K_3 - K_2}c(K_3, T) \geq 0\end{aligned}\quad (16)$$

Finally, the calendar spread inequality requires that for two maturities $T_1 < T_2$ and strike K

$$c(K, T_2) \geq c(K e^{-r(T_2 - T_1)}, T_1)\quad (17)$$

Similar results hold for put options *via* put-call parity that asserts in the case of a stock

$$p(K, T) = c(K, T) + K - S(0)e^{-\eta T}\quad (18)$$

The call spread inequality approximates the probability that the stock exceeds K_1 when we take K_2 close to and above K_1 . The butterfly spread inequality guarantees the existence of a positive pricing density and the calendar spread inequality arranges these densities to be increasing in the convex order with respect to the maturity. When the underlying risk is not a traded asset as occurs, for example, for options on the VIX index where the underlying is the price of the one month variance swap, we lose the calendar spread inequality and the requisite densities are not increasing in the convex order. One can check that on

most days VIX call option prices when deflated by the forward VIX are increasing in maturity for given strikes and we have an empirical increase in the convex order, but there are days when this monotonicity is lost. The conditions for VIX option surfaces to be free of arbitrage are, therefore, not as clear as they are for an underlying stock or a stock index.

References

- [1] Breeden, D. & Litzenberger, R.L. (1978). Pricing of state-contingent claims implicit in option prices, *Journal of Business* **51**, 621–651.
- [2] Carr, P. & Madan, D.B. (2001). Optimal positioning in derivatives, *Quantitative Finance* **1**, 19–37.
- [3] Carr, P. & Madan, D.B. (2005). A note on sufficient conditions for no arbitrage, *Finance Research Letters* **2**, 125–130.

- [4] Carr, P., Geman, H., Madan, D. & Yor, M. (2005). From local volatility to local Lévy models, *Quantitative Finance* **4**, 581–588.
- [5] Dellacherie, C. & Meyer, P. (1980). *Probabilités et Potentiel, Theorie des Martingales*, Hermann, Paris.
- [6] Dupire, B. (1994). Pricing with a smile, *Risk* **7**, 18–20.
- [7] Meyer, P. (1976). Un Cours sur les Intégrales stochastiques, in *Séminaire de Probabilités X*, Lecture Notes in Mathematics, Springer-Verlag, Berlin, Vol. 511.
- [8] Yor, M. (1978). Rappels et Préliminaires Généraux, in *Temps Locaux, Société Mathématique de France*, Astérisque, pp. 17–22, 52–53.

Related Articles

Dupire Equation; Local Volatility Model; Put–Call Parity; Static Hedging; Variance Swap.

DILIP B. MADAN

Barrier Options

Barrier options are vanilla options with path-dependent payoffs, that is, the payoff is not only a function of stock level relative to option strike but also dependent upon whether or not the stock reaches certain prespecified barrier level before maturity. An example will illustrate the idea. Suppose an investor is long an up-and-in at-the-money call option on the S&P 500 index with barrier level at 110% of the initial S&P 500 index level. Before maturity, if the index never reaches 110% of the initial index level, the option never gets knocked in. The investor receives nothing at maturity. However, if the index level reaches 110% at some point before maturity, the investor receives a payoff identical to a vanilla at-the-money call option at maturity. In the latter scenario, the option is “knocked in” on the day when the index reaches 110% level.

There are many types of barrier options. We discuss two common ones, that is, knock-out and knock-in barrier options. For knock-out barrier options, the option will be knocked out and become worthless if the underlying asset crosses a prespecified barrier level. For knock-in barrier options, the barrier option will be knocked in and become a vanilla option only if the underlying asset crosses the prespecified level before maturity. The example used earlier is a knock-in barrier option.

Depending upon the barrier level relative to the initial underlying asset level, we can have an “up” barrier or a “down” barrier. If the barrier is above the initial underlying asset level, it is called an *up barrier*. If the barrier is below the initial underlying asset level, it is called a *down barrier*. Together, we can have four different variations of barrier options, that is, up-and-in, up-and-out, down-and-in, and down-and-out options. Table 1 shows these four variations schematically.

Basic Features

Up-and-in Call/Up-and-in Put

This is the first kind of knock-in barrier options. The up-and-in barrier option has a knock-in barrier level, which is higher than the initial underlying asset level. Before maturity, if the underlying asset goes above

the barrier level, the barrier option will be knocked in and become a vanilla option. Otherwise, the barrier option will expire worthless at maturity. Up-and-in calls are more common. This is because, when the underlying asset increases to knock-in barrier level, it would most likely stay above the initial underlying asset level. Therefore, call options will be more likely be in the money at maturity than put options. Bullish investors can buy up-and-in call options and pay a lower premium than that on the vanilla call options. This makes the on up-and-in calls more leveraged than vanilla calls.

Down-and-in Call/Down-and-in Put

The down-and-in barrier option has a knock-in barrier level, which is below the initial underlying asset level. Before the maturity, if the underlying asset goes below the barrier level the barrier option will be knocked in and become a vanilla option. Otherwise, the barrier option will expire worthless at maturity. Down-and-in puts are more common in this case. Bearish investors can buy down-and-in puts and pay a lower premium than that on the vanilla put options.

Up-and-out Call/Up-and-out Put

This is the first kind of knock-out barrier options. The up-and-out barrier option has a knock-out barrier level above the initial underlying asset level. Before maturity, if the underlying asset crosses the barrier level, the option will be knocked out and become worthless. Otherwise, the barrier option will just be a vanilla option. A bearish investor would buy up-and-out puts to achieve more leverage by paying a lower premium than that on vanilla puts.

Down-and-out Call/Down-and-out Put

The down-and-out barrier option has a knock-out barrier level below the initial underlying asset level. Before maturity, if the underlying asset goes below the barrier level, the option will be knocked out and become worthless. A bullish investor would buy down-and-out calls to achieve more leverage by paying a lower premium than that on vanilla calls.

Some Variations

With increased popularity of barrier options and growth of its market, some other features are being

2 Barrier Options

Table 1 Common barrier types

	Knock in	Knock out
Up barrier	Up-and-in call	Up-and-out call
	Up-and-in put	Up-and-out put
Down barrier	Down-and-in call	Down-and-out call
	Down-and-in put	Down-and-out put

introduced to better meet investors' needs. In the following, we discuss briefly some of the most popular variations to basic knock-in/knock-out barrier options.

Rebates

It is fairly common that a knock-out barrier option pays a rebate on a knock-out event to compensate the investor for the loss of the option. The rebate is typically a small amount and paid either immediately or deferred to the maturity. For knock-in barrier options, a rebate will be paid out at maturity if the barrier is never knocked in.

Double Barrier

A double barrier option is another variation that has two barriers, typically one up barrier and the other a down barrier. For example, investors seeking high leverage would consider double knock-out barrier options if they believe changes in the underlying asset level would be within a narrow range. The double knock-out barrier option would become worthless if the underlying asset drops below the down barrier or rises above the up barrier before maturity. Because of that, the double knock-out barrier option costs less than the single knock-out barrier option and thus provides more leverage.

Another popular variation of double barrier is the knock-in option with a knock-out barrier. The knock-in barrier option can be knocked out either before or after the option is knocked in. If the option can be knocked out even before the option is knocked in, this is called *knock-out dominant barrier option*. Again, this type of double barrier options provides more leverage than a single barrier option because it costs less than the single knock-in barrier option.

Discrete Barrier

The barrier specification varies from contract to contract. Many barrier options have the so-called continuous barrier or intraday barrier. This means that the barrier event can be triggered any time during intraday trading hours. However, some barrier options only allow the barrier event to be triggered by the end-of-day closing price. This is called *discrete barrier*. More generally, a discrete barrier is defined as any barrier type other than the continuous barrier.

American Style Exercise

Although it is not very common, there are a few types of barrier options that have American exercise feature. One example would be a six-month at-the-money call option with installment premium, that is, the premium would be paid monthly instead of a lump sum upfront. The knock-out barrier is at 90% of the initial underlying asset level. The option will be knocked out if the underlying asset drops below the knock-out barrier level before maturity. In addition, the option would be terminated automatically if the investor stops paying the installment premium on monthly reset dates. This installment feature provides more flexible financing for investors because they pay the premium over a period of time instead of paying the lump sum up front. It also allows investors to re-evaluate the market condition and decide if it is optimal to continue the knock-out option or terminate it.

Valuation of Barrier Options

Barrier options are less expensive than vanilla options because of the possibility that the barrier option would be knocked out or knocked in. There are many publications on how to price barrier options (see [2, 7] for a good summary). In general, the valuation of barrier options needs to take into account stock-level dependency of volatility dynamics, that is, local volatility surface. This requires using numerical methods like partial differential equations (see **Finite Difference Methods for Barrier Options**) or Monte Carlo simulation. In certain situations, the barrier options can be priced by using the static replication method (see [3, 4]).

In-Out Parity

Similar to put-call parity for the vanilla options, there is an interesting relationship between knock-in and knock-out barrier option. Using call options as example, an up-and-out call is complimentary to an up-and-in call if both have the same strike and barrier level. A portfolio of one up-and-in call and one up-and-out call will have the same payoff at maturity as a vanilla call option. The reason is simple. If the underlying asset stays below the barrier level before maturity, the up-and-in call will become worthless and the up-and-out call will become a vanilla call. On the other hand, if the underlying asset rises above the barrier level, the up-and-out call will become worthless and the up-and-in call will become a vanilla call. Therefore, for any given scenario of the underlying asset path before maturity, the portfolio will always have the same payoff as a vanilla call. The sum of an up-and-in call and an up-and-out call is the same as a vanilla call. This is so-called in-out parity for barrier options and applies to both the put and call options in the absence of rebates.

Constant Volatility

In the limit of constant volatility, Merton [5] provided the first analytical formula for a down-and-out call option. Later on, Reiner and Rubinstein [6] extended the formula for all eight combinations of barriers (*see Pricing Formulae for Foreign Exchange Options*).

Adjustment for Discrete Barrier

Often, the barrier option has a discrete barrier schedule but the exact valuation is only available for the case of continuous barrier. In the constant volatility

case, Broadie *et al.* [1] proposed that, by adjusting the barrier level in the continuous barrier option valuation, one would obtain a good approximation of the discrete barrier option value. The adjustment to the barrier level is prescribed by the following formula:

$$X_{\text{adj}} = Xe^{\alpha\sigma\sqrt{T/m}} \quad (1)$$

where α is 0.5826 for an “up” barrier and -0.5826 for a “down” barrier, m is the number of discrete samples of the underlying asset price over the term T of the barrier option, and σ is the constant volatility.

References

- [1] Broadie, M., Glasserman, P. & Kou, S.G. (1997). A continuity correction for discrete barrier options, *Journal of Mathematical Finance* **7**, 325–349.
- [2] Briys, E., Bellalah, M., Mai, H.M. & de Varenne, F. (1998). *Options, Futures and Exotic Derivatives—Theory, Application and Practice*, Wiley Frontiers in Finance.
- [3] Carr, P. & Chou, A. (1997). Breaking barriers, *Risk Magazine* **10**, 139–144.
- [4] Derman, E., Ergener, D. & Kani, I. (1997). Static options replication, in *Frontiers in Derivatives*, Irwin Professional Publishing.
- [5] Merton, R. (1971). Theory of rational option pricing, *Bell Journal of Economics and Management* **4**, 141–183.
- [6] Reiner, E. & Rubinstein, M. (1991). Breaking down barriers, *Risk Magazine* **4**, 28–35.
- [7] Wilmott, P. (1998). *Derivatives—The Theory and Practice of Financial Engineering*, John Wiley & Sons.

Related Articles

Finite Difference Methods for Barrier Options; Pricing Formulae for Foreign Exchange Options.

MICHAEL QIAN

Corridor Options

Occupation time derivatives are debt securities that came into existence in 1993 and have attracted some attention from investors and researchers. A defining characteristic of these contracts is a payoff that depends on the time spent by the underlying asset in some predetermined region. Typical specifications consist in interest payments that are proportional to the time in which a reference index rate (most commonly the Libor rate) lies inside a given range. In return for the drawback that no interest will be paid for the time the corridor is left, they offer higher rates than comparable standard products, like floating rate notes. Various claims with features of this type have been studied and have been popularized with different names such as corridor bond or option, range note, range floater, range accrual note, LIBOR range note, fairway bond,^a and hamster option.^b The most common underlyings are stock indices, foreign currencies, and interest rates, such as LIBOR or swap rates. Also spread range notes are common. Range notes pay a coupon proportional to the number of days in which the difference between two interest rates (e.g., 10-year swap rate *versus* 30-year swap rate) is positive. Thus while the value of an interest range note depends on the volatility of the level of the term structure, the value of a range note depends on the volatility of the slope of the term structure. This makes it important to model the correlation of interest rates with different maturities accurately.

One of the most popular structures is the accrual note, typically of one- or two-year maturity, which offers a coupon calculated by the number of days in which three-month dollar Libor, for example, falls inside a predefined range. On these days, the note will typically offer a preassigned spread over the relevant treasury bond. When Libor is outside the range, the payout is zero. The note effectively provides a way for investors to sell volatility, but the binary structure protects them from the unlimited downside that would accompany similar strategies, such as writing a floor and cap on the range. Corridor products offer investors enhanced yield if they have a strong view that rates will stay within a range, and often they are structured to reflect an investor's view that is contrary to a particular forward-rate curve. In its simplest form, the payoff of the corridor bond can

be seen as the sum of the payoffs of digital options expiring on successive days.

Payoff

Let t be the current time (measured in years) and let $0 = T_0 < t < T_1, \dots, T_j, \dots, T_n$ be the payment dates of the coupons $\phi(T_j)$ (T_0 is the last date at which a payment occurred). Let $D(T_j, T_{j+1})$ be the number of days in the coupon period and let $T_{j,i}$ be the date corresponding to i days after date T_j , that is, $T_{j,i} = T_j + i/365$, $i = 1, \dots, D(T_j, T_{j+1})$. Let x_l and x_u be the lower and upper bound of the range (the prespecified range for each observed date can vary daily or across different compounding periods). Finally let $X(t)$ be the value of the reference index at time t . Sometimes, X represents a stock index or a foreign currency rate, and sometimes is taken to be a LIBOR rate of a preassigned tenor (in this case $X(t) = L(t, t + \theta)$, where L is the LIBOR rate and θ is its tenor) or the spread of two swap rates with different maturities.

The range note can have a fixed and a floating version. For a range note, the value of the coupon paid at time T_{j+1} is equal to

$$\phi(T_{j+1}) = C_{j+1} \times \frac{H(T_j, T_{j+1})}{D(T_j, T_{j+1})} \quad (1)$$

where C_{j+1} represents the annual coupon rate for the $(j+1)$ th compounding period and is given by

$$C_{j+1} = \begin{cases} C & \text{fixed range note,} \\ X(T_j) + \Delta & \text{delayed floating range note,} \end{cases} \quad (2)$$

where C is a constant and Δ is the spread to be added to the reference rate, and

$$H(T_j, T_{j+1}) = \sum_{i=1}^{D(T_j, T_{j+1})} 1_{[x_l, x_u]}(X(T_{j,i})) \quad (3)$$

where $1_A(x)$ is the indicator function of the set A , that is, $1_A(x) = 1$ if $x \in A$, otherwise $1_A(x) = 0$. Here the term *delayed* refers to the fact that the option maturity (or reset date) T_j anticipates the payment date T_{j+1} . Sometimes, instead of having the delayed floating range note we have the *in-arrears* floating range note where the coupon is set equal to

$$C_{j+1} = X(T_{j+1}) + \Delta \quad (4)$$

that is, the coupon payment depends on the level of the reference rate at the current coupon payment date (maturity and payment date coincide).

Sometimes, a minimum coupon clause is also included, so that the coupon amounts to

$$\phi(T_{j+1}) = \max\left(C_{j+1} \times \frac{H(T_j, T_{j+1})}{D(T_j, T_{j+1})}, K\right) \quad (5)$$

The standard contract is the fixed rate range note if the underlying is a stock index or a foreign currency and the delayed floating range note if the underlying is a LIBOR rate.

Pricing

In this section, we discuss the pricing problem of the corridor contracts described earlier.

Underlying is a Stock Index

In this case, we model the underlying according to the Geometric Brownian Motion (GBM) model, that is,

$$dX(t) = (r - q)X(t) dt + \sigma X(t) dW(t), X(t) = x \quad (6)$$

where

r : instantaneous risk-free rate,
 q : instantaneous dividend yield,
 σ : percentage volatility,
 x : initial underlying price.

An analytical formula for pricing fixed range notes is promptly available, resorting to the fact that

$$\begin{aligned} & \mathbb{E}_t(1_{[x_l, x_u]}(X(T))) \\ &= \mathbb{E}_t(1_{(-\infty, x_u]}(X(T)) - 1_{(-\infty, x_l]}(X(T))) \\ &= \mathcal{N}(d_{T-t}^{x_u}) - \mathcal{N}(d_{T-t}^{x_l}) \end{aligned}$$

where

$$\begin{aligned} d_{T-t}^w &= \frac{\ln \frac{x}{w} + (r - q)(T - t) - \frac{1}{2}\sigma^2(T - t)}{\sqrt{\sigma^2(T - t)}} \\ \mathcal{N}(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{u^2}{2}} du \end{aligned} \quad (7)$$

Therefore, for a fixed-rate range note, we have

$$\begin{aligned} & \mathbb{E}_t(\phi(T_{j+1})) \\ &= \frac{C_j}{D(T_j, T_{j+1})} \sum_{i=1}^{D(T_j, T_{j+1})} \left(\mathcal{N}(d_{T_{j,i}-t}^{x_u}) - \mathcal{N}(d_{T_{j,i}-t}^{x_l}) \right) \end{aligned} \quad (8)$$

and the price of the fixed range note is

$$\sum_{j=0}^n P(t, T_{j+1}) \mathbb{E}_t(\phi(T_{j+1})) + P(t, T_n) \quad (9)$$

where $P(t, T)$ refers to the price of a zero-coupon bond expiring in T .

From a pricing perspective, the case of the fixed range note with a minimum coupon provision appears more interesting. Indeed, in this case, the pricing formula requires the distribution of the occupation time of the range $[x_l, x_u]$. If the random variable $H(T_j, T_{j+1})/D(T_j, T_{j+1})$ in equation (3) is replaced by its continuously monitored version

$$h(T_j, T_{j+1}; x_l, x_u) = \int_{T_j}^{T_{j+1}} 1_{[x_l, x_u]}(X(s)) ds \quad (10)$$

few analytical results are available. In particular, the distribution of $h(T_j, T_{j+1}; x_l, x_u)$ when $x_l = -\infty$ (or $x_u = \infty$) is obtained in [1], exploiting the Feynman–Kac formula. Related results are presented in [4, 5, 7, 10, 12, 16, 17]. Owing to the stationarity of the Brownian motion, we observe that the distribution of $h(T_j, T_{j+1})$ is the same as the distribution of $h(0, T_{j+1} - T_j; x_l, x_u)$, and this law (when $x_l = -\infty$) is strictly related to the law of the occupation time in the time interval $[0, \tau]$ as

$$Y_{0,x}(u, \tau) = \int_0^\tau \mathbf{1}_{(-\infty, u]}(\delta s + W(s)) ds, W(0) = x \quad (11)$$

where $\delta = (r - q - \sigma^2/2)/\sigma$. Fusai [8] provides the pricing formula in the case of finite values of x_l and x_u , obtaining the Laplace transform in time of the characteristic function of $Y_{0,x}(u, \tau)$.

The real-life case of discrete monitoring is discussed in [9] using Monte Carlo simulation and finite difference methods. In particular, the authors stress that the price of the contract with continuous time

monitoring is the highest (lowest) when the index is inside (outside) the band. This is due to the nature of the contract: if the index is inside the band, and we assume continuous time monitoring, then the passage of time will increase the value of the contract until the moment in which the index crosses the barriers. Instead, if we assume discrete time monitoring, we cannot exploit the passage of time completely: we register the position of the index only at discrete dates and if they are quite distant (e.g., a month), it is possible that the index at the reset date will move outside the band and so the occupation time does not increase. In this case, the time between two monitoring dates is entirely lost. *Vice versa*, in the discrete case, if we are outside the band, and between two monitoring dates the index moves inside the band, then the occupation time increases by the entire time distance between monitoring dates. Instead, with continuous time monitoring we miss every instant before the process crosses the barrier. So the continuous time formula will overvalue (undervalue) the discrete time formula when the index is inside (outside) the band.

We have not discussed here the pricing of the floating range notes with a stock index as underlying because they are not very common. However, using the stock as numéraire and the tower property (to take into account the delay between maturity date and payment date) analytical formula are promptly available. In the case of a minimum coupon, we need the joint law of Brownian motion and its occupation times. Useful formulas for this case are provided in [2, 10, 12].

Underlying is an Interest Rate

In this case, the underlying variable $X(t)$ of the range note is a simple compounded interest rate with tenor θ defined according to the formula

$$1 + X(t)\theta = \frac{1}{P(t, t+\theta)} \quad (12)$$

The pricing of range notes, with the underlying being an interest rate, has begun with Turnbull [18], who assumes that the interest rate dynamics is well represented by a one factor Gaussian Heath–Jarrow–Morton (HJM) model, that is, the dynamics of the price of a zero-coupon bond $P(t, T)$ is given by

$$\frac{dP(t, T)}{P(t, T)} = r(t) dt + \sigma(t, T) dW(t) \quad (13)$$

where $\sigma(t, T)$ is deterministic.

In particular, Turnbull [18] considers the Ho and Lee model ($\sigma(t, T) = \sigma$, constant) and the Hull and White model ($\sigma(t, T) = \sigma e^{-\lambda(T-t)}$, $\lambda > 0$) and obtains a closed form formula for the range note (fixed rate and delayed floating case). A simpler and more intuitive derivation than [18] and for a more general volatility function (albeit still deterministic) is obtained, using the change of numéraire technique, in [14].

The multifactor Gaussian HJM

$$\frac{dP(t, T)}{P(t, T)} = r(t) P(t, T) dt + \sigma(t, T)' \cdot dW(t) \quad (14)$$

where $\cdot$ denotes the inner product in R^n and $W(t) \in R^n$ is an n -dimensional standard Brownian motion as considered in [15]. Such extension is important because it enhances the term structure model calibration to the interest rates covariance matrix “observed” in the market, which, along with the term structure of interest rates, will ultimately determine the price of the range notes under analysis. In fact, in order to price and hedge range notes consistently with the market prices of related plain-vanilla interest rate options (such as caps and/or European swaptions), it is essential to use an interest rate model that is analytically tractable and that provides a good fit to the term structures of interest rates, volatilities, and correlations observed in the market.

Reference [6] further extends these results to a multivariate Lévy term structure model, an extension of a Gaussian HJM model with jump processes.

The main limit of these models is that the rate $X(t)$ can attain negative values with positive probability, which may cause some pricing error in many cases.

An extension to the class of affine term structure model, which encompasses both Gaussian and non-Gaussian models such as Cox–Ingersoll–Ross (CIR) square-root models is introduced in [11]. The extension is useful in ensuring the positivity of the interest rate; moreover, Jangy and Yoon [11] have also consider the pricing spread range notes. The

limit of their analysis is that they consider as underlying asset a continuously compounded interest rate of a given tenor rather than the corresponding simple compounded interest rate. On one hand, this allows to get (i) analytical formulas for a more general class of multifactor models and (ii) positive interest rates. On the other hand, the pricing formulas are not immediately adaptable to real-life contracts because the spread is usually between swap rates of different tenors.

A Libor Market Model is instead adopted in [19]. This has the advantage of enhancing the model calibration to the interest rate covariance matrix observed in the market, and, in addition, it guarantees the positivity of interest rates. However, in this model no analytical formulas are available for floating range notes, which need to be priced by Monte Carlo simulation. However, freezing the drift of forward rates in order to have a dynamics of the LIBOR rate with deterministic drift and volatility, Wu and Chen [19] are able to provide approximate pricing formulae.

Finally, we remark that, with an interest rate as underlying, no pricing formula has so far been made available in the literature concerning the pricing of a range note with a minimum coupon provision.

Related Payoff

Exotic options related to corridor options are quantile options, introduced in [13], and studied in more detail in [1, 4]. Step options studied in [12] are contracts related to range notes, as an alternative to standard barrier options. Barrier options have the drawback of losing all value at the first touch of the barrier. Step options lose value more gradually. The option value decreases as the underlying asset spends more time at lower levels. Another example is the Parisian option (see **Constant Maturity Swap**). A Parisian out option with window D , barrier L , and maturity date T will lose all value if the underlying price has an excursion of duration D above or below the barrier L during the option's life. If the loss of value is prompted by an excursion above (below) the barrier, the option is said to be an up-and-out (down-and-out) Parisian option. Parisian contractual forms were introduced and studied by Chesney *et al.* [3]. Contracts of this type are more robust to eventual price manipulations. The pricing formulas in [3, 12] involve inverse Laplace transforms.

End Notes

^aThe fairway in golf is like the index or interest rate range. The outlook is positive if the ball lands on the fairway. If, however, a ball lands in the rough, the outlook is negative. Source: <http://www.investopedia.com/terms/f/fairwaybond.asp> as of January 2009.

^bThe German noun hamster has the same meaning as the English noun hamster: it is the name of a small rodent. But HAMSTER is also an acronym standing for Hoffnung Auf MarktStabilität in Einer Range (literally: Hope on market stability in a given range). It really is a pun as in German the verb 'hamstern' has the meaning of 'to hoard'. HAMSTER options hoard the fixed amount one gets for each day; the underlying stays in the prespecified range. What is earned cannot be lost any more. Source <http://www.margrabe.com/Dictionary/DictionaryGJ.html#sectH> as January 2009.

References

- [1] Akahori, J. (1995). Some formulae for a new type of path-dependent options, *Annals of Applied Probability* **5**, 383–388.
- [2] Borodin, A.N. & Salminen, P. (1996). *Handbook of Brownian Motion - Facts and Formulae*, Birkhauser.
- [3] Chesney, M., Jeanblanc-Picqué, M. & Yor, M. (1997). Brownian excursions and Parisian barrier options Brownian excursions and Parisian barrier options, *Advances in Applied Probability* **29**(1), 165–184.
- [4] Dassios, A. (1995). The distribution of the quantile of a Brownian motion with drift and the pricing of related path-dependent options, *Annals of Applied Probability* **4**(2), 719–740.
- [5] Douady, R. (1999). Closed form formulas for exotic options and their lifetime distribution, *International Journal of Theoretical and Applied Finance* **2**(1), 17–42.
- [6] Eberlein, E. & Kluge, W. (2006). Valuation of floating range notes in Lévy term-structure models, *Mathematical Finance* **16**(2), 237–254.
- [7] Embrechts, P., Rogers, L.C.G. & Yor, M. (1995). A proof of Dassios' representation of the α -quantile of Brownian motion with drift, *Annals of Applied Probability* **5**(3), 757–767.
- [8] Fusai, G. (2000). Corridor options and Arc-Sine law, *Annals of Applied Probability* **10**(2), 634–663.
- [9] Fusai, G. & Tagliani, A. (2001). Pricing of occupation time derivatives: continuous and discrete monitoring, *Journal of Computational Finance* **5**(1), 1–37.
- [10] Hugonnier, J. (1999). The Feynman-Kac formula and pricing occupation time derivatives, *International Journal of Theoretical and Applied Finance* **2**(2), 153–178.
- [11] Jangy, B.G. & Hee Yoon, J. (2008). *Valuation of Range Notes Under Affine Term Structure Models*, <http://ssrn.com/abstract=1291703>.

-
- [12] Linetsky, V. (1999). Step options: The Feynman-Kac approach to occupation time derivatives, *Mathematical Finance* **9**, 55–96.
 - [13] Miura, R. (1992). A note on lookback options based on order statistics, *Hitotsubashi Journal of Commerce and Management* **27**, 15–28.
 - [14] Navatte, P. & Quittard-Pinon, F. (1999). The valuation of interest rate digital options and range notes revisited, *European Financial Management* **5**(3), 425–440.
 - [15] Nunes, J.P.V. (2004). Multi-factor valuation of floating range notes, *Mathematical Finance* **14**(1), 79–97.
 - [16] Pechtl, A. (1995). Classified information, in *Over the Rainbow*, J. Robert, ed, Risk Publications, pp. 71–74.
 - [17] Takàcs, L. (1996). On a generalization of the arc-sine law, *Annals of Applied Probability* **6**(3), 1035–1040.
 - [18] Turnbull, S.M. (1995). Interest rate digital options and range notes, *Journal of Derivatives* **3**, 92–101.
 - [19] Wu, T.P. & Chen, S.N. (2008). Valuation of floating range notes in a LIBOR market model, *Journal of Futures Markets* **28**(7), 697–710.

Further Reading

Tucker, A.L. & Wei, J.Z. (1997). The latest range, *Advances in Futures and Options Research* **9**, 287–296.

Related Articles

Barrier Options; Corridor Variance Swap; Discretely Monitored Options; Parisian Option.

GIANLUCA FUSAI

Lookback Options

Lookback options are path-dependent options, introduced at first in [25] and [26], characterized by having their settlement based on the minimum or the maximum value of an underlying index as registered during the lifetime of the option. At maturity, the holder can “lookback” and select the most convenient price of the underlying that occurred during this period: therefore they offer investors the opportunity (at a price, of course) of buying a stock at its lowest price and selling a stock at its highest price. Since this scheme guarantees the best possible result for the option holder, he or she will never regret the option payoff. As a consequence, a lookback option is more expensive than a vanilla option with similar payoff function. However, these options do not offer a natural hedge for typical business and are used mainly by speculators. To mitigate their cost, sometimes the lookback feature is mixed with an average feature: for example, the payoff is the best or the worst of past average prices and they are offered as investment product under names such as *Everest*, *Napoleon*, and *Altiplano*.

In the section Payoff Function, we describe lookback options payoff. In the section Pricing, we illustrate the pricing of these options in the Black–Scholes setting and some results on the hedging problem. Thereafter, in the section Non-Gaussian Models, we consider the pricing problem under non-Gaussian models. Finally, in the section Related Payoff, we present payoffs related to lookback options.

Payoff Function

A lookback option can be structured as a put or call. The strike can be either fixed or floating. We now consider two lookback options written on the minimum value achieved by the underlying index during a fixed time window:

- *A fixed strike lookback put* The payoff is given by the difference, if positive, between the strike price and the minimum price over the monitoring period. Therefore, the buyer of this option can sell the asset at the minimum price receiving the strike K .
- *A floating strike lookback call* The payoff is given by the difference between the asset price at the option maturity, which represents the floating strike, and the minimum price over the monitoring period. Therefore, the buyer of this option can buy the underlying asset paying the minimum price.

Notice that floating strike options will always be exercised. Formulae for the payoffs are provided in Table 1, as well as versions involving the maximum and variants denominated *partial price* or *partial time*.

Pricing

In this section, we discuss the pricing problem under the geometric Brownian motion (GBM) assumption, that is,

$$dS(t) = (r - q)S(t)dt + \sigma S(t)dW(t), \quad S(0) = S_0 \quad (1)$$

where r is the instantaneous risk-free rate; q is the instantaneous dividend yield; σ is the percentage volatility; S_0 is the initial underlying price; and we are interested in the distribution of the minimum $m(T)$ and maximum $M(T)$

$$M(T) = \max_{0 \leq u \leq T} S(u), \text{ and } m(T) = \min_{0 \leq u \leq T} S(u) \quad (2)$$

Figure 1 illustrates a simulated path of the underlying asset according to the dynamics in equation 1 and the corresponding trajectories for the maximum and minimum price.

Analytical Solution

Under the GBM assumption, the distribution law of $m(T)$ (as well as the joint density of $m(T)$ and $S(T)$) is known in closed form. This allows one to obtain an analytical solution for standard lookback options as expected value of the discounted payoff; see, for example, [25, 26], and [15]. Therefore, we obtain the following pricing formula for the floating strike lookback call:

$$\begin{aligned} & \mathbb{E}_{t, S_t} e^{-r(T-t)} (S(T) - m(T)) \\ &= S_t e^{-q(T-t)} - e^{-r(T-t)} \mathbb{E}_{t, S_t} m(T) \\ &= S_t e^{-q(T-t)} \mathcal{N}(d_2) \end{aligned}$$

Table 1 Lookback option payoff function

Generic lookbacks			
	Minimum $m(T) = \min_{0 \leq u \leq T} S(u)$	Maximum $M(T) = \max_{0 \leq u \leq T} S(u)$	Range $M(T) - m(T)$
Standard lookbacks			
	Floating strike	Fixed strike	Reverse strike
Call	$(S(T) - m(T))^+$	$(M(T) - K)^+$	$(m(T) - K)^+$
Put	$(M(T) - S(T))^+$	$(K - m(T))^+$	$(K - M(T))^+$
Nonstandard lookbacks			
	Partial price	Partial time	conditions
Call	$(S(T) - \lambda m(T))^+$	$(S(T_2) - m(T_1))^+$	$\lambda \geq 1, T_1 < T_2$
Put	$(\gamma M(T) - S(T))^+$	$(M(T_1) - S(T_2))^+$	$\mu \leq 1, T_1 < T_2$

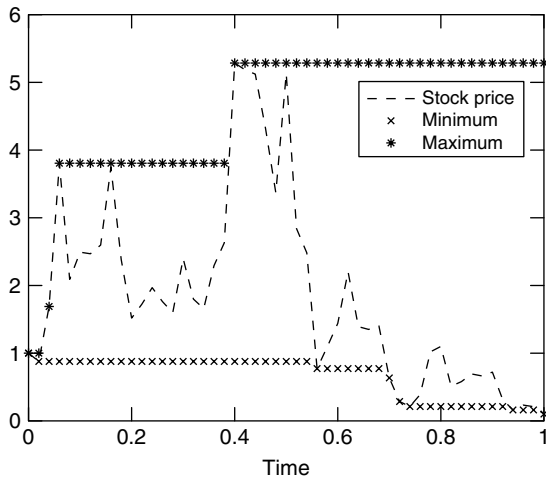


Figure 1 Geometric Brownian motion and its maximum and minimum to date

$$\begin{aligned}
 & -e^{-r(T-t)} m_t \mathcal{N}(d_2 - \sigma \sqrt{T-t}) \\
 & + \frac{\sigma^2 S_t}{2r} \left[e^{-r(T-t)} \left(\frac{S_t}{m_t} \right)^{-2r/\sigma^2} \mathcal{N}(d_3) \right. \\
 & \left. - e^{-q(T-t)} \mathcal{N}(-d_2) \right] \quad (3)
 \end{aligned}$$

with

$$\begin{aligned}
 d_2 = & \frac{1}{\sqrt{\sigma^2(T-t)}} \left(\ln \left(\frac{S_t}{m_t} \right) + (r-q)(T-t) \right. \\
 & \left. + \frac{1}{2} \sigma^2 (T-t) \right),
 \end{aligned}$$

$$d_3 = -d_2 + \frac{2(r-q)}{\sigma} \sqrt{T-t},$$

$$\mathcal{N}(x) = \int_{-\infty}^x e^{-\frac{u^2}{2}} du \quad (4)$$

Notice that the formula above admits a simplification when $t = 0$: indeed, in this case, we have $S(0) = m(0)$. Formula 3 suggests that the lookback call value is given by the sum of the premium of a plain vanilla call with strike equal to the current minimum and the premium of a so-called strike bonus option. This is the expected value of the cash flows necessary to exchange the initial option position for options with successively more favorable strikes. In other words, it measures the potential decrease in the price at which the option allows its holder to buy the security, if and when the security price attains a new minimum.

For pricing the fixed strike put option, we have

$$\begin{aligned}
 & \mathbb{E}_{t, S_t} e^{-r(T-t)} (K - m(T))^+ \\
 & = \mathbf{1}_{(K < m_t)} \left\{ -S_t \mathcal{N}(-d) + e^{-r(T-t)} \right. \\
 & \quad \times K \mathcal{N}(-d + \sigma \sqrt{T-t}) + \left(\frac{\sigma^2}{2r} \right) S_t \\
 & \quad \times \left[e^{-r(T-t)} \left(\frac{S_t}{K} \right)^{-2r/\sigma^2} \mathcal{N}(d_1) - \mathcal{N}(-d) \right] \Big\} \\
 & + \mathbf{1}_{(K \geq m_t)} \left\{ e^{-r(T-t)} (K - m_t) - S_t \mathcal{N}(-d_2) \right\}
 \end{aligned}$$

$$+ e^{-r(T-t)} m_t \mathcal{N}(-d_2 + \sigma \sqrt{T-t}) + \left(\frac{\sigma^2}{2r} \right) S_t \\ \times \left[e^{-r(T-t)} \left(\frac{S_t}{m_t} \right)^{-2r/\sigma^2} \mathcal{N}(d_3) - \mathcal{N}(-d_2) \right] \} \quad (5)$$

with

$$d = \frac{1}{\sqrt{\sigma^2(T-t)}} \left(\ln \left(\frac{S_t}{K} \right) + (r-q)(T-t) \right. \\ \left. + \frac{1}{2} \sigma^2 (T-t) \right), \\ d_1 = \frac{-d + 2(r-q)\sqrt{T-t}}{\sigma} \quad (6)$$

Lookback options on the maximum can be priced by exploiting the relation between maximum and minimum operators.

An important feature of lookbacks is the frequency of observation of the underlying assets for the purpose of identifying the best possible value for the holder. For example, the above expressions are consistent with the assumption that the underlying asset is monitored continuously. Instead, discrete monitoring refers to updating the maximum/minimum price at fixed times (e.g., daily, weekly, or monthly). In this case, we have to replace $M(T)$ and $m(T)$ by $\hat{M}(T)$ and $\hat{m}(T)$, defined as

$$\hat{M}(T) = \max_{0 \leq i \leq n} S(i\Delta), \text{ and } \hat{m}(T) = \min_{0 \leq i \leq n} S(i\Delta) \quad (7)$$

where n is the number of monitoring dates and Δ is the time distance between monitoring dates; with $n\Delta = T$. Nearly all closed-form expressions available for pricing path-dependent options are based on continuous-time paths, but many traded options are based on discrete price fixings. In general, a higher maximum/lower minimum occurs as long as the number n of monitoring dates increases. As noted by Heynen and Kat [29] and Aitsahlia and Leung [1], the discrepancy between option prices under continuous and discrete monitoring of the reference index may have significant effect on the prices of lookback options, but does not introduce new hedging problems. Indeed, the slow convergence of the discrete scheme to the continuous one as the number n of monitoring dates increases is well known. This figure is quantified in an order of proportionality of approximately $1/\sqrt{n}$.

For example, setting $r = 0.1$, $q = 0$, $\sigma = 0.2$, $T = 1$, $S_0 = 100$, the continuous formula returns 19.6456. Assuming a year consists of 250 days and that 10 monitoring dates are available (i.e., monitoring occurs approximately once a month), the discrete formula gives 17.0007: a percentage difference about 15% with respect to the continuous case. Using 10 000 monitoring dates (i.e., monitoring occurs once every 36 minutes), the discrete formula returns 19.5523, a small but still appreciable difference with respect to the continuous case.

Few papers have investigated the analytical pricing of discretely monitored lookback options. A correction to the continuously monitored formula based on the Riemann zeta function is given in [12]. The Riemann zeta function enters indeed in the computations of the moments of the discrete maximum/minimum; see, for example, [23, 27, 31, 32, 34], for different derivations and improvements of the above correction. New results for discrete options have been recently obtained exploiting the Wiener–Hopf factorization and the Spitzer’s identity. For example, [5] and [27] present an exact analytical formula showing how to cast the pricing problem in terms of an integral equation of the Wiener–Hopf type. This equation can be solved in closed form in the Gaussian case. The computational cost is linear in the number of monitoring dates. A related approach based on the Spitzer’s identity (the probabilistic interpretation of the solution of a Wiener–Hopf equation) has been advanced by Borovkov and Novikov [9] and by Petrella and Kou [39]. They propose an algorithm with a computational cost that is quadratic in the number of monitoring dates but has the advantages of being simply adapted to non-Gaussian models provided that they have independent identically distributed (i.i.d.) increments and the pricing formula of plain vanilla calls and puts are available.

Other approaches are mainly numerical and briefly detailed in the following subsections.

Finite Difference Method

The numerical solution of the partial differential equation (PDE) satisfied by the lookback option price is discussed for example in [42] and a detailed treatment for the discrete case is given in [3]. Since lookback options are path-dependent options, their value V does not depend only on the current spot

price and time, but also on the current realized minimum or maximum, and we can write $V = V(S, m, t)$ for options on the minimum (similar discussion holds for lookback on the maximum). Applying Ito's lemma and equating the expected return on the option to the return on a risk-free investment, it can be shown that V solves the following PDE:

$$\frac{\partial V}{\partial t} + (r - q)S \frac{\partial V}{\partial S} + \frac{\sigma^2}{2} S^2 \frac{\partial^2 V}{\partial S^2} = rV \quad (8)$$

which has to be solved for $S \geq m$ and $T \geq t \geq 0$. The above PDE is the standard Black–Scholes PDE, with the change of the domain from $S \geq 0$ to $S \geq m$. Here m appears as parameter delimiting the domain of the spot price. This implies that a boundary condition at $S = m$ is needed. The important point is the observation that when the spot price is near to the running minimum, the probability that at expiry the minimum will be equal to the current minimum m is zero and therefore changes in m do not affect the option value. This allows one to set the boundary condition at $S = m$:

$$\frac{\partial V(S, m, t)}{\partial m} = 0 \text{ when } S = m \quad (9)$$

Together with the payoff condition at $t = T$, equations (8) and (9) allow us to fully characterize the lookback option premium. For discretely monitored lookback options, with monitoring at dates $t_i = i\Delta$ the PDE (8) remains unchanged, while the boundary condition (9) does not apply anymore. Indeed, between monitoring dates, the spot price can freely move in $(0, +\infty)$ and at monitoring dates, the solution is updated according to the rule

$$V(S, m, t_i^+) = V(S, \min(S, m), t_i^-) \quad (10)$$

Equation (8) can be solved numerically using an appropriate numerical scheme such as the Crank–Nicolson one (see **Crank–Nicolson Scheme**). However, exploiting a change of numeraire, the PDE (8) can be simplified to a single state variable [3].

Monte Carlo Simulation

Discretely monitored lookback options can be easily priced by standard Monte Carlo (MC) simulation. The underlying price is simulated at all monitoring

dates exploiting the exact solution of the stochastic differential equation (1):

$$S^{(j)}((i+1)\Delta) = S^{(j)}(i\Delta)e^{(r-0.5\sigma^2)\Delta + \sigma\sqrt{\Delta}\epsilon_i^{(j)}} \quad (11)$$

where $\epsilon_i^{(j)}$ is a standard normal random variate and $S^{(j)}(i\Delta)$ is the spot price at time $t_i = i\Delta$ as sampled in the j th simulation.

The corresponding minimum price $m^{(j)}(i\Delta)$ over the time interval $t_{(i-1)\Delta}, t_{i\Delta}$ is updated at each monitoring date according to the rule

$$\hat{m}^{(j)}(i\Delta) = \min(S^{(j)}(i\Delta), \hat{m}^{(j)}((i-1)\Delta)) \quad (12)$$

with a starting condition $m_0^{(j)} = S_0$. The MC price for a lookback is given by the average of the discounted payoff computed over J simulated sample paths. For a lookback option with fixed strike K , the MC price is

$$e^{-rt_n} \frac{1}{J} \sum_{j=1}^J (K - \hat{m}^{(j)}(n\Delta))^+ \quad (13)$$

Similarly, for a floating strike lookback option, the MC price is

$$e^{-rt_n} \frac{1}{J} \sum_{j=1}^J (S^{(n\Delta)} - \hat{m}^{(j)}(n\Delta)) \quad (14)$$

Unfortunately, the procedure cannot be immediately applied to continuously monitored options shrinking the time step Δ . Indeed, owing to fact that we can only sample at discrete times, we lose information about the parts of the continuous-time path that lie between the sampling dates. This procedure will be systematically biased in the sense that the continuous minimum (maximum) will be always overestimated (underestimated). Andersen and Brotherton-Ratcliffe [4] show that for a one-year lookback with 256 discrete monitoring points this bias is around 5% of the option price and suggest a procedure to correct it.

Binomial and Tree Methods

As for PDEs, the implementation of a tree for path-dependent options involves two state variables and the need to keep track of current extreme values will cause the number of calculations to grow substantially faster than the number of nodes. However, under the GBM assumption and exploiting

a change of numeraire, the pricing of lookback options can be reduced to one-state binomial model (having a reflecting barrier at 0); see, for instance, [6] and [14], as well as **Tree Methods**. Using such models cuts the amount of computation remarkably and also makes it straightforward to deal with the early exercise feature.

Hedging

We conclude this section, mentioning the hedging problem of lookback options. When $r = q + \sigma^2/2$, lookback options can be exactly replicated using a self-financing strategy based on straddles (an ordinary put plus an ordinary call) with an exercise price equal to the initial extremum (maximum or minimum) [25]. If, over the life of the option, the stock price never rises above or falls below its initial value, the initial straddle would exactly satisfy the writer's terminal obligation. When the stock price is at an extremum and then achieves a new maximum or minimum, the straddle should be sold and a new portfolio established, a straddle with exercise price equal to the new maximum (minimum). This strategy is self-financed and replicates the lookback option. However, in general, the replicating strategy requires the computation of the Δ coefficient, that is, units of stocks needed to replicate the contingent claim, by taking the derivative of the pricing formula with respect to the current spot price. An alternative method is put forward by the use of the Malliavin calculus approach, and the main component for this method to work is a sort of representation theorem, called the *Clark–Ocone formula*. This formula allows us to identify a formal expression for the replicating portfolio of basically any contingent claim. This has been exploited in [8] to obtain the replicating portfolio.

A different approach is taken in [30] that finds bounds on the prices of the lookback option, in terms of the (market) prices of call options. This is achieved without making explicit assumptions about the dynamics of the price process of the underlying asset, but rather by inferring information about the potential distribution of asset prices from the call prices. Thus the bounds and the associated hedging strategies are model independent and represent limits on the possible price of the lookback, which are necessary for the absence of arbitrage.

Non-Gaussian Models

The GBM model in equation (1) is one of the most successful and widely used models in financial economics. Unfortunately, deviations between the model and empirical evidence are well known in literature and therefore offer new opportunities for the development of more-realistic models and pricing formulas for exotic options. With reference to lookback options, extensions have been obtained replacing the GBM model by the constant elasticity of variance (CEV) model (see **Constant Elasticity of Variance (CEV) Diffusion Model**) or by an exponential Lévy model (see **Exponential Lévy Models**). In the following sections we discuss such extensions.

CEV

The study of lookback options in the CEV model has started with [10] and [11]. Their approach consists in approximating the CEV process by a trinomial lattice and uses it to value barrier and lookback options numerically. A binomial tree is also used in [16]. While these approaches are purely numerical, in [20] closed-form solutions for the Laplace transforms of the probability distributions of the maximum and minimum are obtained. Lookback prices are recovered by inverting the Laplace transforms and integrating against the option payoff. An analytical inversion of the Laplace transforms for lookback options in terms of spectral expansions associated with the infinitesimal generator of the CEV diffusion is given in [36]. All these studies point out that the differences in prices of these exotic options under the CEV and geometric Brownian motion assumptions can be far more significant than the differences for standard European options.

Lévy Processes

Lookback options have been priced also under Lévy processes. For example, in [13] a very powerful algorithm is proposed that can be adapted for pricing discrete lookback under a jump-diffusion model. For more general Lévy processes, very efficient algorithms have been proposed in [22, 39] and [24]. An analytical expression in terms of the Laplace transform is found for continuously monitored options, under a double-exponential model, in [35].

Related Payoff

Several exotic options can be thought of as a modification of lookback options. We have mentioned a few of them here.

In *partial lookback* options, the lookback feature is limited to only the first (for entry timing) or the last (for exit timing) part of the options' life. This product, even with a relatively short lookback periods, appears to offer a good solution to most timing problems at a reasonable price. Kat and Heynen [33] provide closed-form pricing formulas for such options. Analysis shows how the prices of such "partial lookback options" respond to a change in the monitoring period.

Quanto lookback refers to a payoff structure where the terminal payoff of the quanto option depends on the realized extreme value of a stock denominated in a foreign currency but the payoff is paid in a domestic currency. These contracts have been studied in [18].

Double lookbacks or *Range options* include calls and puts with the underlying being the difference between the maximum and minimum prices of one asset over a certain period, and calls or puts with the underlying being the difference between the maximum prices of two correlated assets over a certain period. Analytical expressions of the joint probability distribution of the maximum and minimum values of two correlated geometric Brownian motions are derived in [28] and used in the valuation of double lookbacks. An option on the spread between the maximum and minimum price of a single stock over a given interval of time captures the idea of an option on price volatility; see, for example, the related literature on financial econometrics devoted to estimate the volatility using range estimators. Therefore, such an option might be of interest to traders who want to bet on price volatility or hedge an existing position that is sensitive to price volatility.

Quantile options have a payoff at maturity depending on the order statistics of the underlying asset price. The j -quantile process $q(n, j)$ is defined as the level at which the return process stays below for j periods out of n . In particular, we have $q(n, n) = \hat{M}(n\Delta)$ and $q(n, 0) = \hat{m}(n\Delta)$. The quantile payoff is obtained replacing the extreme appearing in Table 1 by the quantile $q(n, j)$. These options can be priced by making use of what is known as the *Dassios–Port–Wendel identity* [19], that allows one

to write the quantile as sum

$$q(n, j) \stackrel{d}{=} \hat{m}((n - j)\Delta) + \hat{M}(j\Delta) \quad (15)$$

where $\stackrel{d}{=}$ means equality in distribution and $\hat{m}((n - j)\Delta)$ and $\hat{M}(j\Delta)$ are independent processes. From this identity, it follows that the density of the quantile is the convolution of the densities of $\hat{m}((n - j)\Delta)$ and $\hat{M}(j\Delta)$ and, at the issue date of the contract, the quantile price can be obtained as expected discounted payoff under the risk-neutral measure; see [5] and [19] for details. Another way of exploiting the Dassios–Port–Wendel identity consists in, conditioning on the minimum, representing the quantile option price as average of lookback option prices written on the maximum and with random strike. The average is taken with respect to the density of the discrete minimum. Finally, we remark that the pricing off the inception date is not straightforward, because the quantile process is non-Markovian. A discussion can be found in [19]. Other relevant references are [2, 38], and [7].

A *drawdown (drawup) option* is defined as the drop (increase) of the asset price from its running maximum (minimum), $D(t) = M(t) - S(t)$, ($U(t) = S(t) - m(t)$). Maximum drawdown $MDD(t)$ is defined as the maximal drop of the asset price from its running maximum over a given period of time:

$$MDD(t) = \max_{0 \leq s \leq t} D(s) \quad (16)$$

In a similar manner, we can define the maximum drawup. Maximum drawdown measures the worst loss of an investor who enters the market at a certain point and leaves it at some following point within a given time period; this means that he or she buys the asset at a local maximum and sells it at the subsequent lowest point, and this drop is the largest in the given time period. A derivative contract on the maximum drawdown, introduced in [41], can serve as an important risk measure indicator: when the market is in a bubble, it is reasonable to expect that the prices of drawdown contracts would be significantly higher. On the other hand, when the market is stable, or when it exhibits mean reversion behavior, the prices of drawdown contracts would become cheaper. When the market experiences a crash, the lookback option may expire close to worthless if the final asset value is near its running maximum. Momentum traders believe

that the realized maximum drawdown (maximum drawup) will be larger than expected, and thus they are natural buyers of this contract. On the other hand, selling the (unhedged) contract is equivalent to taking the opposite strategy, namely, buying the asset when it is setting its new low. This is known as *contrarian trading*. Contrarian traders believe that the realized maximum drawdown (maximum drawup or range) will be smaller than expected, and they are natural sellers of this contract. The distribution of the maximum drawdown of Brownian motion is studied in [37].

Russian options are perpetual American Options with lookback payoff, introduced in [40]. Russian options can be regarded as a kind of perpetual American fixed strike lookback option with zero strike price and their pricing can be derived by using a probability approach [21], or a PDE approach [17].

Finally, we mention the class of structures named *mountain options*, having names such as Himalaya, Everest, Altiplano and so on (see **Atlas Option**; **Himalayan Option**; **Altiplano Option**). Here, the extrema over a given period of a given asset is replaced by the best or the worst performer over different periods of assets in a given basket. Sometimes, a global floor on the return of the product is also introduced. It is clear that MC simulation is needed to be able to price this type of products, that, in general, are very sensible to the crosscorrelation of the assets. In addition, the Greeks of these contracts can change markedly as the trade progresses.

References

- [1] Aitsahlia, F. & Leung, L.T. (1998). Random walk duality and the valuation of discrete lookback options, *Applied Mathematical Finance* **5**(3/4), 227–240.
- [2] Akahori, J. (1995). Some formulae for a new type of path-dependent option, *Annals of Applied Probability* **5**, 383–388.
- [3] Andreasen, J. (1998). The pricing of discretely sampled Asian and lookback options: a change of numeraire approach, *Journal of Computational Finance* **2**(1), 5–30.
- [4] Andersen, L. & Brotherton-Ratcliffe, R. (1996). Exact exotics, *Risk Magazine* **9**, 85–89.
- [5] Atkinson, C. & Fusai, G. (2007). Discrete extrema of Brownian motion and pricing of exotic options, *Journal of Computational Finance* **10**(3), 1–43.
- [6] Babbs, S. (2000). Binomial valuation of lookback options, *Journal of Economic Dynamics and Control* **24**(11–12), 1499–1525.
- [7] Ballotta, L. & Kyprianou, A. (2001). A note on the Alpha-Quantile option, *Applied Mathematical Finance* **8**, 137–144.
- [8] Bermin, H.-P. (2000). Hedging lookback and partial lookback options using Malliavin calculus, *Applied Mathematical Finance* **39**, 75–100.
- [9] Borovkov, K. & Novikov, A. (2002). On a new approach for option pricing, *Journal of Applied Probability* **39**, 1–7.
- [10] Boyle, P.P. & Tian, Y. (1999). Pricing lookback and barrier options under the CEV process, *Journal of Financial and Quantitative Analysis* **34**, 241–264.
- [11] Boyle, P.P., Tian, Y. & Imai, J. (1999). Lookback options under the CEV process: a correction, *Journal of Finance and Quantitative Analysis* web site <http://www.jfqa.org/>. In: Notes, Comments, and Corrections.
- [12] Broadie, M., Glasserman, P. & Kou, S. (1999). Connecting discrete and continuous path-dependent options, *Finance and Stochastics* **3**, 55–82.
- [13] Broadie, M. & Yamamoto, Y. (2005). A double-exponential fast Gauss transform algorithm for pricing discrete path-dependent options, *Operations Research* **53**(5), 764–779.
- [14] Cheuck, T.H.F. & Vorst, T.C.F. (1997). Currency lookback options and observation frequency: a binomial approach, *Journal of International Money and Finance* **16**(2), 173–187.
- [15] Conze, A. & Vishwanathan, R. (1991). Path-dependent options: the case of lookback options, *Journal of Finance* **46**, 1893–1907.
- [16] Costabile, M. (2006). On pricing lookback options under the CEV process, *Decisions in Economics and Finance* **29**, 139–153.
- [17] Dai, M. (2000). A closed-form solution for perpetual American floating strike lookback options, *Journal of Computational Finance* **4**(2), 63–68.
- [18] Dai, M., Kwok, Y.K. & Wong, H.Y. (2004). Quanto lookback options, *Mathematical Finance* **14**(3), 445–467.
- [19] Dassios, A. (1995). The distribution of the quantile of a Brownian motion with drift & the pricing of related path-dependent options, *Annals of Applied Probability* **5**, 389–398.
- [20] Davydov, D. & Linetsky, V. (2001). The valuation and hedging of barrier and lookback options under the CEV process, *Management Science* **47**, 949–965.
- [21] Duffie, D. & Harrison, J.M. (1993). Arbitrage pricing of Russian options and perpetual lookback options, *The Annals of Applied Probability* **3**(3), 641–651.
- [22] Feng, L. & Linetsky, V. (2009). Computing exponential moments of the discrete maximum of a Levy process and lookback options, *Finance and Stochastics*, available at SSRN:<http://ssrn.com/abstract=1260934>.
- [23] Fusai, G., Abrahams, I.D. & Sgarra, C. (2006). An exact analytical solution for discrete barrier options, *Finance and Stochastics* **10**(1), 1–26.

- [24] Fusai, G., Marazzina, D., Marena, M. & Ng, M. (2008). *Maturity Randomization and Option Pricing*. w.p. SEMeQ.
- [25] Goldman, M.B., Sosin, H.B. & Gatto, M.A. (1979). Path-dependent options: buy at the low, sell at the high, *Journal of Finance* **34**, 1111–1127.
- [26] Goldman, M.B., Sosin, H.B. & Shepp, L. (1979). On contingent claims that insure ex-post optimal stock market timing, *Journal of Finance* **34**, 401–413.
- [27] Green, R., Fusai, G. & Abrahams, I.D. (2009). The Wiener-Hopf technique & discretely monitored path dependent option pricing, *Mathematical Finance*, to appear.
- [28] He, H., Keirstead, W. & Rebholz, J. (1998). Double lookbacks, *Mathematical Finance* **8**, 201–228.
- [29] Heynen, R.C. & Kat, H.M. (1995). Lookback options with discrete and partial monitoring of the underlying price, *Applied Mathematical Finance* **2**, 273–284.
- [30] Hobson, D.G. (1998). Robust hedging of the lookback option, *Finance and Stochastics* **2**(4), 329–347.
- [31] Hörfelt, P. (2003). Extension of the corrected barrier approximation by Broadie, Glasserman, and Kou, *Finance and Stochastics* **7**(2), 231–243.
- [32] Howison, S. & Steinberg, M. (2007). A matched asymptotic expansions approach to continuity corrections for discretely sampled options. Part 1: barrier options, *Applied Mathematical Finance* **14**, 63–89.
- [33] Kat, H.M. & Heynen, R.C. (1994). Selective memory. *Risk Magazine* **7**(11), 73–76.
- [34] Kou, S.G. (2003). On pricing of discrete barrier options, *Statistica Sinica* **13**, 955–964.
- [35] Kou, S.G. & Wang, H. (2003). First passage times of a jump diffusion process, *Advances in Applied Probability* **35**, 504–531.
- [36] Linetsky, V. (2004). Lookback options and diffusion hitting time: a spectral expansion approach, *Finance and Stochastics* **8**, 373–398.
- [37] Magdon-Ismail, M., Atiya, A., Pratap, A. & Abu-Mostafa, Y. (2004). On the maximum drawdown of a Brownian motion, *Journal of Applied Probability* **41**(1), 147–161.
- [38] Miura, R. (1992). A note on lookback options based on order statistics, *Hitotsubashi Journal of Commerce & Management* **27**, 15–28.
- [39] Petrella, G. & Kou, S.G. (2004). Numerical pricing of discrete barrier and lookback options via Laplace transforms, *Journal of Computational Finance* **8**, 1–37.
- [40] Shepp, L. & Shiryaev, A.N. (1993). The Russian option: reduced regret, *Annals of Applied Probability* **3**, 631–640.
- [41] Vecer, J. (2006). Maximum drawdown and directional trading, *Risk Magazine* **19**(12), 88–92.
- [42] Wilmott, P., Dewynne, J.N. & Howison, S. (1993). *Option Pricing: Mathematical Models and Computation*, Oxford Financial Press.

Related Articles

Barrier Options; Corridor Options; Discretely Monitored Options; Parisian Option.

GIANLUCA FUSAI

Parisian Option

Parisian options are barrier options that are activated or canceled—depending on the type of option—if the underlying asset has been continuously traded above or below the barrier level long enough. A down-and-out Parisian option denotes a contract that expires worthless if the underlying asset reaches a prespecified level L and remains constantly below this level for a time interval longer than a fixed number D , called the *window*. Its price (for a call option) at time 0 is given by

$$\phi(T, K) := e^{-rT} \mathbb{E} \left[(S_T - K)^+ \mathbb{1}_{T_L^{D,-}(S) > T} \right] \quad (1)$$

where $T_L^{D,-}(S)$ is the first time the asset S makes an excursion longer than D below L . Parisian-style options are mostly encountered in convertible bonds with “soft-call” provision for conversion. For example, the bond’s specifications may be such that conversion will be allowed if and only if the share price remains above a theoretical price for a given amount of time, for example, 20 business days prior to the conversion date (this is Parisian option). Other covenants stipulate that the average share price trades for n days above the trigger level. While the latter does not correspond *sensu stricto* to a Parisian option, the motivation is similar: to render the conversion rule more stable—and less prone to manipulation—basing it on the behavior of the stock price over a window of time as opposed to basing it on the (more volatile) spot price. The pioneer paper on that topic is owed to Chesney *et al.*, [5]. Pricing Parisian options is a challenging issue and several methods have been proposed in the literature: Monte Carlo simulations, Laplace transforms, lattices, and partial differential equations.

Monte Carlo Method

As for standard barrier options, using simulations leads to a biased problem, owing to the choice of the discretization time step in the Monte Carlo algorithm. Baldi *et al.*, [3] have developed a method based on sharp large deviation estimates, which improves the usual Monte Carlo procedure. It consists in providing an approximation of p^ϵ —the probability that a Brownian bridge reaches a time-dependent barrier

over a time of length ϵ —by studying its asymptotic behavior when ϵ tends to 0. They derive precise estimates of $g_{L,t}^S := \sup\{u \leq t | S_u = L\}$, which is, for a down-and-out Parisian option, related to $T_L^{D,-}(S)$ by the following formula: $T_L^{D,-}(S) := \inf\{t > 0 : (t - g_{L,t}^S) \mathbb{1}_{S_t < L} > D\}$. This procedure still works when the asset follows a diffusion process with general coefficients.

Laplace Transforms

The idea of using Laplace transforms for pricing Parisian options is owed to Chesney *et al.*, [5]. By using the Brownian excursion theory, they get closed formulas for

$$\int_0^\infty dt e^{-\lambda t} \phi(t, K) \quad (2)$$

the Laplace transform of the price with respect to the maturity time.

For models with constant parameters, when considering a down and in call option, one rewrites $\phi(T, K)$ as

$$e^{-\left(r + \frac{m^2}{2}\right)T} \mathbb{E}_{\mathbb{P}} \left(\mathbb{1}_{T_b^{D,-} < T} (xe^{\sigma Z_T} - K)^+ e^{mZ_T} \right) \quad (3)$$

where Z is a $\mathbb{P}$ -Brownian motion, m depends on r and σ , and $T_b^{D,-} := T_b^{D,-}(Z)$ is the first time Z makes an excursion below $b := \frac{1}{\sigma} \log(L/S_0)$ longer than D . By using the Brownian motion excursion theory, notably the Azéma martingale and the Brownian meander, the density of $Z_{T_b^{D,-}}$ can be obtained and it can be shown that $T_b^{D,-}$ and $Z_{T_b^{D,-}}$ are independent. There is no explicit formula for the density of $T_b^{D,-}$, but we only know its Laplace transform. The strong Markov property enables to introduce $Z_{T_b^{D,-}}$ in equation (3). We rewrite equation (3) as

$$\int_{-\infty}^\infty \mathbb{E}_{\mathbb{P}} (\mathbb{1}_{T_b^{D,-} < T} \mathcal{P}_{T-T_b^{D,-}}(f_x)(z)) v^-(dz) \quad (4)$$

where v^- denotes the law of $Z_{T_b^{D,-}}$, $f_x(z) = e^{-(r+m^2/2)T} e^{mz} (xe^{\sigma z} - K)^+$, and $\mathcal{P}_t(f_x)(z) = 1/\sqrt{2\pi t} \int_{-\infty}^\infty f_x(u) \exp(-(u-z)^2/2t) du$. It remains to compute the Laplace transform of equation (4) with respect to the maturity. A change of variables

introduces the Laplace transform of $T_b^{D,-}$, which is explicitly known. This leads to a closed formula.

We refer the reader to [1] for the description of a fast and accurate numerical inversion of the Laplace transforms. By studying the regularity of the Parisian option prices with respect to the maturity time, Labart and Lelong [9] justify the accuracy of the numerical inversion. Except for particular values of the barrier, the prices are of class C^∞ . Their study relies on the existence and the regularity of a density for the Parisian time $T_b^{D,-}$.

This algorithm is implemented in [4] and is compared to a procedure for approximating a general Laplace transform with one that can be easily inverted. The Laplace transform approach is very specific to the problem, but practically we see that the lack in the flexibility of the method is compensated by its accuracy and computational speed.

Lattices

Costabile [6] presents a discrete time algorithm to evaluate Parisian options. The evaluation method is based on a combinatorial approach used to count the number of trajectories of a particle which, moving in a binomial lattice, remains constantly above an upper barrier for time intervals strictly smaller than a prespecified window period. Once this number has been computed, it can be used to derive a binomial algorithm, based on the Cox–Ross–Rubinstein (CRR) model (*see Binomial Tree or Tree Methods*). It enables to evaluate Parisian options with a constant or an exponential barrier. Avellaneda and Wu [2] model and price Parisian-style options by a trinomial lattice method, which changes with the value of the asset with respect to the barrier.

Partial Differential Equations

Pricing of Parisian options can be done using partial differential equations. Let τ define the time the underlying asset has continuously spent in the excursion. For a down Parisian option, $\tau := t - \sup\{t' \leq t | S_{t'} \geq L\}$. The dynamics of τ is

$$d\tau_t = \begin{cases} dt & \text{if } S_t < L, \\ -\tau_t & \text{if } S_t = L, \\ 0 & \text{if } S_t < L \end{cases} \quad (5)$$

The new state variable τ can be viewed as a clock that starts ticking as soon as the share price crosses the barrier level and is immediately reset when the share price returns above L . We assume that the asset follows a log normal Brownian motion given by $dS_t = \mu S_t dt + \sigma S_t dW_t$. The option price is a function of S, t, τ . If $S \geq L$, the governing equation is the standard Black Scholes equation:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} - rV = 0 \quad (6)$$

If $S \leq L$, τ is ticking. The new governing equation is

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} + \frac{\partial V}{\partial \tau} - rV = 0 \quad (7)$$

The boundary conditions are the following: the path-wise continuity of V in $S = L$ leads to $V(L, t, \tau) = V(L, t, 0)$ for all t , and

$$\begin{aligned} V(S, T, \tau) &= (S_T - K)^+ & \text{if } \tau < D, \\ V(S, T, \tau) &= 0 & \text{otherwise} \end{aligned} \quad (8)$$

In the study of Haber *et al.* [8], the numerical solution to equations (6) and (7) is implemented using an explicit finite difference scheme. In the case of a discrete monitoring of the contract, Vetzal and Forsyth [7] develop an algorithm based on the numerical solution of a system of one-dimensional PDEs. It is assumed that τ only changes at observation dates with the value of S with respect to the barrier. Away from observation dates, the PDE satisfied by V does not depend on τ . Then, the pricing problems consist of a small number of one-dimensional PDEs, which exchange information only at observation dates (we impose the continuity of V).

These methods have one major benefit: they are flexible enough to be easily modified to price more general options, like Parisian (i.e., when the recorded duration is cumulative rather than continuous).

Double Parisian

There exists a double barrier version of the standard Parisian options. Double Parisian options are barrier options that are activated or canceled if the underlying asset continuously remains outside a range

$[L_1, L_2]$ long enough. The price of a double Parisian out call at time 0 is given by

$$e^{-rT} \mathbb{E} \left[(S_T - K)^+ \mathbb{1}_{T_{L_1}^{D,-}(S) > T} \mathbb{1}_{T_{L_2}^{D,+}(S) > T} \right] \quad (9)$$

These double Parisian options can be priced using the Monte Carlo procedure improved with the sharp large deviation method proposed by Baldi, Caramellino, and Iovino [3]. Labart and Lelong [9] give analytical formulas for the Laplace transforms of the prices with respect to the maturity time.

References

- [1] Abate, J., Choudhury, L.G. & Whitt, G. (1999). An introduction to numerical transform inversion and its application to probability models, in *Computational Probability*, W. Grassman, ed., Kluwer, Boston, pp. 257–323.
- [2] Avellaneda, M. & Wu, L. (1999). Pricing Parisian-style options with a lattice method, *International Journal of Theoretical and Applied Finance* **2**(1), 1–16.
- [3] Baldi, P., Caramellino, L. & Iovino, M.G. (2000). Pricing complex barrier options with general features using sharp large deviation estimates, in *Monte Carlo and Quasi-Monte Carlo Methods 1998 (Claremont, CA)*, Springer, Berlin, pp. 149–162.
- [4] Bernard, C., LeCourtis, O. & Quittard-Pinon, F. (2005). A new procedure for pricing Parisian options, *The Journal of Derivatives* **12**(4), 45–53.
- [5] Chesney, M., Jeanblanc-Picqué, M. & Yor, M. (1997). Brownian excursions and Parisian barrier options, *Advances in Applied Probability* **29**(1), 165–184.
- [6] Costabile, M. (2002). A combinatorial approach for pricing Parisian options, *Decisions in Economics and Finance* **25**(2), 111–125.
- [7] Forsyth, P.A. & Vetzal, K.R. (1999). Discrete Parisian and delayed barrier options: A general numerical approach, *Advances in Futures Options Research* **10**, 1–16.
- [8] Haber, R.J., Schonbucher, P.J. & Wilmott, P. (1999). Pricing Parisian options, *Journal of Derivatives* **6**(3), 71–79.
- [9] Labart, C. & Lelong, J. Pricing Double Parisian options using Laplace transforms, *International Journal of Theoretical and Applied Finance* (to appear), <http://hal.archives-ouvertes.fr/hal-00220470/fr/>.

Related Articles

Barrier Options; Discretely Monitored Options; Finite Difference Methods for Barrier Options; Lattice Methods for Path-dependent Options; Partial Differential Equations.

CÉLINE LABART

Cliquet Options

Cliquet options can be broadly characterized as contracts whose economic value depends on a series of periodic settlement values. Each settlement period has an associated strike whose value is set at the beginning of the period. This periodic resetting of the strike allows the cliquet option to remain economically sensitive across wide changes in market levels.

The Market for Cliquet Options

The early market in cliquet options featured vanilla contracts that were simply a series of forward starting at-the-money options. Rubinstein [4] provided pricing formulae for forward-start options in a Black–Scholes framework resulting in Black–Scholes pricing for vanilla cliquets. Cliquet products now trade on exchanges and the fore-runner to these listings were reset warrants, whose first public listings in the United States appeared in 1993 [5] and 1996 [1, 2]. Cliquet options are equally effective in capturing bullish (call) and bearish (put) market sentiments.

The current market for cliquet options accommodates a rich variety of features, which are sometimes best illuminated in discussions of pricing methods [6, 7]. The most actively traded cliquets are return-based products that accumulate periodic settlement values and pay a cash flow at maturity. The return characteristics and the price appeal of a cliquet can be tailored by adding caps and floors to the period returns and by introducing a strike moniness factor different from one. Defining the i th settlement value, R_i , in a call style cliquet by

$$R_i = \text{Max} \left(\text{floor}_i, \text{Min} \left(\frac{S_i}{S_i^0} - k_i, \text{cap}_i \right) \right) \quad (1)$$

where floor_i is the one-period(local) return floor for period i ; cap_i is the one-period(local) return cap for period i ; S_i is the market level on the settlement date for period i ; S_i^0 is the market level on the strike setting date for period i ; and k_i is a strike moniness factor for period i .

The payoff at maturity is given by

$$\text{payoff} = \text{Notional} \cdot \text{Max} \left(GF, \text{Min} \left(\sum_i^n R_i, GC \right) \right) \quad (2)$$

where GF is a global floor; GC is a global cap; and notional is the principal amount of the investment.

The investor forgoes returns above the local cap and is protected against returns below the local floor. For the same investment cost, investors can participate in more of the upside return by raising the local cap at the expense of a lowered local floor and the increased exposure to downside returns.

Applications of Cliquet Options

The periodic strike setting feature of a cliquet enables an investor to implement a strategy consistent with rolling options positions but without exposure to volatility movements. For example, an investor could buy a cliquet to implement a rolling three-month put strategy and be immunized against the future increase in options premiums that would accompany increases in volatility throughout the life of the strategy. Hence a cliquet provides cost certainty, whereas the rolling put strategy does not.

Cliquet products are often embedded in principal-protected notes, which combine certain aspects of fixed-income investing with equity investing. These notes guarantee the return of principal at maturity with the investment upside provided by the cliquet return. Retail notes would generally base investment gains on a broad market index such as the S&P 500 index. Principal-protected notes may further guarantee a minimum investment yield, which compounds to the value of the global floor at maturity. The guaranteed yield may be considered as part of the equity return, as it is in equation (2), or it can be considered as part of the fixed-income return. In the latter case, the equity payoff in equation (2) would be modified as in equation (3):

$$\text{payoff} = \text{Notional} \cdot \text{Max} \left(0, \text{Min} \left(\sum_i^n R_i, GC \right) - GF \right) \quad (3)$$

where the global floor now sets a strike on the sum of periodic returns.

Summary

We have discussed the general characteristics of cliquet options and illustrated the payoff for one commonly traded type of the cliquet. Numerous variations

exist and can be tailored to give very different risk-reward profiles. Some are distinguished in the market by specific names, for example reverse cliquets [3]. The customizability of cliquet options likely means we will continue to see product innovation in this area in the future.

References

- [1] Conran, A. (1996). *IFC Issues S&P 500 Index Bear Market Warrants*, November 26, 1996 Press Release, <http://www.ifc.org/ifcext/media.nsf/Content/PressReleases>.
- [2] Gray, S.F. & Whaley, R.E. (1997). Valuing S&P 500 bear market warrants with a periodic reset, *Journal of Derivatives* **5**(1), 99–106.
- [3] Jeffrey, C. (2004). Reverse cliquets: end of the road? *RISK* **17**(2), 20–22.
- [4] Rubinstein, M. (1991). Pay now, choose later, *RISK* **4**, 13.
- [5] Walmsley, J. (1998). *New Financial Instruments*, 2nd edition, John Wiley & Sons, New York.
- [6] Wilmott, P. (2002). Cliquet options and volatility models, *Wilmott Magazine*, 6.
- [7] Windcliff, H., Forsyth, P.A. & Vetzal, K.R. (2006). Numerical methods and volatility models for valuing cliquet options, *Applied Mathematical Finance* **13**, 353.

RICK L. SHYPIT

Basket Options

Equity basket options are derivative contracts that have as underlying asset a basket of stocks. This category may include (broadly speaking) options on indices as well as options on exchange-traded funds (ETFs), as well as options on bespoke baskets. The latter are generally traded over the counter, often as part of, or embedded in, structured equity derivatives.

Options on broad market ETFs, such as the Nasdaq 100 Index Trust (QQQQ) and the S&P 500 Index Trust (SPY), are the most widely traded contracts in the US markets. As of this writing, their daily volumes far exceed those of options on most individual stocks. Owing to this wide acceptance, QQQQ and ETF options have recently been given quarterly expirations in addition to the standard expirations for equity options. Options on sector ETFs, such as the S&P Financials Index (XLF) or the Merrill Lynch HOLDR (SMH), are also highly liquid.

If we denote by B the value of the basket of stocks at the expiration date of the option, a basket call has payoff given by $\max(B - K, 0)$ and a basket put has payoff $\max(K - B, 0)$, where K is the strike price. Most exchange-traded ETF options are physically settled. Index options tend to be cash settled. Over-the-counter basket options, especially those embedded in structured notes, are cash settled.

The fair value price of a (bespoke) basket option is determined by the joint risk-neutral distribution of the underlying stocks. If we write the value of the basket as

$$B = \sum_{i=1}^n w_i S_i \quad (1)$$

where w_i , S_i denote respectively the number of shares of the i th stock and its price, the returns satisfy

$$\frac{dB}{B} = \sum_{i=1}^n \frac{w_i S_i}{B} \frac{dS_i}{S_i} = \sum_{i=1}^n p_i \frac{dS_i}{S_i}, \quad \text{with} \quad p_i \equiv \frac{w_i S_i}{B} \quad (2)$$

Here, p_i represents the instantaneous capitalization weight of the i th stock in the basket, that is, the percentage of the total dollar amount of the

basket associated with each stock. If we assume that these weights are approximately constant which is reasonable, it follows that the volatility of the basket and the volatilities of the stocks satisfy the relation

$$\sigma_B^2 = \sum_{i,j=1}^n p_i p_j \sigma_i \sigma_j \rho_{ij} \quad (3)$$

where σ_B is the volatility of the basket, σ_i are the volatilities of the stocks, and ρ_{ij} is the correlation matrix of stock returns. If we assume lognormal returns for the individual stocks, then the probability distribution for the price of the basket is not lognormal. Nevertheless, the distribution is well approximated by a lognormal and equation (3) represents the natural approximation for the implied volatility of the basket in this case.

The notion of *implied correlation* is sometimes used to quote basket option prices. The market convention is to assume (for quoting purposes) that $\rho_{ij} \equiv \rho$, a constant. It then follows from equation (1) that the implied correlation of a basket option is

$$\begin{aligned} \rho &\equiv \frac{\sigma_B^2 - \sum_{i=1}^n p_i^2 \sigma_i^2}{\sum_{i \neq j} p_i p_j \sigma_i \sigma_j} = \frac{\sigma_B^2 - \sum_{i=1}^n p_i^2 \sigma_i^2}{\left(\sum_{i=1}^n p_i \sigma_i \right)^2 - \sum_{i=1}^n p_i^2 \sigma_i^2} \\ &\approx \frac{\sigma_B^2}{\left(\sum_{i=1}^n p_i \sigma_i \right)^2} \end{aligned} \quad (4)$$

Implied correlation is the market convention for quoting the implied volatility of a basket option as a fraction of the weighted average of implied volatilities of the components.

For example, if the average implied volatility for the components of the QQQQ for the December at-the-money options is 25% and the corresponding QQQQ option is trading at an implied volatility of 19%, the implied correlation is $\rho \approx (19/25)^2 = 58\%$.

This convention is sometimes applied to options that are not at the money as well. In this case, in the calculation of implied correlation for the basket option, the implied volatilities for the component stocks are usually taken to have the same moneyness as the index in percentage terms. Other

2 Basket Options

conventions for choosing the volatilities of the components, such as equal-delta or “beta-adjusted” money, are sometimes used as well. Since the corresponding implied correlations can vary with strike price, market participants sometimes talk about the *implied correlation skew* of a series of basket options.

Further Reading

Avellaneda, M., Boyer-Olson, D., Busca, J. & Friz, P. (2002). Reconstructing volatility, *Risk* **15**(10).

Haug, E.G. (1998). *The Complete Guide to Option Pricing Formulas*, McGraw-Hill.

Hull, J. (1993). *Options Futures and Other Derivative Securities*, Prentice Hall Inc., Toronto.

Related Articles

Correlation Swap; Exchange-traded Funds (ETFs).

MARCO AVELLANEDA

Call Spread

A call spread is an option strategy with limited upside and limited downside that uses call options of two different strikes but the same maturity on the same underlying. More details and pricing models can be found in [1]. Market considerations can be found in [3–5]. The call spread produces a structure that at maturity pays off only in scenarios where the price of the underlying is above the lower strike. One can think of this strategy as buying a low-strike call option and financing part of the upfront cost by selling a higher strike call option. The effect of selling the higher strike option is to limit the upside potential, but reduce the cost of the structure. It should be used for expressing a bullish view that the underlying will rise in price above the lower strike. As with all options, choosing the strike and maturity will depend on one's view of how much the underlying will move and how quickly it will move there. An example is shown, in detail, in Figure 1.

In the example shown in Figure 1, we look at a 790/810 call spread on the S&P 500 index, SPX. With the underlying SPX index at 770 and with three months to expiration, a 790 strike call price is 49.44 and an 810 strike call price is 41.79. The spread cost is $49.44 - 41.79 = 7.65$. Thus, the cost for a call spread is significantly reduced from the outright cost of a call option with the same strike. This upfront cost for the call spread is the most one can lose in a call spread. We subtract this initial investment from all other valuations, as shown in Figure 1, to get a total value. On the other hand, if both options expire in the money, one will earn $20 = 810 - 790$ on the call spread. Then the maximum profit is the spread minus the initial cost, or $20 - 7.65 = 12.35$.

Note that with three months to expiration, the call spread value is fairly insensitive to the underlying price. However, as the option gets closer to expiration, the sensitivity of the price of the call spread becomes greater, especially in the range of the spread itself. This sensitivity to underlying price or delta (*see Delta Hedging*) is illustrated in Figure 2.

The delta of the call spread stays relatively flat until relatively close to expiration of the option. When the call spread is close to expiration, the delta is very unstable around the two strikes.

Another important consideration is the volatility implied by the market (*see Implied Volatility: Market Models*). When the strikes are out of the money, the call spread price increases when volatility increases because the probability of finishing in the money increases. On the other hand, when the strikes are in the money, the call spread price decreases when volatility increases because the probability of finishing out of the money increases. This is illustrated in Figure 3.

The Relationship with Digital Options and Skew

The value of a European call spread structure can be written in terms of the difference of two call options. If we let $p(S)$ denote the probability distribution of the underlying at the time of option expiry, then we have

$$\begin{aligned} CallSpread &= e^{-r\tau} \int_{K_1}^{\infty} (S - K_1) p(S) dS - e^{-r\tau} \\ &\quad \times \int_{K_2}^{\infty} (S - K_2) p(S) dS \end{aligned} \quad (1)$$

$$\begin{aligned} CallSpread &= e^{-r\tau} \int_{K_2}^{\infty} (K_2 - K_1) p(S) dS + e^{-r\tau} \\ &\quad \times \int_{K_1}^{K_2} (S - K_1) \cdot p(S) dS \end{aligned} \quad (2)$$

If we now take the strikes very close to each other, the second term becomes insignificant. Next, if we lever up by $1/(K_2 - K_1)$, the payoff approximates the payoff of a digital option, which pays one if the underlying at termination is greater than the strike and zero otherwise. In this case,

$$\begin{aligned} DigitalOption &= e^{-r\tau} \int_K^{\infty} p(S) dS \\ &= e^{-r\tau} (1 - \Phi(K)) \end{aligned} \quad (3)$$

where $\Phi(K)$ is the cumulative probability distribution at termination of the underlying. For the original paper, see [6]. Also see [2, 7, 8] for more details. We can state equation (3) in words as follows. The

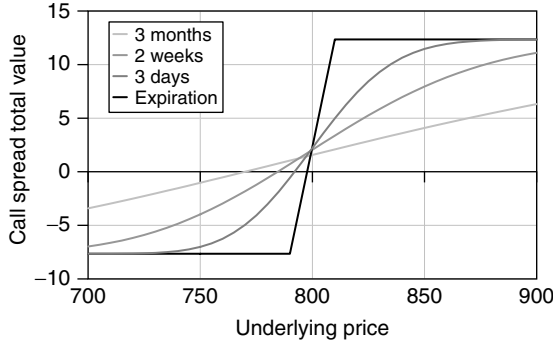


Figure 1 The value of a call spread at various times before expiration

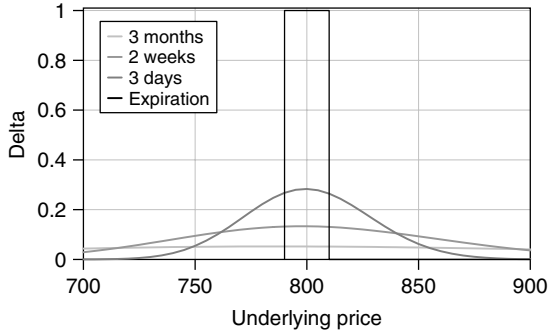


Figure 2 The delta of a call spread at various times before expiration

probability distribution function of the underlying at termination can be inferred from market prices as the derivative of digital options prices with respect to the strike. As these digital option prices come from call spreads with close strikes, we can conclude that the probability distribution function can be inferred from vanilla option prices.

Equation (2) shows that for close strikes or long expiries, the value of a call spread is approximately the strike difference times the probability that the underlying finishes above the spread:

$$CallSpread \approx e^{-r\tau} (K_2 - K_1) \cdot (1 - \Phi(K_2)) \quad (4)$$

This can be used as a crude first-order estimate for the value of a call spread. The second term in equation (2) can be approximated as (similar to the

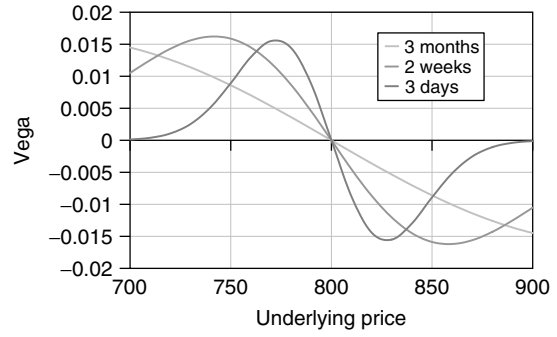


Figure 3 The vega of a call spread at various times before expiration

area of a triangle)

$$e^{-r\tau} \int_{K_1}^{K_2} (S - K_1) p(S) dS \approx e^{-r\tau} \frac{(K_2 - K_1)}{2} \times (\Phi(K_2) - \Phi(K_1)) \quad (5)$$

This gives a better approximation

$$CallSpread \approx e^{-r\tau} (K_2 - K_1) \times \left(1 - \frac{\Phi(K_1) + \Phi(K_2)}{2} \right) \quad (6)$$

This is a very intuitive formula as it is just the payoff of the call spread times the “average” probability the call spread finishes in the money.

References

- [1] Hull, J. (2003). *Options, Futures, and Other Derivatives*, 5th Edition, Prentice Hall.
- [2] Lehman Brothers (2008). *Listed Binary Options*, available at <http://www.cboe.com/Institutional/pdf/ListedBinaryOptions.pdf>
- [3] The Options Industry Council (2007). *Option Strategies in a Bull Market*, available at www.888options.com.
- [4] The Options Industry Council (2007). *Option Strategies in a Bear Market*, available at www.888options.com.
- [5] The Options Industry Council (2007). *The Equity Options Strategy Guide*, January 2007, available at www.888options.com
- [6] Reiner, E. & Rubinstein, M. (1991). Breaking down the barriers, *Risk Magazine* 4, 28–35.
- [7] Taleb, N.N. (1997). *Dynamic Hedging: Managing Vanilla and Exotic Options*, Wiley Finance.

- [8] Wikipedia (undated). *Binary Option*, available at http://en.wikipedia.org/wiki/Binary_option

Related Articles

Call Options.

Further Reading

Haug, E.G. (2007). *Option Pricing Formulas*, 2nd Edition, McGraw Hill.

ERIC LIVERANCE

Butterfly

A butterfly spread is an option strategy with limited upside and limited downside that uses call options of three different strikes but the same maturity on the same underlying. Specifically, a butterfly is a structure that is a long position in 1 low-strike call, a short position in 2 midstrike calls, and a long position in 1 high-strike call. More details and pricing models can be found in [2]. Market considerations can be found in [4–6]. The butterfly spread produces a structure that at maturity pays off only in scenarios where the price of the underlying is between the lowest and highest strikes. One can think of this strategy as buying an option on the underlying being in a range. The butterfly has limited upside potential, but a significantly reduced cost compared to that of an outright call option. It should be used for expressing a bullish view that the underlying will trade in a range. As with all options, choosing the strike and maturity will depend on one's view of how much the underlying will move and how quickly it will move there. An example is shown in detail in Figure 1.

In the example shown in Figure 1, we look at a 780/800/820 butterfly on the S&P 500 index, SPX. With the underlying SPX index at 770 and with three months to expiration, the butterfly cost is close to 1.00. The call option with the 800 strike is 70.14; thus, the cost for a butterfly is significantly reduced from the outright cost of a call option with the same strike. This upfront cost for the butterfly is the maximum that this butterfly position can lose. We subtract this initial investment from all other valuations, as shown in Figure 1, to get a total value. If the underlying is exactly 800 at expiration, the position will earn 20 on the butterfly from the low-strike option. The maximum position profit then is the strike spread minus the initial cost, or $20 - 1.00 = 19.00$.

Note that with three months to expiration, the butterfly value is fairly insensitive to the underlying price and is difficult to distinguish from the x -axis. However, as the option gets closer to expiration, the sensitivity of the price of the call spread becomes greater, especially in the range of the butterfly strikes. This sensitivity to underlying price or delta (*see Delta Hedging*) is illustrated in Figure 2.

The delta of the butterfly stays relatively flat until relatively close to expiration of the option. When the butterfly is close to expiration, the delta is very unstable around the three strikes.

Another important consideration is the volatility implied by the market (*see Implied Volatility: Market Models*). The vega profile of a butterfly is shown in Figure 3. When the underlying is close to the strikes, the vega is negative because when volatility increases the probability that the underlying expires out of the money increases. For this reason, it is common to use a butterfly with relatively long expiries and with strikes centered around at-the-money to take a view that implied volatility will decline while still holding a position with relatively small delta (insensitive to changes in the underlying). When the underlying is away from the money, the butterfly is long vega because when volatility increases, the probability that the underlying finishes in the money increases.

The Relationship with Distribution of the Underlying

A butterfly can be thought of as a long call spread plus a short call spread, with overlapping strikes and the same strike spread. An approximation for the value of a call spread can be found in **Call Spread**:

$$\text{Call Spread} \approx e^{-r\tau} (K_2 - K_1) \cdot \left(1 - \frac{\Phi(K_1) + \Phi(K_2)}{2} \right) \quad (1)$$

where $\Phi(x)$ is the cumulative distribution function of the underlying. Applying equation (1) to a butterfly, we have

$$\begin{aligned} \text{Butterfly} &\approx e^{-r\tau} (K_2 - K_1)^2 \left[\frac{\Phi(K_3) - \Phi(K_1)}{K_3 - K_1} \right] \\ &\approx e^{-r\tau} (K_2 - K_1)^2 p(K_2) \end{aligned} \quad (2)$$

where $p(x)$ is the probability distribution function of the underlying at option expiration and is the derivative of $\Phi(x)$ (Figure 4).

We can apply this formula in the following way. We convert the triangle in the lower part of Figure 3 to a square. Then we let the value of the payoff of the

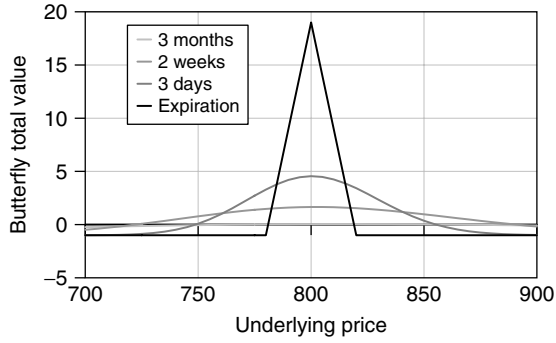


Figure 1 The value of a butterfly at various times before expiration

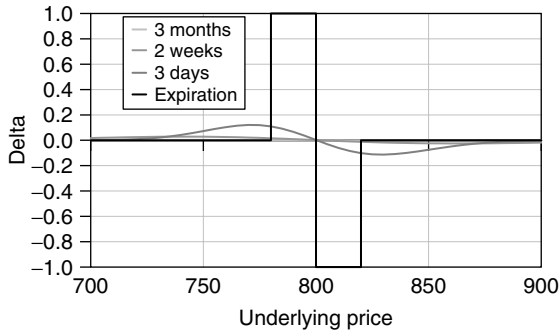


Figure 2 The delta of a butterfly at various times before expiration

butterfly to be represented as the probability times the area of the square, as in equation (2). Then turning around equation (2), we have

$$\text{prob}(K_a < S < K_b) \approx e^{r\tau} \frac{\text{Butterfly}}{(K_2 - K_1)} \quad (3)$$

The relationship between option prices and the distribution of the underlying was first pointed out in [1], but see also [3, 7]. The use of call spreads and butterflies to impute the market-implied underlying probability distribution can be related to taking derivatives with respect to the strike of the call price. A call spread is like a first derivative and a butterfly is like a second derivative. Formally, we have

$$\text{Call} = e^{-r\tau} \int_K^\infty (S - K) p(S) dS \quad (4)$$

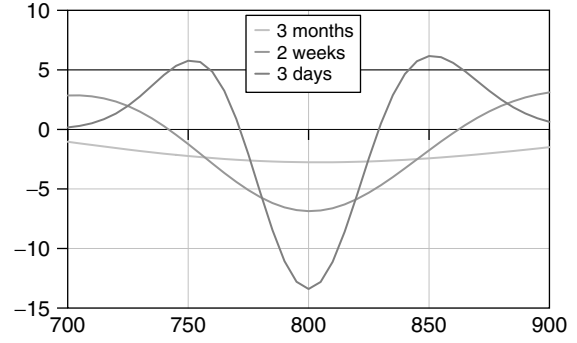


Figure 3 The vega of a butterfly at various times before expiration

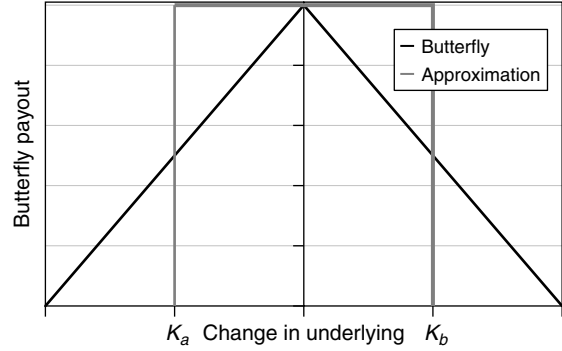


Figure 4 Using a butterfly to infer the underlying probability distribution

$$\frac{\partial}{\partial K} \text{Call} = -e^{-r\tau} \int_K^\infty p(S) dS \quad (5)$$

$$\frac{\partial^2}{\partial K^2} \text{Call} = -e^{-r\tau} p(K) \quad (6)$$

References

- [1] Breeden, D. & Litzenberger, R. (1978). Prices of state-contingent claims implicit in option prices, *Journal of Business* **51**, 621–651.
- [2] Hull, J. (2003). *Options, Futures, and Other Derivatives*, 5th Edition, Prentice Hall.
- [3] Jackwerth, J.C. (1999). Option-implied risk-neutral distributions and implied binomial trees: a literature review, *The Journal of Derivatives* **7**, 66–82.

-
- [4] The Options Industry Council (2007). *Option Strategies in a Bull Market*, available at www.888options.com.
 - [5] The Options Industry Council (2007). *Option Strategies in a Bear Market*, available at www.888options.com
 - [6] The Options Industry Council (2007). *The Equity Options Strategy Guide*, January 2007, available at www.888options.com
 - [7] Rubenstein, M. (1994). Implied binomial trees, *The Journal of Finance* **49**, 771–818.

Related Articles

Corridor Options; Risk-neutral Pricing; Variance Swap.

ERIC LIVERANCE

Gamma Hedging

Why Hedging Gamma?

Gamma is defined as the second derivative of a derivative product with respect to the underlying price. To understand why gamma hedging is not just the issue of annihilating a second-order term in the Taylor expansion of a portfolio, we review the profit and loss (P&L)^a explanation of a delta-hedged self-financing portfolio for a monounderlying option and its link to the gamma.

Let us consider an economy described by the Black and Scholes framework, with a riskless interest rate r , a stock S with no repo or dividend whose volatility is σ , and an option O written on that stock.

Let Π be a self-financing portfolio composed at t of

- the option O_t ;
- its delta hedge: $-\Delta_t S_t$ with $\Delta_t = \frac{\partial O}{\partial S}$; and
- the corresponding financing cash amount $-O_t + \Delta_t S_t$.

We note $\delta\Pi$ the P&L of the portfolio between t and $t + \delta t$ and we set $\delta S = S_{t+\delta t} - S_t$. Directly, we have that the delta part of the portfolio P&L is $-\Delta_t \delta S$ and that the P&L of the financing part is $(-O_t + \Delta_t S_t)r\delta t$. Regarding the option P&L, δO , we have, by a second-order expansion,

$$\delta O \approx \frac{\partial O}{\partial t} \delta t + \frac{\partial O}{\partial S} \delta S + \frac{1}{2} \frac{\partial^2 O}{\partial^2 S} (\delta S)^2 \quad (1)$$

Furthermore, the option satisfies the Black and Scholes equation (see **Black–Scholes Formula**):

$$\frac{\partial O}{\partial t} + rS \frac{\partial O}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 O}{\partial^2 S} = rO \quad (2)$$

Combining these two equations and writing the P&L of the portfolio as the sum of the three terms, we get

$$\delta\Pi \approx \frac{1}{2} S^2 \frac{\partial^2 O}{\partial^2 S} \left(\left(\frac{\delta S}{S} \right)^2 - \sigma^2 \delta t \right) \quad (3)$$

where $\partial^2 O / \partial^2 S$ is the gamma of the option part of the portfolio (in terms of definition, $S^2(\partial^2 O / \partial^2 S)$ is called the *cash gamma* because it is expressed

in currency and can be summed over several stock positions, whereas the direct gamma cannot).

As no condition was put on the relation of the volatility to time and space, equation (3) is easily extended to a local volatility setting (see **Local Volatility Model**). Practitioners call this equation the *breakeven relation* and $\sigma\sqrt{\delta t}$ the *breakeven* for it represents the move in performance the stock has to make in the time δt to ensure a flat P&L (e.g., if we consider that a year is composed of 256 open days, a stock having an annualized volatility of 16% needs to make a move of 1%, at which the delta is rebalanced, to ensure a flat P&L between two consecutive days). Figure 1 shows the portfolio P&L for a position composed of an option with a positive gamma.

Equation (3) leads to two important remarks. First, it is a local relation, both in time and space, and the fact that the gamma is gearing the breakeven relation implies that the global P&L of a positive gamma position, hedged according to the Black and Scholes self-financing strategy, can very well be negative if a stock makes large moves in a region where the gamma is small and makes small moves in a region where the gamma is maximum, even if the realized variance of the stock is higher than the pricing variance σ^2 . Secondly, in the long run, the realized variance is usually smaller than the implied variance, which can lead practitioners to build negative gamma positions. Yet, Figure 1 shows that a positive gamma position is of finite loss and possibly infinite gain, whereas it is the opposite for a negative gamma position. Practically, this is why traders tend naturally to a gamma neutral position.

A specific aspect of the equity market is the presence of dividends. One can wonder if, on the date the stock drops by the dividend amount, a positive gamma position is easier to carry than a negative gamma position. It is, of course, linked to the dividend representation chosen in the stock modeling. It can be shown that the only consistent way of representing the dividends is the one proposed in **Dividend Modeling**, where the stock is modeled as in Black and Scholes between two consecutive dividend dates. It is the only representation in which equation (3) stands (on the dividend date, the P&L term coming from the cash dividend part is offset by a term arising from the adapted Black and Scholes equation). In others, either the gamma carries a dividend part (dividend yield models) that leads to a

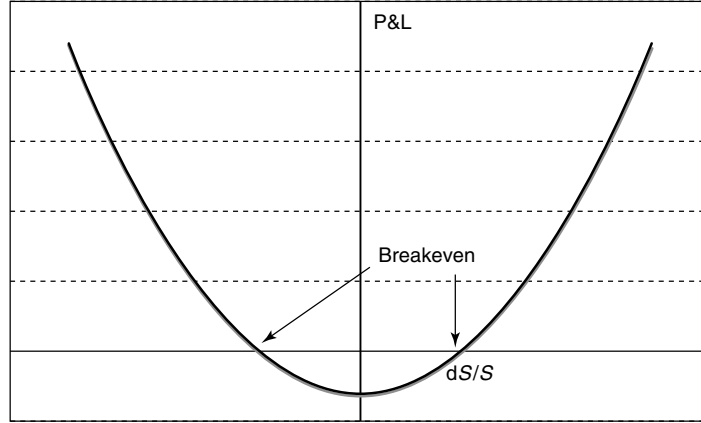


Figure 1 The P&L of a self-financing portfolio composed of an option with a positive gamma in the interval δt

false breakeven on the dividend date or equation (3) is not associated with the stock but with the variable that is stochastic (model in which the stock is described as a capitalized exponential martingale minus a capitalized dividend term, for example). This is why practitioners use the model proposed in **Dividend Modeling** rather than any other. This is also why it is, indeed, a general framework we put ourselves in by excluding dividends and repo (which is usually represented by a drift term whose P&L impact is also offset by a term arising from the adapted Black and Scholes equation) in our analysis.

Practical Gamma Hedging

We have seen why traders usually try to build a gamma-neutral portfolio. Yet, there is no pure gamma instrument in the market, and neutralizing the gamma exposure always brings a vega exposure to the portfolio. Without trying to be exhaustive, we briefly review here some natural gamma hedging instruments.

Hedging Gamma with Vanilla Options

European calls and puts have the same gamma (and the same vega). Hence, they are equivalent hedging instruments. Figure 2 shows the gamma of a European option for two different maturities and Figure 3 shows the compared evolutions with respect to the maturity of the gamma and of the vega.

These two figures show that, to efficiently hedge his or her gamma exposure, a trader would rather use a short-term option to avoid bringing too much vega to his or her position. Moreover, the gamma of an “at-the-money” option is increasing as one gets closer to the maturity, whereas the gamma of an “out-of-the-money” option is decreasing.

The Put Ratio Temptation

As equation (3) shows, the gamma and the theta (first derivative of a derivative product with respect to time) of a portfolio are of opposite signs. Moreover, in the equity market, the implied volatility is usually described by a skew, meaning that if we consider two puts P_1 and P_2 for the same maturity T , having two strikes K_1 and K_2 with $K_1 < K_2$, we classically have $\sigma_{K_1} > \sigma_{K_2}$. If we now build a self financing portfolio Π that is composed of $P_2 - \alpha P_1$ with $\alpha = \Gamma_2 / \Gamma_1$, the ratio of the two gammas, we get from equation (3) that $\delta \Pi \approx \frac{1}{2} S^2 \Gamma_2 (\sigma_{K_1}^2 - \sigma_{K_2}^2) \delta t > 0$.

This result is not in contradiction with arbitrage theory; it only demonstrates that equation (3) is strictly a local relation. As shown by Figure 2, to keep this relation through time, the trader would have to continuously sell the put P_2 , as α increases as time to maturity decreases, and, in case of a market drop down, he or she would find himself in a massive negative gamma situation. Still, practitioners commonly use put ratios to improve the breakeven of their position.

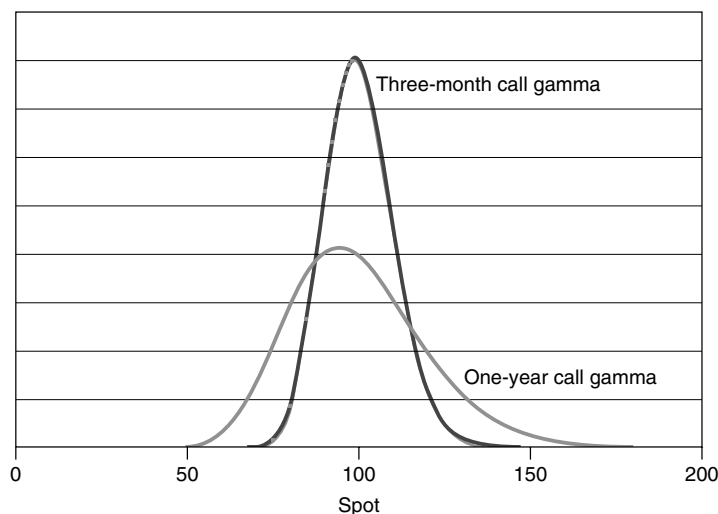


Figure 2 Gamma of a European call as a function of the spot for two maturities (strike is equal to 100)

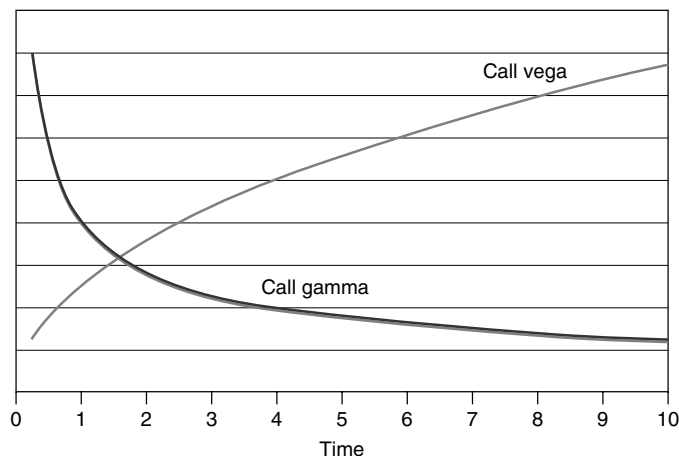


Figure 3 Compared evolution of the gamma and vega of an at-the-money European call as a function of maturity (scales are different)

Hedging Gamma with a Variance Swap or a Gamma Swap

As explained in **Variance Swap** a variance swap is equivalent to a log contract. Hence, its cash gamma is constant. It is therefore an efficient gamma hedging instrument for a portfolio whose gamma is not particularly localized (as opposed to a portfolio of vanilla options whose gamma is locally described by Figure 2). Gamma swaps (*see* **Gamma Swap**) have the same behavior. Their specificity is to have a constant gamma.

Extending the Definition of Gamma

In the market, the implied volatility changes with the spot moves (*see* **Implied Volatility Surface**). It is in contradiction with the use of a Black and Scholes model whose volatility is constant, but, to avoid the multiplicity of risk sources, and to keep them observable, traders tend to rely on that model, nonetheless (and therefore hedge their vega exposure). Nevertheless, to take this dynamics into consideration, some traders incorporate a “shadow” term into their sensitivities. The shadow gamma [1]

is defined as

$$\frac{\partial^2 O}{\partial^2 S} + \frac{\partial}{\partial S} \left(\frac{\partial O}{\partial \sigma} \frac{\partial \sigma}{\partial S} \right) \quad (4)$$

The second term, the shadow term, depends on the chosen dynamics of the implied volatility.

The problem with the shadow approach is that we cannot rely anymore on a self-financing strategy in the Black and Scholes framework to define the breakeven. One solution, in order to build a self-financing strategy that incorporates volatility surface into the dynamics, is to use a stochastic volatility model (see **Heston Model**) instead of a Black and Scholes model. For example, one can use the following model:

$$\begin{aligned} dS &= rS dt + \sigma S dW_t^1 \\ d\sigma &= \mu dt + \nu dW_t^2 \\ d\langle W^1, W^2 \rangle_t &= \rho dt \end{aligned} \quad (5)$$

Using the same arguments as in the Black and Scholes framework, the P&L of a delta-hedged self-financing portfolio (now with a first-order hedge for the volatility factor using a volatility instrument like a straddle, for example) in this model is

$$\begin{aligned} \delta \Pi &\approx \frac{1}{2} S^2 \frac{\partial^2 O}{\partial^2 S} \left(\left(\frac{\delta S}{S} \right)^2 - \sigma^2 \delta t \right) \\ &+ \frac{1}{2} \frac{\partial^2 O}{\partial^2 \sigma} ((\delta \sigma)^2 - \nu^2 \delta t) \\ &+ S \frac{\partial^2 O}{\partial S \partial \sigma} \left(\left(\frac{\delta S}{S} \right) \delta \sigma - \rho \sigma \nu \delta t \right) \end{aligned} \quad (6)$$

Two other “gamma” terms appear in this equation, which proves that incorporating the dynamics of the volatility is not as simple as the addition of a shadow term in the Black and Scholes breakeven relation. It also shows that controlling the P&L leads to a more complex gamma hedge, as it is now necessary to annihilate two more terms (the second and third ones, for which natural hedging instruments are strangles and risk reversals).

Another popular way of integrating the volatility surface dynamics in the model is to use Levy processes (see **Exponential Lévy Models**). We do not give the P&L explanation in that case, but, like in the stochastic volatility framework, it is the sum of the term presented in equation (3), for the Brownian

part, and of a term coming from the pure jump part. The hedge of the latter is very complex because it is not localized in space (one needs to use a strip of gap options, e.g., to control it).

Finally, a possible way of controlling the volatility surface dynamics is to make no assumption on the volatility except that it is bounded. This framework is known as *uncertain volatility modeling* and is presented in **Uncertain Volatility Model**. The analysis leads to the conclusion that instead of one breakeven volatility, there are, in fact, two: the upper bound for positive gamma regions and the lower bound for negative gamma ones. In that case, and supposing that the effective realized volatility stays locally between these two bounds, gamma hedging is not necessary, as the P&L of the delta-hedged self financing portfolio is naturally systematically positive.

Multiunderlying Derivatives

We consider a multidimensional Black and Scholes model of N stocks S_i with volatility σ_i . ρ_{ij} represents the correlation between the Brownian motions controlling the evolution of S_i and S_j . We do not discuss the issue of multicurrency (see **Quanto Options**) and, using the same mechanism as in the monounderlying framework, we can express the P&L of a delta-hedged self-financing portfolio as

$$\begin{aligned} \delta \Pi &\approx \frac{1}{2} \sum_{i=1}^N S_i^2 \frac{\partial^2 O}{\partial^2 S_i} \left(\left(\frac{\delta S_i}{S_i} \right)^2 - \sigma_i^2 \delta t \right) \\ &+ \sum_{i < j} S_i S_j \frac{\partial^2 O}{\partial S_i \partial S_j} \left(\left(\frac{\delta S_i}{S_i} \right) \left(\frac{\delta S_j}{S_j} \right) - \rho_{ij} \sigma_i \sigma_j \delta t \right) \end{aligned} \quad (7)$$

The first term can be controlled by the hedging instruments we have previously reviewed. The cross ones, which incorporate the “cross gammas”, can also be controlled, using so-called correlation swaps (typically, a basket option minus the sum of the individual options).

Conclusion

Controlling the gamma exposure of a position is one of the main concerns of traders. Hedging instruments

are common options but it is not possible to simply hedge the gamma without modifying the vega exposure of the position. Also, integrating the volatility surface dynamics in the model leads to a more complex gamma-hedging issue than in a Black and Scholes model, but it still can be addressed. Moreover, we remark that although we have considered the equity market in our study, this analysis can easily be extended to other complete markets in which there is no arbitrage and where the price process is modeled by a Brownian motion of any dimension.

End Notes

^aP&L stands for “profit and loss” and represents the evolution of the portfolio value between two dates due to time and to the market activity between these dates.

Reference

- [1] Taleb, N. (1996). *Dynamic Hedging: Managing Vanilla and Exotic Options*, John Wiley & Sons, pp. 138–146.

Related Articles

Correlation Swap; Delta Hedging; Exponential Lévy Models; Gamma Swap; Heston Model; Uncertain Volatility Model; Variance Swap.

CHARLES-HENRI ROUBINET

Delta Hedging

The delta of an asset or a portfolio broadly means net market exposure to an asset class. This may be for a single asset, like an option, or for a portfolio, like an S&P 500 benchmarked mutual fund. Delta hedging is the process of reducing the size of this exposure to a target level to reduce the amount of risk exposure due to the delta present in the portfolio.

Delta hedging is a term used in two broad categories: delta hedging of investment portfolios and delta hedging of financial derivatives.

Delta Hedging of Investment Portfolios

The delta of a portfolio is determined based on the assets in the portfolio. For equity portfolios, it may mean net equity exposure, for fixed income portfolios it may mean portfolio duration, while for commodity portfolios it could be exposure to the underlying commodity such as bushels of corn or barrels of oil. To delta hedge a portfolio, the portfolio manager determines what target net delta level is desired for the portfolio and uses a broad market instrument like a futures or swaps to achieve that desired level of delta. In addition, some portfolio managers might be more precise by hedging sensitivity to one factor (beta hedging) or multiple factors.

Delta Hedging of Financial Derivatives

Financial derivatives provide linear or nonlinear exposure to an underlying asset price level. While it is possible that both the buyer and the seller of a specific derivative are interested in identically opposite exposures, it is more common that one of the parties to the transaction is a financial intermediary or market maker that plans to hedge or reduce some of the risks of the transaction.

Delta hedging is the simplest form of hedging financial derivatives. This hedging aims to neutralize the direct exposure to the underlying asset, or delta, while maintaining second-order exposures to convexity, volatility, and time. In this discussion, we assume that a market maker enters into a financial derivative transaction and decides to hedge the delta

arising from the transaction. If the net delta exposure between the financial derivative and the hedge is zero, the position is said to be *delta neutral*.

The Process of Delta Hedging

The process of delta hedging incorporates some or all of the following steps.

Calculation of Delta

Typically, this is based on risk neutral valuation of the product. As a result the delta may vary depending on the underlying model used in the valuation. This variation may be small for simple products but may result in material differences for complex products. For example, a vanilla equity call option at-the-money (ATM) might have a similar delta based on Black–Scholes [1] inputs or a local volatility model, while a knock-out put, might have materially different deltas if measured by the two models above.

Determining the Appropriate Hedging Instrument

Entering into a hedge has various costs, the key aspects of which are mentioned below:

1. **Liquidity of the hedging instrument**

It is ideal that the hedging instrument is easily tradable with a low bid/ask spread. For derivatives, where the underlying is less liquid or there might be some market impact in executing the hedge, the price of the financial derivative is based on the average realized execution price level of the hedge.

2. **Basis risk in the hedging instrument**

In practice, some of the hedging instruments may have a small basis risk *versus* the actual underlying, which needs to be hedged. This usually happens when the actual underlying has significant transaction or liquidity costs and the basis tracking risk is low.

3. **Financing costs**

There might be costs associated with going long or short an asset as a hedge or entering into a financial swap transaction (e.g., the cost of posted collateral).

4. **Stability of the hedge**

If the hedge involves going long an asset or entering into a long-term linear over the counter (OTC)

derivative, the hedge is considered stable. However, if the market maker requires to borrow the hedge to go short (e.g., short stocks or bonds), then the hedge may be subject to a lack of availability to borrow in the marketplace.

Incorporating the Cost of Hedging into the Price of the Financial Derivative

1. Market makers who seek to hedge a position may incorporate the cost of hedging using adjustments to the financing rates, expected loss owing to basis risk, or future rollover costs of short-term hedges (like futures) into the pricing of the derivative.
2. In addition, the market maker may seek contractual obligations to ensure that if he is unable to hedge the contract, he has a right to unwind the derivative at fair market price.

Examples of Delta Hedging

Below are some examples of delta hedging.

Example 1. Listed put option on S&P 500 index.

Market Maker Derivatives Inc. (MMDI) sells an exchange listed put on the S&P 500 (SPX) Index on a \$100mm notional with a strike price of \$1350 and one year expiration to a client when the SPX is trading at \$1400. The Black–Scholes model calculates the delta of the position at 38.5%. To delta hedge and neutralize the position to P&L swings from changes in the level of the SPX, MMDI needs to sell \$38.5mm of SPX Index. MMDI can do one of the following to achieve this:

1. Sell \$38.5mm in SPX index futures (1100 futures at \$1400 with a 250 multiplier). Since these futures expire every three months, the market maker needs to roll this exposure into a new contract every three months.
2. Sell \$38.5mm of a one-year swap on the SPX Index with another counterparty for one year or enter into an OTC put/call combo transaction (buying a put and selling a call with the same strike price to replicate a forward using put–call parity) for \$38.5 mm notional. This is done in the OTC market place and appropriate International Swaps and Derivatives Association, Inc. (ISDA) (<http://www.isda.org/>) documents and collateral agreements need to be in place before this is done.

3. Sell \$38.5mm a basket of stock that comprises S&P 500 index, paying a borrowing fee and executing stock orders on the exchange.

Example 2. Long total return swap on XYZ commodity index.

MMDI sold a five-year total return swap on the XYZ commodity index to a client for \$10mm notional, which has a delta of \$10mm. To hedge the position, MMDI can do one of the following:

1. Buy an equal and offsetting swap with another client or market counterparty or a basket of commodities swaps for each of the components of the XYZ index from another counterparty which is the perfect hedge (assuming no counterparty risk).
2. Hedge the XYZ index with a much more actively traded index like the CRB index taking into account the tracking risk and weighting differences of the components between the two indices.
3. Maintain a portfolio of long futures on commodities, which compromises the XYZ index and rolling futures on the position for the next five years. The market maker assumes the risk of rolling the futures positions to changes in the shape of the commodity forward curve.

Rehedging Delta with Time, Spot Moves

The delta of a financial product may change with time or the levels of the different market parameters such as volatility, underlying price, interest rates, or skew. It is then necessary to periodically adjust the size of the hedge to maintain the delta at a preset level. The rate of change of the delta is proportional to the gamma (*see Gamma Hedging*).

Reference

- [1] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* 81(May-June), 637–659.

Related Articles

Hedging; Hedging of Interest Rate Derivatives; Option Pricing: General Principles.

VIJU JOSEPH

Dispersion Trading

Dispersion trading refers to the practice of selling index variance while buying variance of its constituents at the same time. The reverse strategy (buying index variance while selling constituents variance) can also be employed, but it is not as popular.

To understand dispersion trading, consider the index as a basket of stocks:

$$S_I = \sum_{i=1}^n w_i S_i \quad (1)$$

where w_i is the weight of S_i in the basket.

The variance of the index is related to that of the individual stocks by

$$\sigma_I^2 = \sum_{i=1}^n w_i^2 \sigma_i^2 + 2 \sum_{i=1}^n \sum_{j>i}^n \rho_{ij} w_i w_j \sigma_i \sigma_j \quad (2)$$

The variance is defined as

$$\sigma_i^2 = \frac{1}{T} \sum_{t=1}^T (S_i^t - \bar{S}_i)^2 \quad (3)$$

with $\bar{S}_i = \frac{1}{T} \sum_{t=1}^T S_i^t$, and the correlation ρ_{ij} is defined as

$$\rho_{ij} = \frac{1}{T} \sum_{t=1}^T (S_i^t - \bar{S}_i)(S_j^t - \bar{S}_j) / (\sigma_i \sigma_j) \quad (4)$$

If we hold the realized variances of every component stock constant, the maximum for the index variance is reached when the correlation between all the components is 100%. If the correlation between stocks is not perfectly 100%, the index variance is lower. The more “dispersed” the stocks are, the lower is the index variance.

A measure of dispersion—the dispersion spread—can be defined as

$$D = \sqrt{\left(\sum_{i=1}^n w_i \sigma_i \right)^2 - \sigma_I^2} \quad (5)$$

or, alternatively, it has also been defined as

$$D = \sum_{i=1}^n w_i \sigma_i - \sigma_I \quad (6)$$

$D = 0$ corresponds to the case when there is no dispersion—all correlations are 100%.

So, to long dispersion is equivalent to be short correlation and *vice versa*.

To characterize the correlation between the constituents, one can define the average correlation as if the correlation is the same between every pair of stocks in the basket

$$\bar{\rho} = \frac{\sigma_I^2 - \sum_{i=1}^n w_i^2 \sigma_i^2}{2 \sum_{i=1}^n \sum_{j>i}^n w_i w_j \sigma_i \sigma_j} \quad (7)$$

However, sometimes, it is easier to calculate the less accurate “correlation proxy,” which is defined as

$$\tilde{\rho} = \frac{\sigma_I^2}{\left(\sum_{i=1}^n w_i \sigma_i \right)^2} \quad (8)$$

This correlation proxy can be interpreted as the “average” of all correlations between all pairs of stocks in the index including a stock with itself (which we know should be 100%). When the number of stocks in the index n is high, it can be seen that $\sum_{i=1}^n w_i^2 \sigma_i^2$ is much less compared to the retained terms:

$$\sum_{i=1}^n w_i^2 \sigma_i^2 \ll \sigma_I^2 \quad (9)$$

$$\sigma_I^2 \simeq \sigma_I^2 - \sum_{i=1}^n w_i^2 \sigma_i^2 \quad (10)$$

$$2 \sum_{i=1}^n \sum_{j>i}^n w_i w_j \sigma_i \sigma_j \simeq 2 \sum_{i=1}^n \sum_{j>i}^n w_i w_j \sigma_i \sigma_j + \sum_{i=1}^n w_i^2 \sigma_i^2 \quad (11)$$

The correlation proxy and correlation are very close to each other. The implied correlation can be simply inferred from the ratio of average volatilities. Sometimes it is also convenient to calculate the mean variance ratio that is more directly related to the trade profit/loss (P/L).

By definition, realized correlation is the correlation calculated using realized volatilities, and implied correlation is the correlation calculated using implied volatilities. Implied volatilities decide the price of the traded instruments like vanilla options and variance swaps.

The success of dispersion trades relies on the fact that statistically the realized correlation tends to be below the implied correlation. Historically, if one were long dispersion, on average, one made more money than the amount one lost. There are many different reasons for this phenomenon, for example, one may argue that there is more market demand for index volatility than that of the individual stock, which means usually there is more premium for equity stock volatility. More importantly, correlation jumps to a very high level when extreme market conditions exist, namely, global recession and market crash, while it stays low in a normal and uneventful market.

To long the volatility of each component stock, and short the index volatility, one can either trade vanilla options or variance swaps. The variance swaps provide direct exposure to variance without the unnecessary cost and hassle of hedging against daily stock movements.

One issue in dispersion trade is to decide the relative weight for index and constituents variances. There is no single “correct” relative weight to use. For example, vega neutral weights aim to make the sum of constituents vega and index vega zero, so that the trade is hedged against fluctuations in level of volatility. “Premium neutral” weights make the initial premium of buying constituents and selling index cancel each other.

In reality, it is impractical to trade all constituents. Often, a selection of names in the index (or even those not in the index) is used. This is called a *proxy basket*. One can build the proxy basket by selecting, for example, the names that have the largest weights in the index, or the names that are judged relatively “cheap”, or the names that are mostly likely to “disperse” against each other, or simply by the stock fundamentals.

Related Articles

Basket Options; Correlation Swap.

YONG REN

Correlation Swap

A correlation swap is a type of exotic derivative security that pays off the observed statistical correlation between the returns of several underlying assets, against a preagreed price. At the time of writing, it is traded over-the-counter (OTC) on equity and foreign exchange derivatives markets. This article focuses on equity correlation swaps, which appeared in the early 2000s, as a means to hedge the parametric risk exposure of exotic trading desks to changes in correlation.

Payoff

Similar to **variance swaps**, the correlation swap payoff involves a notional (the amount to be paid/received per *correlation point*^a), a realized correlation component (the formula used to calculate the level of observed statistical correlation between the underlying assets), and a strike price:

$$\begin{aligned} \text{Correlation swap payoff} &= \text{notional} \\ &\times (\text{realized correlation} - \text{strike}) \end{aligned}$$

For example, a one-year correlation swap contract on the constituents of the Dow Jones Euro Stoxx 50 index would include the following terms:

- underlying assets—each of the 50 constituent stocks denoted by $S_1, \dots, S_N$ ($N = 50$);
- notional—€100 000 per *correlation point*;
- realized correlation— $\frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \rho_{i,j}$, where $\rho_{i,j} = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{Var}(X_i) \times \text{Var}(X_j)}} \times 100$ is the familiar pairwise coefficient of correlation between the time series X_i and X_j of daily log returns observed in the year following the trade date;
- strike—52.0 *correlation points*.

Thus, if after one year the arithmetic average of pairwise correlation coefficients between the 50 underlying assets is equal to 58.3 *correlation points*, the swap seller will pay a net cash flow of €630 000 to the swap buyer.

Realized Correlation

There are mainly two types of realized correlation formulas currently found on over-the-counter (OTC) markets:

- equally weighted realized correlation—the formula used in the above example;
- weighted realized correlation— $\frac{\sum_{1 \leq i < j \leq N} w_i w_j \rho_{i,j}}{\sum_{1 \leq i < j \leq N} w_i w_j}$,

where $w_1, \dots, w_N$ are preagreed positive weights summing to 1. In the above example, one would typically take the “index weights” as of the trade date, that is, the stock quantities that a portfolio manager would invest in to track one unit of the Dow Jones Euro Stoxx 50 index.

Several technical reports have investigated how the above weighted realized correlation (WRC) formula relates to other proxy formulas that are popular in econometrics, when the underlying assets and weights correspond to an equity index. Tierens–Anadu [4] give empirical evidence that in the case of the S&P 500 index,

$$WRC \approx \frac{\sum_{i < j} w_i w_j \text{Cov}(X_i, X_j)}{\sum_{i < j} w_i w_j \sqrt{\text{Var}(X_i) \text{Var}(X_j)}} \quad (1)$$

In addition, Bossu [1, 2] derives the following limit-case proxy formula, subject to some conditions on the weights:

$$\begin{aligned} &\frac{\sum_{i < j} w_i w_j \text{Cov}(X_i, X_j)}{\sum_{i < j} w_i w_j \sqrt{\text{Var}(X_i) \text{Var}(X_j)}} \\ &\stackrel{N \rightarrow \infty}{\sim} \frac{\text{Var}\left(\sum_{i=1}^N w_i X_i\right)}{\left(\sum_{i=1}^N w_i \sqrt{\text{Var}(X_i)}\right)^2} \quad (2) \end{aligned}$$

The proxy formula on the right-hand side is remarkable because we can interpret the numerator as “index variance” and the denominator as “average constituent variance”, which is more straightforward than the average of $N \times (N - 1)/2$ pairwise correlation coefficients.

Fair Value

At the time of writing, little is known about the “fair value” of correlation swaps. Owing to the typically large number of underlyings, the popular Monte Carlo engine with or without local volatility surfaces requires an $N \times N$ correlation matrix as additional input parameter. There are two problems with this approach: a practical one and a theoretical one. The practical problem is that individual correlation coefficients cannot be implied from listed option markets.^b The theoretical problem is that, even if one could come up with a sensible implied correlation matrix, a meaningful dynamic replication strategy for the correlation swap payoff would still be missing.^c

Ongoing research (see, e.g., the working papers of Bossu [1] and Jacquier [3]) aims to identify the linkages between **dispersion trading** and the dynamic hedging of correlation swaps, especially when the underlying assets are the constituent stocks of an equity index. This approach exploits the proxy formulas above to rewrite the correlation swap payoff as a function of tradable variance swaps; in this framework the correlation swap becomes a multiasset volatility derivative (see **Realized Volatility Options**), rather than a classical multiasset derivative.

End Notes

^aIn market jargon, a *correlation point* is equal to 0.01. With this convention, the value of a correlation coefficient is comprised between -100 and $+100$ correlation points.

^bTo imply the value for $\rho_{i,j}$, one needs three option prices: a vanilla option on S_i , a vanilla option on S_j and, for

example, a vanilla option on a portfolio made of 50% S_i and 50% S_j . The former two options are listed, the latter is not.

^cFor example, in a two-asset extension of the Black–Scholes model with instantaneous correlation ($d\ln S_i^1)(d\ln S_j^2) = \rho dt$, the forward value of an equally weighted correlation swap is simply $\rho - \text{strike}$, which would be hedged purely with cash!

References

- [1] Bossu, S. (2007). *A New Approach For Modelling and Pricing Correlation Swaps*, Dresdner Kleinwort Equity Derivatives report (working paper). Available at <http://math.uchicago.edu/~sbossu/CorrelationSwaps7.pdf>
- [2] Bossu, S. & Gu, Y. (2004). *Fundamental Relationship Between an Index's Volatility and the Average Volatility and Correlation of its Components*, JPMorgan Equity Derivatives report (working paper). Available at <http://math.uchicago.edu/~sbossu/CorrelFundamentals.pdf>
- [3] Jacquier, A. (2007). *Variance Dispersion and Correlation Swaps*, Working paper. Available at SSRN: <http://ssrn.com/abstract=998924>
- [4] Tierens, I. & Anadu, M. (2004). *Does it Matter Which Methodology you use to Measure Average Correlation Across Stocks?* Goldman Sachs Equity Derivatives Strategy: Quantitative Insights, 13 April 2004.

Related Articles

Basket Options; Correlation Risk; Dispersion Trading; Variance Swap.

SÉBASTIEN BOSSU

Stock Pinning

Stock pinning, or simply *pinning*, is formally the occurrence of a closing stock print, on option expiration day, which exactly matches the denominated value of a strike price. As an example, let stock XYZ have strikes 30, 32.5, 35, and 40. If on Friday, May 16, 2008 at 4:00 pm EDT (USA), the third Friday of the month and thus an expiration day for listed

options, the closing print of XYZ is \$35, then stock XYZ is said *to pin*. Prices of \$34.27, \$31.60, or even \$32.48, are said not to have pinned. Figure 1 is a tick price graph of KO (Coca Cola Corporation) showing the last several days prior to a pinning expiration.

As a practical matter, it may be useful experimentally to consider pinning to have occurred if the stock expires within a certain interval of a strike price. There are several reasons for this looser definition. Empirically there may be several “closing prints”

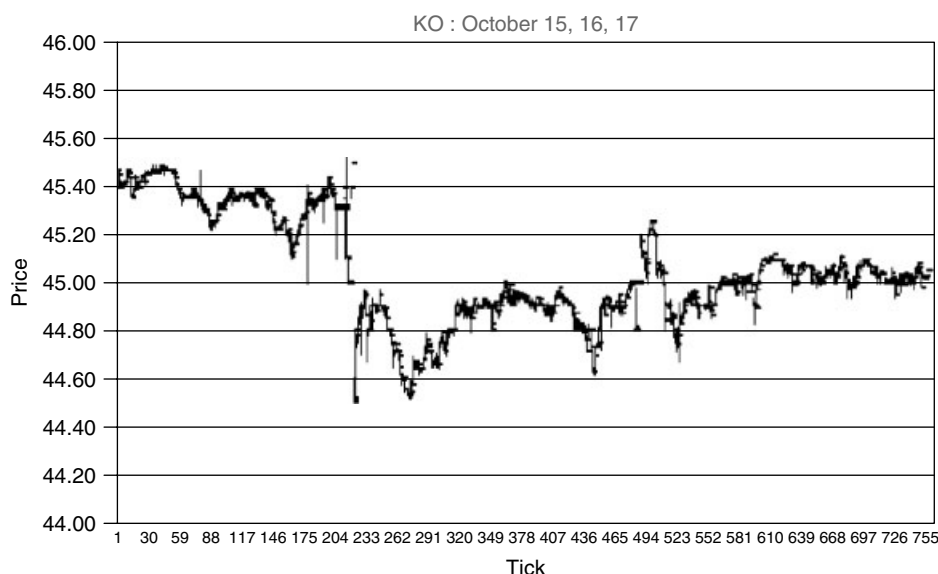


Figure 1 KO (Coca Cola) tick data for a pinning expiration, October 17, 2003

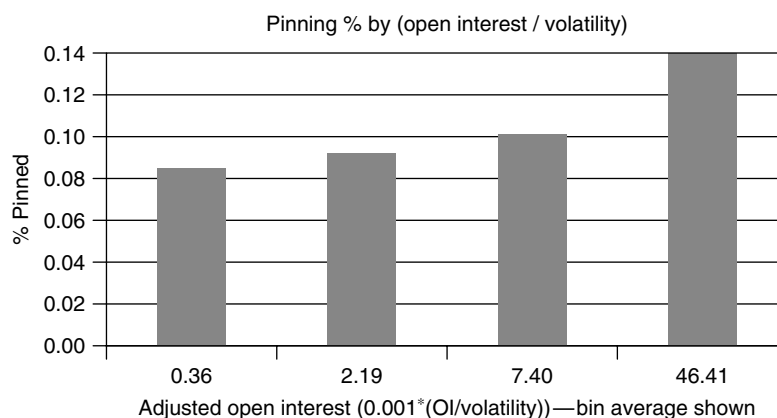


Figure 2 All optionable stocks in 2002 divided into quartiles by pinning strength, β . As predicted, the probability of pinning increases with beta. (Pinning criteria \$0.15, courtesy Bart Rothwell)

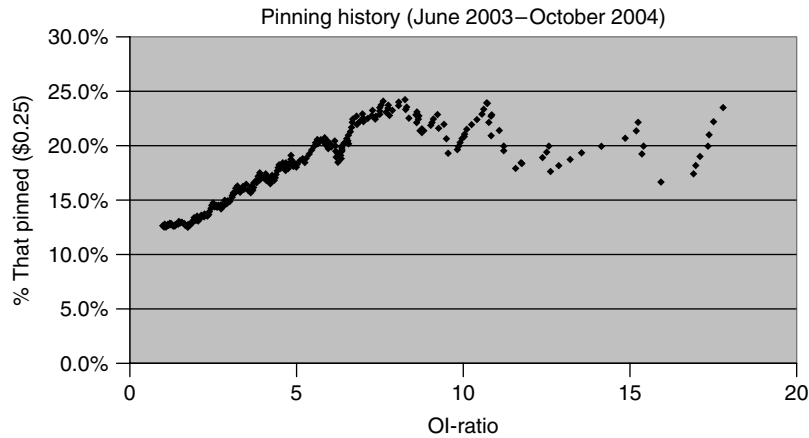


Figure 3 Cumulative distribution function of pinning of stocks, which are within \$1 of a strike with 1 week to go to expiration as a function of the parameter, β . (Courtesy Tom MacFarland)

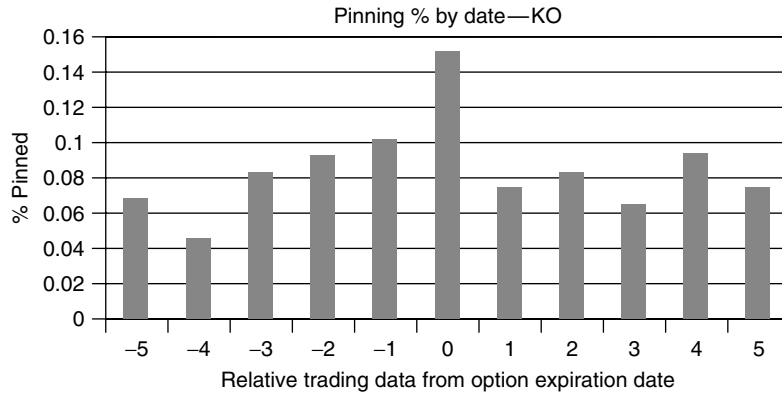


Figure 4 The percentage of days KO closed within \$0.15 of a strike in a 10-year period January 1, 1996 to January 1, 2005. 0 is expiration day, negative integers are days prior to expiration, positive integers are days following expiration. (Courtesy Bart Rothwell)

making a choice of the closing price arbitrary. Tick data shows that a stock may be effectively *pinned* over the last several minutes before expiration but then have a closing print just off the strike. In the first example, this might happen if the last quote were \$34.98 bid, at \$35.01, and the closing price stayed in the interval but was not precisely \$35.

Two additional reasons for a looser definition are the automatic exercise conditions mandated by the OCC (Options Clearing Corporation) and the consequent *pin risk*, which attends expiring short positions on the (nearly) pinned strike. The OCC has traditionally fixed an interval about a strike

outside of which in-the-money puts and calls would be automatically exercised by the clearing process; options within the interval would require *exercise notice* by the holder. Over time, the OCC has reduced the interval to the current \$0.01 (from \$0.05 before June 2008 expiration); traders will declare a stock to have pinned if it falls within the OCC interval. Pin risk attends to any short position inside this interval because an uncertain number of options may be assigned and thus an uncertain postexpiration stock position exists in the positions of those short the expiring at-the-money options.^a

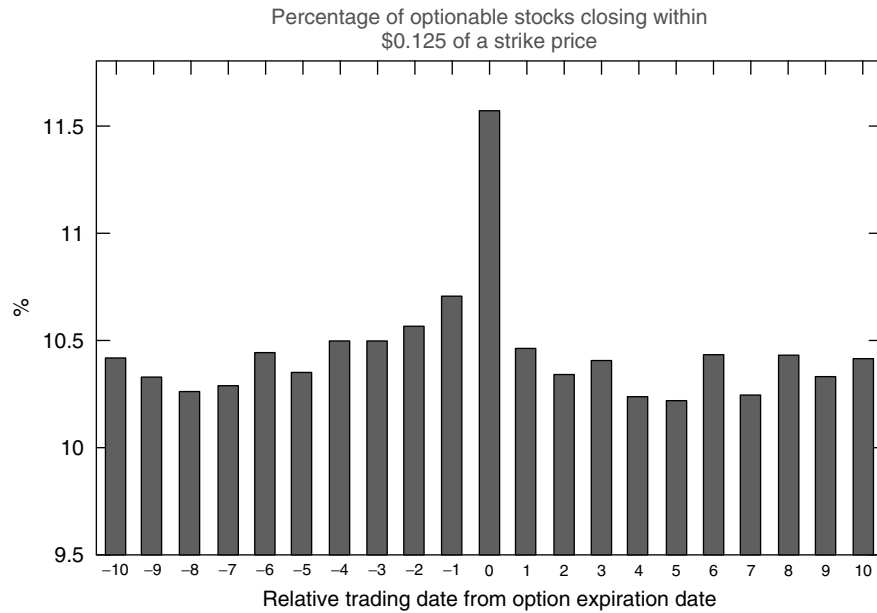


Figure 5 All stocks, January 1996 to September 2002 [Reproduced with permission from Stock price clustering on option expiration dates, Ni *et al.*, Journal of Financial Econometrics, © Elsevier 2005.]

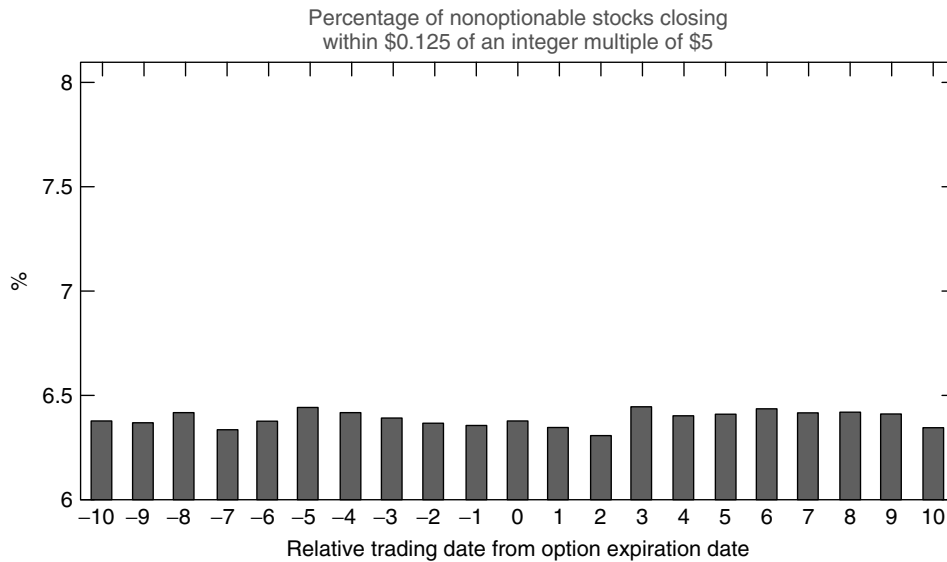


Figure 6 Nonoptionable stocks do not pin, January 1996 to September 2002 [Reproduced with permission from Stock price clustering on option expiration dates, Ni *et al.*, Journal of Financial Econometrics, © Elsevier 2005.]

So far we have defined a single instance of pinning. Complementing the notion of an individual instance of stock pinning is an ensemble assertion

of stock pinning. In this perspective, a stock, or stocks, is said to pin if, no matter how small an interval one chooses about a strike price, there

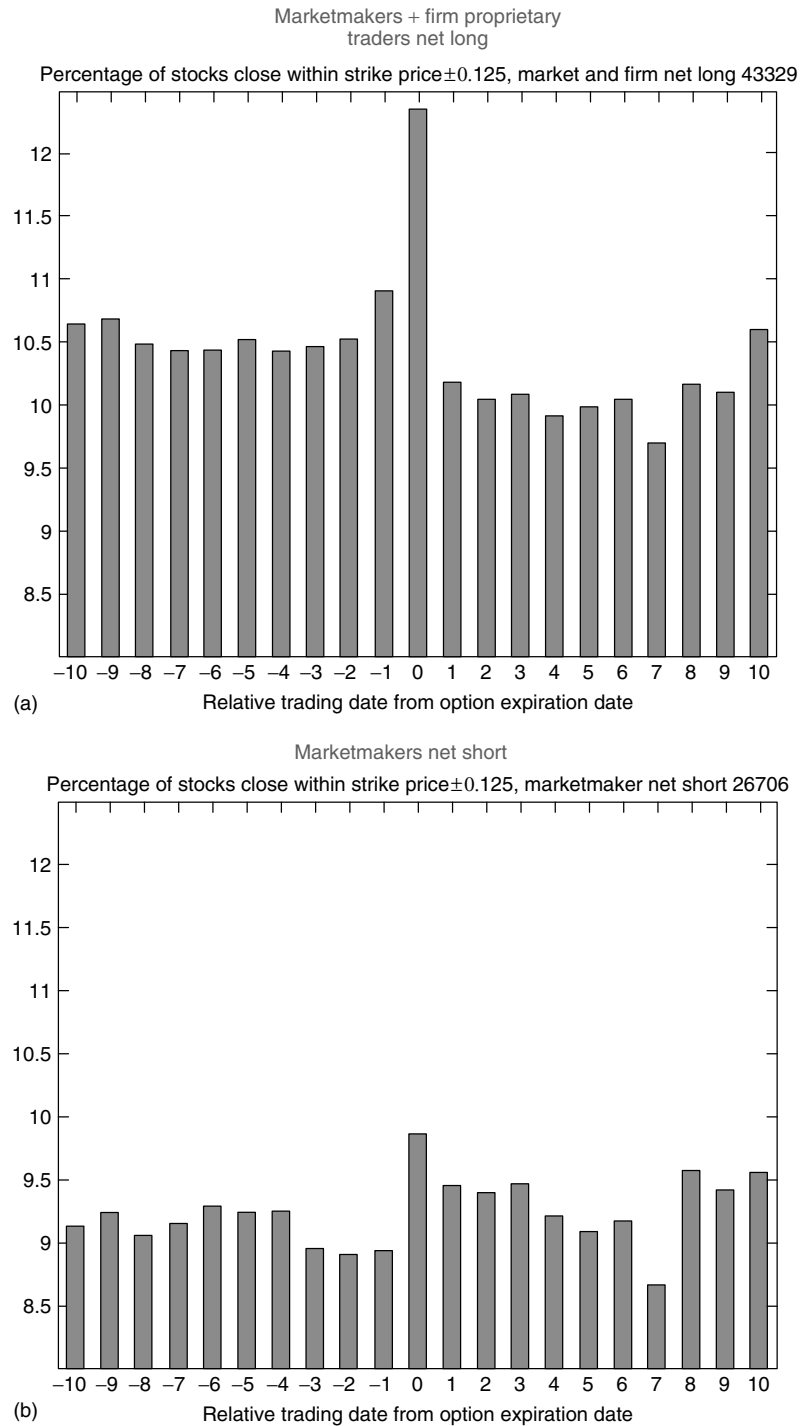


Figure 7 All optionable stocks, January 1996 to September 2002. (Pinning criterion \$0.125): pinning when professional traders are (a) long and (b) short the expiring at-the-money strike

is a finite probability of finding closing prints within the interval. To compute this limit, expressed mathematically as

$$\lim_{\varepsilon \rightarrow 0} P(|S-K| \leq \varepsilon) > 0 \quad (1)$$

where ε is an interval about (any) strike K , S is the stock price at expiry, and P is the probability among all expiration closes, one can do empirical experiments or theoretical calculations. It is important to note that standard models of option pricing such as Black–Scholes, Heston, SABR, and stochastic volatility models, in general, cannot exhibit pinning mathematically.

Although traders had long believed pinning to be a real phenomenon, little theoretical or experimental effort was made to examine the subject through the 1990s. Krishnan and Nelkin [3] looked at the data set of MSFT (Microsoft Corporation) expirations and found evidence of pinning. They proposed a model that combined a Gaussian random walk with a Brownian bridge process in order to force pinning to a strike. The model *perforce* guaranteed pinning, but suffered from many obvious weaknesses: stocks do not always pin; they can pin at many possible strikes; and the choice of the amount of Brownian bridge component was exogenously (and arbitrarily) imposed.

Then Avellaneda and Lipkin [1], and nearly simultaneously, Ni *et al.* [4], produced theoretical and experimental arguments for pinning. In the former work, Avellaneda and Lipkin proposed an asymmetric hedging strategy for professional traders—aggressive hedging of long gamma positions and weak hedging of short gamma positions. This hedging strategy coupled with a stock impact function (simplistically assumed to be linear^b) led directly to pinning (with nonzero probability), which depended naturally and endogenously on the option open interest, the intrinsic stock volatility, the (logarithmic) distance to the strike and the time to expiration. In dimensionless form, the *strength parameter*, β , is proportional to the open interest and inversely proportional to the volatility. Figures 2 and 3 show, experimentally, the monotonic growth of pinning probability with β .

Ni *et al.* used the IVY® and CBOE databases to check pinning frequencies. Figures 4 and 5, typical Ni *et al.* graphs, indicate the excess clustering

of stock prices near a strike on expiration days for KO and for the entire market over extended periods. Figure 6 demonstrates the absence of pinning in nonoptionable stocks. Finally, Ni *et al.* lent support for the hedging assumptions of Avellaneda and Lipkin. Figure 7(a,b) shows the difference between pinning when professional traders are long (a) and short (b) the expiring at-the-money strike.

Since 2004, other research groups, for example, Jeannin, *et al.* [2], have continued to explore the details of pinning.

End Notes

^a. Practitioners define *pin risk* as the uncertain deltas which an otherwise balanced position might have postexpiration due to the assignment of calls or puts on a near pinning strike. For example, a position long 50 calls and short 50 puts on the \$25 strike, for a stock which expires near \$25, may be assigned from 0 to 5000 shares of stock due to the uncertain number of puts which may have been exercised. This amount of stock thus assigned is independent of the number of calls the trader chooses to exercise himself.

^b. Following the initial 2003 work, Gennady Kasyan (with Avellaneda and Lipkin, unpublished) showed that any impact function stronger than square-root would result in pinning. This suggests that weaker impact functions may be contradicted by the extensive market evidence of stock pinning.

References

- [1] Avellaneda, M. & Lipkin, M.D. (2003). A market-induced mechanism for stock pinning, *Quantitative Finance* **3**, 417–425.
- [2] Jeannin, M., Iori, G. & Samuel, D. (2008). The pinning effect: theory and a simulated microstructure model, *Quantitative Finance* **8**, 823–831.
- [3] Krishnan, H. & Nelken, I. (2001). The effect of stock pinning upon option prices, *Risk* December.
- [4] Ni S, Pearson N., Poteshman A. (2004). Stock price clustering on option expiration dates, *SSRN* August 27.

Related Articles

Price Impact.

MIKE LIPKIN

Variance Swap

Definition

A variance swap is a volatility derivative that pays off on realized volatility of some underlying:

$$\text{Payoff} = (\sigma_R^2 - K_{\text{var}}) \times N \quad (1)$$

where σ_R^2 is the realized variance, K_{var} is the fair value of the realized variance at inception, and N is the notional amount, a leverage factor. The realized variance may be defined differently in different markets depending not only on the “default model” but also on specific contract specifications. One standard way for stocks [9] is to define realized variance as

$$\sigma_R^2 = \frac{252}{n} \sum_{i=1}^n u_i^2, \quad u_i = \ln(S_i / S_{i-1}) \quad (2)$$

This requires $n + 1$ observations of the daily closing stock price S_i . The factor 252 is the approximate number of business days in a year, which gives an annualized variance. Also note that this formula assumes the mean of the returns to be zero. This means that it is distinct from the statistical variance of the returns. The mean of returns is typically small and setting it to zero makes returns on variance swaps additive over time. Using log returns to measure variance makes this formula compatible with the standard Black–Scholes option pricing formula.

The long position in a variance swap receives N dollars for every point by which the stock’s realized variance σ_R^2 has exceeded the inception fair value K_{var} . See [6] for one of the earliest, but most comprehensive, references on variance swaps. In this reference, the authors discuss replication problems due to strike spacing and gapping of the underlying, for example.

It is market practice to define the variance notional in volatility terms:

$$\text{Variance Notional} = \frac{\text{VegaNotional}}{2 \times \sigma_{\text{strike}}} \quad (3)$$

where the strike volatility σ_{strike} is equal to the square root of the strike variance K_{var} . With this adjustment, if the realized volatility is 1 percentage point above

σ_{strike} at maturity, the payoff is approximately equal to the vega notional.

Uses of variance swaps include trading the level of volatility (variance swaps are a more pure way to do this compared to straddles), trading the realized/ implied vol spread, hedging of volatility exposures, or trading volatility on a forward basis via forward variance swaps. See [1] for more details about variance swaps, market practices, and their use in vol spread trading and correlation trading.

Fair Value and Skew

In [6], it is shown that a variance swap can be statically replicated with calls, puts, and a forward contract. The payoff for the variance swap then comes from delta hedging of this portfolio of options with the underlying. The fair value is determined by the cost of the replicating portfolio of options. The payoff in terms of realized volatility is achieved by delta hedging with the underlying. If the realized vol is exactly equal to the expected vol at inception, the hedging profits will be exactly equal to the cost of the option portfolio and the payout will be zero. As the portfolio of options (in theory) includes option of every strike, the fair value cost is affected significantly by the skew. The skew is defined to be the way the implied volatility changes as the strike changes, all else being equal.

The fair value can be derived using the volatility formula discussed in [4]. In this reference, the authors show the remarkable formula that the expected value of any smooth payoff function $f(S_T)$ in terms of the terminal stock price S_T can be written in terms of stock and option prices.

$$\begin{aligned} E_t[f(S_T)] &= f(\kappa)B_0 + f'(\kappa)[C_t(\kappa) - P_t(\kappa)] \\ &\quad + \int_{-\infty}^{\kappa} f''(s)P_t(s)ds \\ &\quad + \int_{\kappa}^{\infty} f''(s)C_t(s)ds \end{aligned} \quad (4)$$

An application of Ito’s lemma to the usual Black diffusion yields the variance differential:

$$2 \left[\frac{dS_t}{S_t} - d(\log S_t) \right] = \sigma^2 dt \quad (5)$$

Applying equation (4) to the log contract [11] in equation (5) shows that the replicating portfolio

consists of out-of-the-money calls and puts of all strikes K , where each option has the weight $1/K^2$ plus a dynamic delta hedge using a forward contract on the stock. The fair value of the variance strike is

$$K_{\text{var}} = \frac{2}{T} \left[rT - \left(\frac{S_0}{S_*} e^{rT} - 1 \right) - \log \frac{S_*}{S_0} \right. \\ \left. + e^{rT} \int_0^{S_*} \frac{1}{K^2} P(K) dK \right. \\ \left. + e^{rT} \int_{S_*}^{\infty} \frac{1}{K^2} C(K) dK \right] \quad (6)$$

Here, $C(K)$ and $P(K)$ are the prices of the call and put, respectively, with strike K .

In [6], we find the following formulas for fair value in terms of simple skew models. For the skew that is linear in the strike,

$$\sigma(K) = \sigma_{\text{ATM}} - b \frac{K - S_F}{S_F} \quad (7)$$

the fair value is

$$K_{\text{var}} = \sigma_{\text{ATM}}^2 (1 + 3Tb^2 + \dots) \quad (8)$$

For the skew that is linear in the delta (here Δ_p is the put delta),

$$\sigma(\Delta_p) = \sigma_{\text{ATM}} + b \left(\Delta_p + \frac{1}{2} \right) \quad (9)$$

the fair value is

$$K_{\text{var}} = \sigma_{\text{ATM}}^2 \left(1 + \frac{1}{\sqrt{\pi}} b \sqrt{T} + \frac{1}{12} \frac{b^2}{\sigma_{\text{ATM}}^2} + \dots \right) \quad (10)$$

In [7], we have a particularly elegant formula for the fair value given directly in terms of the skew. Denote

$$z(k) = d_2 = -\frac{k}{\sigma_{\text{BS}}(k)\sqrt{T}} + \frac{\sigma_{\text{BS}}(k)\sqrt{T}}{2} \quad (11)$$

Intuitively, z measures the log-moneyness of an option in implied standard deviations. Then

$$K_{\text{var}} = \int_{-\infty}^{\infty} dz \cdot N'(z) \sigma_{\text{BS}}^2(z) T \quad (12)$$

Market Risk

In [5], the authors derive general results on market risk for variance swaps. Because variance is additive, a variance swap partway through its life is valued partly by realized vol (already observed) and partly by unrealized/implied vol, which is yet to be observed.

$$V(t_0, T)(T - t_0) = \underbrace{V(t_0, t)(t - t_0)}_{\text{Realized}} \\ + \underbrace{V(t, T)(T - t)}_{\text{Unrealized}} \quad (13)$$

It follows from this that at time t a variance swap with notional N has value

$$M(t) = N e^{-rT} [\lambda(V(t_0, t) - K_0) \\ + (1 - \lambda)(K_t - K_0)], \\ \lambda = (t - t_0)/(T - t_0) \quad (14)$$

The first piece is just the time-weighted value of realized variance against the strike. The second piece is the time-weighted difference of fair value variance strikes. The same formula can be used to decompose the daily price change into three risk components: gamma, and theta:

$$\Delta M(t) = M(t + \Delta t) - M(t) \\ = N \left(\underbrace{\frac{1}{\tau} V(t, t + \Delta t) \Delta t}_{\text{Gamma}} \right. \\ \left. + \underbrace{(1 - \lambda_{t+\Delta t}) \cdot \Delta K}_{\text{Vega}} - \underbrace{\frac{1}{\tau} K_t \Delta t}_{\text{Theta}} \right) \quad (15)$$

The biggest independent risk is in the fair value of the strike, which we have put in the vega component. This component also entails the skew risk.

Volatility Swaps

Although variance swaps can be statically replicated, volatility swaps (*see Volatility Swaps*) cannot. In [3], the authors show that there is an approximate dynamic replicating strategy for volatility

swaps. Before this, volatility swap valuation had been thought to be highly model dependent [2]. See also [8, 10], where the authors give a closed formula for valuing vol swaps using a GARCH model.

References

- [1] Bossu, S., Strasser, E. & Guichard, R. (2005). *Just What You Need to Know About Variance Swaps*, JP Morgan Equity Derivatives Research Publication.
- [2] Brockhaus, O. & Long, D. (2000). *Volatility Swaps Made Simple*, RISK Magazine, pp. 92–95.
- [3] Carr, P. & Lee, R. (2008). Robust replication of volatility derivatives, *Mathematics in Finance* Working Paper #2008-3. Courant Institute of Mathematical Sciences.
- [4] Carr, P. & Madan, D. (1998). Towards a theory of volatility trading, in *Volatility* ed., R.A. Jarrow, Risk Books.
- [5] Chriss, N. & Morokoff, W. (1999). *Market Risk for Volatility and Variance Swaps*, Risk Magazine.
- [6] Demeterfi, K., Derman, E., Kamal, M. & Zou, J. (1999). A guide to volatility and variance swaps, *Journal of Derivatives* **6**, 9–32.
- [7] Gatheral, J. (2006). *The Volatility Surface: A Practitioner's Guide*, Wiley Finance.
- [8] Haug, E.G. (2007). *Option Pricing Formulas*, 2nd Edition, McGraw Hill.
- [9] Hull, J. (2003). *Options, Futures, and other Derivatives*, 5th Edition, Prentice Hall.
- [10] Javaheri, A., Wilmott, P. & Haug, E.G. (2002). *Garch and Volatility Swaps*, published on Wilmott.com.
- [11] Neuberger, A. (1994). The log contract, *The Journal of Portfolio Management* **20**, 74–80.

Related Articles

Correlation Swap; Corridor Variance Swap; Gamma Swap; Realized Volatility and Multipower Variation; Realized Volatility Options; Volatility; Volatility Index Options; Volatility Swaps; Weighted Variance Swap.

ERIC LIVERANCE

Volatility Swaps

Volatility swaps are very similar to variance swaps, both in concept and in application. Like variance swaps, volatility swaps can be used by hedge funds to speculate on volatility movements or by portfolio managers to hedge other products against volatility fluctuations. Since their introduction in 1998, they have seen rapid growth and there is currently a sizable market in both the equity and foreign exchange markets. Volatility swaps are also traded in interest rate and commodity markets.

Technically speaking, a volatility swap is not really a swap, but a forward contract on the realized volatility. At maturity, the buyer receives from the seller the difference between the realized volatility and the fixed strike amount, multiplied by the dollar notional (quoted in dollar per volatility point):

$$\text{Volatility swap} = \text{Notional} \times (\sigma_{\text{realized}} - K) \quad (1)$$

while

$$\text{Variance swap} = \text{Notional} \times (\sigma_{\text{realized}}^2 - K^2) \quad (2)$$

The fixed strike payment is usually referred to as the *fixed leg* of the swap and the realized volatility is referred to as the *floating leg*. The swap contract contains the detailed specifications on how the volatility is calculated. Typically, the floating leg is calculated as

$$\sqrt{\text{Annualization} \times \sum_{i=1}^N \left(\ln \frac{S_i}{S_{i-1}} \right)^2} \quad (3)$$

where S_i should be adjusted by discrete dividends across dividend payment days. In the most accepted convention, the floating leg is reset and computed daily.

Besides variance swaps, another product closely related to volatility swaps is the VIX contract, which is written as the square root of the sum of expected future variances.

Because variance is the square of volatility, the payoff of a variance swap is convex in volatility while the payoff of a volatility swap is linear in volatility. A volatility swap is thus cheaper than the corresponding variance swap. More specifically, the fair strike for

the volatility swap is slightly lower than that of a variance swap. The difference between the two is known as the *convexity adjustment* and gets larger as volatility of volatility gets larger. The convexity adjustment can be calculated, for example, in the Heston model.

The variance swap is preferred in the equity market due to the fact that it can be replicated with a linear combination of vanilla options and a dynamic position in futures (*see Variance Swap*; [2]). In other markets, the volatility swap is actually more liquid than the variance swap. Although a position in variance swap can be replicated, a position in volatility swap cannot. This means that different models that correctly calibrate to the vanilla option surface will give the same price for variance swaps but not for the volatility swaps. In other words, the price of a variance swap is model independent, but the price of volatility swap is not. In practice, the volatility swap and variance swap admit almost equal pricing for short-term maturities. Recent research also suggest that the model dependence is not as large as it is commonly believed and volatility swaps can also be approximately replicated by trading vanilla options (*see Volatility Index Options; Realized Volatility Options*; [1]). Newer models have been developed that can price volatility derivatives including volatility swaps, while remaining consistent with the entire volatility surface [3].

References

- [1] Carr, P. & Lee, R. (2007). Realised volatility and variance: option via swaps, *Risk* **20**(5), 76–83.
- [2] Demeterfi, K., Derman, E., Kamal, M. & Zou, J. (1999). *More than You Ever Wanted to Know About Volatility Swaps*, Goldman Sachs Quantitative Strategies Research Notes.
- [3] Ren, Y., Madan, D. & Qian, M. (2007). Calibrating and pricing with embedded local volatility models, *Risk* **20**(9), 138–143.

Related Articles

Corridor Variance Swap; Realized Volatility Options; Variance Swap; Volatility Index Options; Weighted Variance Swap.

YONG REN

Static Hedging

Liquid traded put and call options can be used as hedge instruments for over-the-counter traded products. Barrier options are the most common exotic options, and, for these contracts, static hedging works out particularly well. In the Black–Scholes model there are simple methods (conceptually straightforward and/or closed form) for constructing replicating portfolios that do not require dynamic trading; they are set up at initiation of the barrier option, and liquidated at either knockout or expiry. They are, thus, static hedges. Inspired by Allen and Padovani [1], we describe how to find static hedges for barrier options in the Black–Scholes model in a way that encompasses both Derman’s [6] intuitive calendar-spread algorithm and Carr’s [4] strike-spread hedges stemming from put–call symmetry.

Construction of Static Hedges

Unless we explicitly say otherwise, we consider a Black–Scholes model throughout this article. This means that the interest rate is constant, and all options are written on some underlying asset S that follows a geometric Brownian motion. A zero-rebate, knockout barrier option is a contract that pays off as a plain vanilla option if S stays within a specified barrier for the whole life of the barrier option, but becomes worthless if the barrier is hit or crossed (*see also Barrier Options*). Recurrent examples are the down-and-out call and the up-and-out call. The value of a still-alive barrier option is of the form $F(S_t, t)$, where the function F solves the Black–Scholes partial differential equation with 0 as boundary condition along the barrier (*see Finite Difference Methods for Barrier Options*). This is illustrated in Figure 1, which is useful to keep in mind when the method for constructing static hedges is described in the following.

Construction of Static Hedges

A portfolio of puts and calls that (approximately) replicates the barrier option can be found as the solution to a linear system of equations, and constructing it does not require knowledge/implementation of barrier option valuation formulas. The idea

is to match the barrier option’s value at expiry and along the barrier.

To illustrate, consider a down-and-out call with strike K , expiry T , and barrier B . (An up-and-out call is treated similarly, except strike-above-the-barrier calls are used as hedge instruments.) Let Put , $Call$ (spot, time | strike, expiry) denote put and call values.

Suppose that we have specified a grid of time points $0 = t_0 < t_1 < \dots < t_n = T$, and n pairs of put with strikes $K_j \leq B$ and expiries $T_j \leq T$. Find the solution α to

$$A\alpha + u = 0 \quad (1)$$

where A is an $n \times n$ matrix with entries $A_{i,j} = Put(B, t_i | K_j, T_j)$ and u is an n -vector with entries $Call(B, t_i | K, T)$. A portfolio with the (K, T) -call and α_j in the (K_j, T_j) -put then matches the barrier option’s zero value at the t_i -points along the barrier, and its expiry payoff above the barrier. So the barrier option is—to a good approximation when the match points, the t_i ’s, are close—replicated buying this portfolio at time 0 and selling it either when the barrier option is knocked out (because sample paths are continuous, this can only happen if the barrier is actually hit) or when it expires. In other words, this represents a static hedge. There is freedom of choice regarding strikes and expiries of the hedge instruments. Derman [6] suggests calendar-spread hedging with strikes along the barrier, that is, using $T_j = t_{j-1}$ and $K_j = B$. This makes the A -matrix triangular so that we can solve for α_j ’s in one easy-to-explain backward-working pass. Another choice—closely related to Carr’s work [4]—is to use strike spreads, that is, $T_j = T$ for all j and K_j ’s that are different and below the barrier.

Example 1. Table 1 gives a numerical comparison of the performance of different hedge portfolios for three-month barrier options; a typical lifetime of a barrier option in foreign exchange markets. Looking at the results in Table 1 for the down-and-out call, we see the appeal of using options as hedge instruments; very few options are needed in the static hedges to achieve a hedge quality that is several orders of magnitude better than usual dynamic Δ -hedging. The numbers for the up-and-out call demonstrate one problem that static hedging does not immediately solve: the up-and-out call is a *reverse* or *live-out* option meaning that the underlying call is in the money when the barrier option knocks out. This

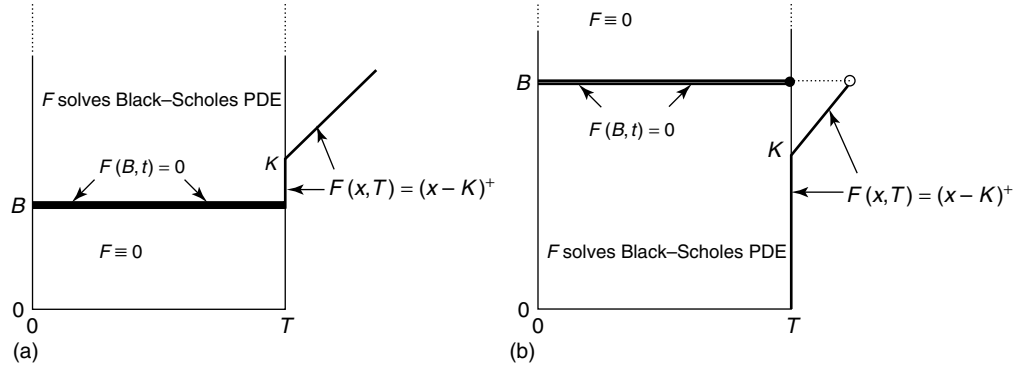


Figure 1 The PDEs for (a) down-and-out and (b) up-and-out call options

Table 1 Performance of dynamic and static hedge strategies in the Black–Scholes with 15% volatility and zero interest rate and dividends. The columns show the initial price of the hedge portfolio and the standard deviation of the benchmarked discounted hedge error, that is, the value of hedge portfolio at liquidation minus barrier option payoff relative to the initial value of the barrier option. All static hedges use three options besides the (K, T) -call. The time points for value matching, the t_i s, and the expiries for the calendar spreads are from the list $(0, 1/12, 2/12, 3/12)$. The strike spreads use calls with strikes $(110, 112, 114)$ for the up-and-out case, and puts with strikes $(90.25 = B^2/K, 88.25, 86.25)$ for the down-and out case. The Δ -hedge is adjusted daily and all portfolios are continuously monitored

Hedge method	Barrier option type			
	Down-and-out call $K = 100, T = 1/4, B = 95$		Up-and-out call $K = 100, T = 1/4, B = 110$	
	Cost	Standard deviation (%)	Cost	Standard deviation (%)
Dynamic; Δ	2.6964	11	1.0358	81
Static; strike spreads	2.6964	0	1.0674	19
Static; calendar spreads	2.6704	1.0	1.3468	94

discontinuity creates a large gap risk, and hedge quality deteriorates. To alleviate this, a number of regularization techniques have been suggested, for instance [10] using singular value decomposition when solving equation (1).

Beyond Black–Scholes Dynamics

Constructing static hedges by solving linear equations like equation (1) goes well beyond the Black–Scholes model. For constant elasticity of variance (asset volatility $\sigma S_t^{\gamma-1}$) and local volatility (asset volatility of the form $\sigma(S_t, t)$) models, the system carries over verbatim; the entries of the A -matrix are just calculated with a different formula/method. For jump-diffusion models [2], one needs to extend the grid of match points to space points beyond the barrier, and for stochastic volatility models [8], an extra

dimension is needed to match different volatility levels at knockout. By using both strike and calendar spreads, asymptotically perfect static hedges can be found in these two cases. It should be stressed that the static hedges are model and parameter dependent, but experimental and empirical evidence [7, 9] suggests a high degree of robustness to model risk.

Put–Call Symmetry and Static Hedges

In a number of papers [3–5], Peter Carr and coauthors have derived put–call symmetries and shown how they can be used to create static hedges for barrier options. In its basic form [4] [page 1167], the put–call symmetry states that in the zero-dividend, zero-interest rate Black–Scholes

model, we have

$$\begin{aligned} \text{Call}(S_t, t|K, T) &= (K/S_t) \times \text{Put}(S_t, t|S_t^2/K, T) \\ &\text{for all } S_t, t, K, \text{ and } T \end{aligned} \quad (2)$$

So a down-and-out call is replicated by buying one strike- K call, selling K/B puts with strike B^2/K , liquidating this position the first time that $S_t = B$, and if that does not happen, holding it until the options expire. More general symmetry relations enable one to find static hedges for such contracts as up-and-out calls, barrier options with rebates, lookback options, and double barrier options ([11] is a survey). Those static hedges will typically involve a continuum of plain vanilla options. Put–call symmetries also exist in models with nonzero interest rates and dividends, and more general dynamics than geometric Brownian motion (see [5]). Note that the strike-spread approach from the previous section finds the symmetry-based static hedges without explicit knowledge of closed-form results, and that the perfect replication of the down-and-out call in Table 1—where the strike- B^2/K put is included as a hedge instrument—demonstrates the basic put–call symmetry.

References

- [1] Allen, S. & Padovani, O. (2002). Risk management using quasi-static hedging, *Economic Notes* **31**, 277–336.
- [2] Andersen, L., Andreasen, J. & Eliezer, D. (2002). Static replication of barrier options: some general results, *Journal of Computational Finance* **5**, 1–25.

- [3] Bowie, J. & Carr, P. (1994). Static simplicity, *Risk Magazine* **7**(8), 44–50.
- [4] Carr, P., Ellis, K. & Gupta, V. (1998). Static hedging of exotic options, *Journal of Finance* **53**, 1165–1190.
- [5] Carr, P. & Lee, R. (2008). Put-call symmetry: extensions and applications, *Mathematical Finance* forthcoming.
- [6] Derman, E., Ergener, D. & Kani, I. (1995). Static options replication, *Journal of Derivatives* **2**, 78–95.
- [7] Engelmann, B., Fengler, M., Nalholm, M. & Schwendner, P. (2007). Static versus dynamic hedges: an empirical comparison for barrier options, *Review of Derivatives Research* **9**, 239–264.
- [8] Fink, J. (2003). An examination of the effectiveness of static hedging in the presence of stochastic volatility, *Journal of Futures Markets* **23**, 859–890.
- [9] Nalholm, M. & Poulsen, R. (2006). Static hedging and model risk for barrier options, *Journal of Futures Markets* **26**, 449–463.
- [10] Nalholm, M. & Poulsen, R. (2006). Static hedging of barrier options under general asset dynamics: unification and application, *Journal of Derivatives* **13**, 46–60.
- [11] Poulsen, R. (2006). Barrier options and their static hedges: simple derivations and extensions, *Quantitative Finance* **6**, 327–335.

Related Articles

Barrier Options; Finite Difference Methods for Barrier Options; Hedging; Put–Call Parity;

ROLF POULSEN

Corridor Variance Swap

A corridor variance swap, with corridor C , on an underlying Y is a *weighted variance swap* on $X := \log Y$ (unless otherwise specified), with weight given by the corridor's indicator function:

$$w(y) := \mathbb{1}_{y \in C} \quad (1)$$

For example, one may define an *up*-variance swap by taking $C = (H, \infty)$ and a *down*-variance swap by taking $C = (0, H)$, for some agreed H .

In practice, the corridor variance swap monitors Y discretely, typically daily, for some number of periods N , annualizes by a factor such as $252/N$, and multiplies by notional, for a total payoff

$$\text{Notional} \times \text{Annualization} \times \sum_{n=1}^N \mathbb{1}_{Y_n \in C} \left(\log \frac{Y_n}{Y_{n-1}} \right)^2 \quad (2)$$

If the contract makes dividend adjustments (as typical for contracts on single stocks but not on indices), then the term inside the parentheses becomes $\log((Y_n + D_n)/Y_{n-1})$, where D_n denotes the dividend payment, if any, of the n th period.

Corridor variance swaps accumulate only the variance that occurs while price is in the corridor. The buyer therefore pays less than the cost of a full variance swap. Among the possible motivations for a volatility investor to accept this trade-off and to buy up (or down) variance are the following. First, the investor may be bullish (bearish) on Y . Second, the investor may have the view that the market's downward volatility skew is too steep (flat), making down-variance expensive (cheap) relative to up-variance. Third, the investor may be seeking to hedge a short volatility position that worsens as Y increases (decreases).

Model-free Replication and Valuation

The continuously monitored corridor variance swap admits model-free replication by a static position in options and dynamic trading of shares, under conditions specified in **Weighted Variance Swap**, which include all positive continuous semimartingale

share prices Y under deterministic interest rates and proportional dividends.

Explicitly, one replicates using equation (7) of that article, with λ derived in [3]:

$$\lambda(y) = \int_{K \in C} \frac{2}{K^2} \text{Van}(y, K) dK \quad (3)$$

where $\text{Van}(y, K) := (K - y)^+ \mathbb{1}_{K < \kappa} + (y - K)^+ \mathbb{1}_{K > \kappa}$ for an arbitrary put/call separator κ .

Therefore, in the case that the interest rate equals the dividend yield (otherwise, *see Weighted Variance Swap*), a replicating portfolio statically holds $2/K^2 dK$ vanilla calls or puts at each strike K in the corridor C . The corridor variance swap model-independently has the same initial value as a claim on the time- T payoff $\lambda(Y_T) - \lambda(Y_0)$. Additionally, the replication strategy trades shares dynamically according to a “zero-vol” delta-hedge, meaning that its share holding equals the negative of what would be the European portfolio's delta under zero volatility.

For corridors of the type $C = (0, H)$ or $C = (H, \infty)$ where $H > 0$, taking $\kappa := H$ in equation (3) yields

$$\lambda(y) = (-2 \log(y/H) + 2y/H - 2) \mathbb{1}_{y \in C} \quad (4)$$

This λ , with H chosen arbitrarily, is also valid for the variance swap $C = (0, \infty)$

Further Properties

1. For a small interval $C = (a, b)$, the corridor variance swap approximates a contract on local time, in the following sense. Corridor variance satisfies

$$\begin{aligned} V_T^{(a,b)} &:= \int_0^T \mathbb{1}_{X_t \in (\log a, \log b)} d\langle X \rangle_t \\ &= \int_{\log a}^{\log b} L_T^x dx \end{aligned} \quad (5)$$

by the occupation time formula, where L_T^x denotes (an x -cadlag modification of) the *local time* of X . Therefore, at any point a ,

$$\frac{1}{\log b - \log a} V_T^{(a,b)} \longrightarrow L_T^a \quad \text{as } b \downarrow a \quad (6)$$

2. Corridor variance can arise from imperfect replication of variance. The replicating portfolio for

a standard variance swap holds options at all strikes $K \in (0, \infty)$. In practice, not all of those strikes actually trade. If we truncate the portfolio to hold only the strikes in some interval C , then the resulting value does not price a full variance swap but rather a C -corridor variance swap. (Moreover, in practice not even an interval of strikes actually trade, but rather a finite set, which can replicate instead a strike-to-strike notion of corridor variance, as shown in [1].)

3. In the case $C = (H, \infty)$, where $H > 0$, we rewrite equation (4) as

$$\lambda(y) = \frac{2}{H}(y - H)^+ - 2(\log y - \log H)^+ \quad (7)$$

Thus, the replicating portfolio is long calls on Y_T and short calls on $\log Y_T$.

Let F_{X_T} be the characteristic function of $X_T = \log Y_T$. Then techniques in [4, 5] price the calls on Y_T and $\log Y_T$, respectively. Specifically, assuming zero interest rates and dividends, we have the following semiexplicit formula for the corridor variance swap's fair strike:

$$\begin{aligned} \mathbb{E}\lambda(Y_T) - \lambda(Y_0) &= \frac{2}{H\pi} \int_{0-\alpha i}^{\infty-\alpha i} \operatorname{Re} \left(\frac{F_{X_T}(z-i)}{iz-z^2} e^{-iz \log H} \right) dz \\ &\quad + \frac{2}{\pi} \int_{0-\beta i}^{\infty-\beta i} \operatorname{Re} \left(\frac{F_{X_T}(z)}{z^2} e^{-iz \log H} \right) dz - \lambda(Y_0) \end{aligned} \quad (8)$$

for arbitrary positive α, β such that $\alpha + 1, \beta < \sup\{p : \mathbb{E}Y_T^p < \infty\}$, where $\mathbb{E}$ denotes expectation with respect to martingale measure.

In the case $C = (0, \infty)$, equation (4) implies the fair strike formula

$$\begin{aligned} \mathbb{E}\lambda(Y_T) - \lambda(Y_0) &= -2\mathbb{E} \log(Y_T/Y_0) \\ &= 2i F'_{X_T}(0) + 2 \log Y_0 \end{aligned} \quad (9)$$

In the case $C = (H_1, H_2)$, where $0 \leq H_1 < H_2$, subtract the formula for $C = (H_2, \infty)$ from the formula for $C = (H_1, \infty)$.

In the case of nonzero interest rates or dividends, add to equation (8) a correction involving payoffs at all expiries in $(0, T)$, as specified in equation (7a) in **Weighted Variance Swap**, and in equation (9) replace Y_0 by the forward price.

4. With discrete monitoring, the question arises, how to define up-variance and down-variance, and in particular how much variance to recognize, given a discrete move that takes Y across H . Definition (2) recognizes the full square of each move that ends in the corridor. Alternatively, the contract specifications in [2] treat the movements of Y across H by recognizing a fraction of the squared move. The fraction is defined in a way that admits approximate discrete hedging, in the sense that the time-discretized implementation of the continuous replication strategy has in each period a hedging error of only third order in that period's return.

References

- [1] Carr, P. & Lee, R. (2008). *From Hyper Options to Variance Swaps*, Bloomberg LP, University of Chicago.
- [2] Carr, P. & Lewis, K. (2004). Corridor variance swaps, *Risk* **17**(2), 67–72.
- [3] Carr, P. & Madan, D. (1998). Towards a theory of volatility trading, in *Volatility*, R., Jarrow, ed, Risk Publications, pp. 417–427.
- [4] Carr, P. & Madan, D. (1999). Option valuation using the fast Fourier transform, *Journal of Computational Finance* **3**, 463–520.
- [5] Lee, R. (2004). Option pricing by transform methods: extensions, unification, and error control, *Journal of Computational Finance* **7**(3), 51–86.

Related Articles

Delta Hedging; Gamma Swap; Realized Volatility Options; Variance Swap; Volatility Swaps; Weighted Variance Swap.

ROGER LEE

Gamma Swap

A gamma swap on an underlying Y is a *weighted variance swap* (see **Weighted Variance Swap**) on $\log Y$, with weight function

$$w(y) := y/Y_0 \quad (1)$$

In practice, the gamma swap monitors Y discretely, typically daily, for some number of periods N , annualizes by a factor such as $252/N$, and multiplies by notional, for a total payoff

$$\text{Notional} \times \text{Annualization} \times \sum_{n=1}^N \frac{Y_n}{Y_0} \left(\log \frac{Y_n}{Y_{n-1}} \right)^2 \quad (2)$$

If the contract makes dividend adjustments (as typical for single-stock gamma swaps but not index gamma swaps), then the term inside the parentheses becomes $\log((Y_n + D_n)/Y_{n-1})$, where D_n denotes the dividend payment, if any, of the n th period.

Gamma swaps allow investors to acquire variance exposures proportional to the underlying level. One application is dispersion trading of a basket's volatility against its components' single-name volatilities; as a component's value increases, its proportion of the total basket value also increases and, hence, so does the desired volatility exposure of the single-name contract. This variable exposure to volatility is provided by gamma swaps, according to point 1 of the "Further Properties" below. A second application is to trade the volatility skew; for example, to express a view that the skew slopes too steeply downward, the investor can go long a gamma swap and short a variance swap, to create a weighting $y/Y_0 - 1$, which is short downside variance and long upside variance. A third application is to trade single-stock variance without the caps often embedded in variance swaps to protect the seller from crash risk; in a gamma swap, the weighting inherently dampens the downside variance, so caps are typically regarded as unnecessary.

Model-free Replication and Valuation

The continuously monitored gamma swap admits model-free replication by a static position in options

and dynamic trading of shares, under conditions specified in **Weighted Variance Swap**, which include all positive continuous semimartingale share prices Y under deterministic interest rates and proportional dividends.

Explicitly, one replicates by using equation (7) of the above article, with

$$\begin{aligned} \lambda(y) &= \frac{2}{Y_0} [y \log(y/\kappa) - y + \kappa] \\ &= \int_0^\infty \frac{2}{Y_0 K} \text{Van}(y, K) dK \end{aligned} \quad (3)$$

where $\text{Van}(y, K) := (K - y)^+ \mathbb{1}_{K < \kappa} + (y - K)^+ \mathbb{1}_{K > \kappa}$ for an arbitrary put/call separator κ . Forms of this payoff were derived in, for instance, [2, 3].

Therefore, in the case that the interest rate equals the dividend yield (otherwise, see the weighted variance swap article), a replicating portfolio statically holds $2/(Y_0 K) dK$ vanilla calls or puts at each strike K . The gamma swap model independently has the same initial value as a claim on the time- T payoff $\lambda(Y_T) - \lambda(Y_0)$. Additionally, the replication strategy trades shares dynamically according to a "zero-vol" delta-hedge, meaning that its share holding equals the negative of what would be the European portfolio's delta under zero volatility.

Further Properties

Points 2–5 follow from equation 3. Point 1 uses only the definition 1.

1. For an index $Y_t := \sum_{j=1}^J \theta_j Y_{j,t}$, let $\alpha_{j,t} := \theta_j Y_{j,t} / Y_t$ be the fraction of total index value due to the quantity θ_j of the j th component $Y_{j,t}$. Define the cumulative *dispersion* D_t by

$$dD_t = \sum_{j=1}^J \alpha_{j,t} d[\log Y_j]_t - d[\log Y]_t \quad (4)$$

Going long $\alpha_{j,0}$ gamma swaps (non-dividend-adjusted) on each Y_j and short a gamma swap on Y creates the payoff

$$\begin{aligned} &\sum_{j=1}^J \alpha_{j,0} \int_0^T \frac{Y_{j,t}}{Y_{j,0}} d[\log Y_j]_t - \int_0^T \frac{Y_t}{Y_0} d[\log Y]_t \\ &= \int_0^T \frac{Y_t}{Y_0} dD_t \end{aligned} \quad (5)$$

as noted in [2]. Hence, a static combination of gamma swaps produces cumulative index-weighted dispersion.

2. By Corollary 2.7 in [1], if the implied volatility smile is symmetric in log-moneyness, and the dividend yield equals the interest rate ($q_t = r_t$), and there are no discrete dividends, then a gamma swap has the same value as a variance swap.
3. Assuming that $Y_T = Y_t R_{t,T}$ for all t , where the time- t conditional distribution of each $R_{t,T}$ does not depend on Y_t , the gamma swap has time- t gamma equal to a discounting/dividend-dependent factor times

$$\frac{2}{Y_0} \mathbb{E}_t \left(\frac{\partial^2}{\partial y^2} \Big|_{y=Y_t} y R_{t,T} \log(y R_{t,T}) \right) = \frac{2 \mathbb{E}_t R_{t,T}}{Y_0 Y_t} \quad (6)$$

where $\mathbb{E}$ denotes expectation with respect to martingale measure. Therefore, the *share gamma*, defined to be Y_t times the gamma, does not depend on Y_t . This property motivates the term *gamma swap*.

4. Within the family of weight functions proportional to $w(y) = y^n$, the gamma swap takes $n = 1$. In that sense, the gamma swap is intermediate between the usual logarithmic variance swap (which takes $n = 0$) and an arithmetic variance swap (which, in effect, takes $n = 2$). Expressed in terms of put and call holdings, the replicating portfolios in these three cases

hold, at each strike K , a quantity proportional to K^{n-2} . The gamma swap $O(1/K)$ is intermediate between logarithmic variance $O(1/K^2)$ and arithmetic variance $O(1)$.

5. Let F be the characteristic function of $\log Y_T$. If $\mathbb{E} Y_T^p < \infty$ for some $p > 1$ then

$$\mathbb{E} Y_T \log Y_T = -i F'(-i) \quad (7)$$

Gamma swap valuations are therefore directly computable in continuous models for which F is known, such as the Heston model (*see Heston Model*).

References

- [1] Carr, P. & Lee, R. Put–call symmetry: extensions and applications, *Mathematical Finance*, Forthcoming.
- [2] Mougeot, N. (2005). *Variance Swaps and Beyond*, BNP Paribas.
- [3] Overhaus, M., Bermúdez, A., Buehler, H., Ferraris, A., Jorindson, C. & Lamnouar, A. (2007). *Equity Hybrid Derivatives*, John Wiley & Sons.

Related Articles

Corridor Variance Swap; Delta Hedging; Realized Volatility Options; Variance Swap; Volatility Swaps; Weighted Variance Swap.

ROGER LEE

Atlas Option

In late 1990s, Societe Generale introduced a series of options on baskets of assets which are now commonly referred to as “Mountain Range” options [2]. They were introduced in part to replicate certain portfolio strategies and in part to extend single-name options to portfolios. What these options share is a strong dependence on the correlation structure of the assets, brought about by their nonlinear and path-dependent payoffs. But beyond this similarity, each type has its own distinct payoff tailored to its own risk profile and usage, making each deserving a study of its own. In a series of three articles, we look at the three commonly traded types of Mountain Range options—the Atlas option, the Himalayan option, and the Altiplano option.

We start with the Atlas option, which being the only non-path-dependent option in this group, is somewhat easier to analyze than the other two. This article is organized as follows. We first provide a description of the Atlas option and discuss the financial motivations for and strategies of its usage. We then discuss modeling, valuation, and risk issues that Atlas options share with all Mountain Range options, and conclude with a brief analysis of the risk profile that is unique to it. We remark that although the following discussion holds for a wide class of assets, such as foreign exchange (FX) and commodities, these options are traded mostly on baskets of stocks.

Contract Description

The payoff of the Atlas option is simply a call (or a put) option on the performance of a portfolio at maturity with the best and worst performing names removed. More precisely, given a portfolio, or basket of n stocks, let $S_i(t)$ be the price of the stock $i = 1, \dots, n$ at time t , with 0 being the start of the option and T its maturity. Furthermore, assume that the indices are such that $S_1(T)/S_1(0), \dots, S_n(T)/S_n(0)$, that is, the performance of the stocks is in increasing order. Given a strike K , the number of underperforming assets w and outperforming ones b to be removed,

the payoff of the Atlas option is

$$\max \left(\frac{1}{n - (w + b)} \sum_{i=1+w}^{n-b} \frac{S_i(T)}{S_i(0)} - K, 0 \right) \quad (1)$$

with the obvious condition that $b + w < n$.

If $b < w$, with b possibly 0, or if $b > w$ with w possibly 0, then this option becomes the best of or worst of option, respectively. On the other hand, if $b = w$, with equal number of underperforming and outperforming stocks removed, this option becomes, in effect, a “middle of the road” or an “average of averages” option. By removing the outliers, we are removing extreme risk and lowering the premium, while making it more favorable to risk-averse customers. For example, it can provide protection against defaults for the price of missing out on top performers.

Modeling

In the simplest of implementations, as in single-name options, the asset price processes are modeled as lognormal processes but with a correlation matrix. In more advanced implementations, to account for the volatility smile, some versions of stochastic volatility models are often used. One may even model some form of default component. However, in these more complex models, the modeling of correlation and its estimation become more complex as well.

Valuation and Risk

The number of assets in Mountain Range options generally ranges from a low of 4 or 5 to a high of about 20. Owing to their complex payoff and path-dependency, idiosyncratic characteristics of each asset need to be taken into account. Hence one cannot assume homogeneity of assets for neither small nor large baskets, making any closed-form approximation (especially in light of path dependence) intractable. Consequently, Mountain Range options, even the non-path-dependent Atlas options, are calculated using Monte Carlo simulation [1]. Monte Carlo methods, especially for high-dimensional payoffs with large number of assets, are slow to converge, and usually one or more variance-reduction techniques are employed. This problem is exacerbated further when calculating first and second order Greeks.

But even in the simple lognormal model, the sheer size of the correlation matrix can become a challenge. Since for n assets there can be $n(n-1)/2$ distinct correlations, even for a modest basket of 10 assets, 45 different correlations are possible. Moreover, it is not clear how one can obtain the correlation numbers themselves. If, theoretically speaking, there existed $n(n-1)/2$ traded spread options on each pair, their implied correlations could be used with the spread options as hedges. However, it is unlikely that every pair of assets in a basket would have a traded spread option. Even if they did, their sheer number would make transaction costs prohibitive, even for moderate bid-ask spreads. Hence historical correlations are more often used, even though as with all historical estimates, they are hard to hedge and can change with macro- and microeconomic shifts. When all assets belong to the same sector, a single correlation number is commonly used. This high amount of asset interdependence makes cross-gammas (*see* **Gamma Hedging**) important, adding further to the hedging complexity.

Risk Profile

When $w = b = 0$, the Atlas option is simply a call option on the average performance of a basket. As in a vanilla call on a single stock, the higher the volatility, the higher the price. But the analysis gets more interesting when we start removing good and bad performers at maturity.

For simplicity, we look at a homogeneous portfolio with identical pairwise correlations given by a single number in a simple lognormal model. For very high correlations, the basket behaves as a single asset—thus removing assets has a small effect on

the option payoff. Since it behaves as a single asset, as in single-asset calls, the option's payoff generally increases with volatility.

For low correlations, on the other hand, the basket has a high dispersion at maturity—on average, a few stocks will have a high price and the rest low—and higher the volatility, the higher the dispersion. Since the expectation of the sum of asset prices at maturity is independent of both volatility and correlation, having a few high asset prices implies that many others are very low in most paths. But it is precisely these high-contribution assets that are removed from the basket, leaving the basket with low-priced assets, thus reducing the price of the option. So for low correlation, with $b \neq 0$, increasing volatility does not necessarily increase the price. We remind the reader that this simple analysis applies to a homogeneous basket. Individual volatilities, dividends, and a non-constant correlation can affect the payoff in ways not always easily explained.

References

- [1] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*, Springer, New York.
- [2] Mountain Range Options, document downloadable from global-derivatives.com

Related Articles

Altiplano Option; Basket Options; Correlation Risk; Himalayan Option.

REZA K. GHARAVI

Himalayan Option

In the late 1990s, Societe Generale introduced a series of options on baskets of assets that are now commonly referred to as *Mountain Range* options [2]. They were introduced in part to replicate certain portfolio strategies and in part to extend single-name options to portfolios. What these options share is a strong dependence on the correlation structure of the assets, brought about by their nonlinear and path-dependent payoffs. But beyond this similarity, each type has its own distinct payoff tailored to its own risk profile and usage, making each deserving a study of its own. In this second of three articles on Mountain Range options, we look at the Himalayan option. Unlike Atlas options, Himalayans are path dependent, and are usually longer dated than Atlas options.

This article is organized as follows. We first provide a description of the Himalayan option and discuss the financial motivations for and strategies of its usage. We then discuss modeling, valuation, and risk issues that Himalayan options share with all Mountain Range options, and conclude with a brief analysis of the risk profile that is unique to it. We remark that although the following discussion holds for a wide class of assets, such as foreign exchange (FX) and commodities, these options are traded mostly on baskets of stocks.

Contract Description

The payoff of the Himalayan option is best described in words. Given n assets and contractual times $T_1, \dots, T_n$ (usually yearly), at first time T_1 we take the best performing stock, record its performance, and then remove it from the basket. We continue until at maturity T_n we are left with the last stock. The payoff of the option is the sum of the performances on these n contractual times. In some variations, the top two or three are removed at each time.

What this option emulates is the greedy strategy of liquidating a portfolio—the best performing stocks are sold first. Since this is a derivative product, Himalayan options may offer regulatory and tax benefits compared to actual holding of a portfolio.

Modeling

As in single-name options, in the simplest of implementations, the asset price processes are modeled with lognormal processes endowed with a correlation matrix. In more sophisticated implementations, to account for the volatility smile, some versions of stochastic volatility models are often used. One may even include some form of default component in the asset price. However, in these more complex models, the modeling of correlation and its estimation become more complex as well.

Valuation and Risk

The number of assets in Mountain Range options generally ranges from a low of 4 or 5 to a high of about 20. Owing to their complex payoff and path dependency, idiosyncratic characteristics of each asset need to be taken into account. Hence one cannot assume homogeneity of assets for either small or large baskets, making any closed-form approximation (especially in light of path dependence) intractable. As a result, Mountain Range options, specially the path-dependent varieties such as the Himalayan, are calculated using Monte Carlo simulation [1]. In light of their high-dimensional payoffs due to large number of assets and time points, to speed up convergence, specially for first- and second-order Greeks, usually one or more variance-reduction techniques are employed.

The other challenge posed by these options is the correlation. Even in simple lognormal models, the sheer size of the correlation matrix can become a challenge. Since for n assets there can be $n(n-1)/2$ distinct correlations, even for a modest basket of 10 assets, 45 distinct correlations are possible. Moreover, obtaining the pairwise correlations themselves is not straightforward. If, theoretically speaking, there existed $n(n-1)/2$ traded spread options on each pair, their implied correlations could be used with the spread options as hedges. However, it is unlikely that every pair of assets in a basket would have a traded spread option. Even if they did, their sheer number would make transaction costs prohibitive, even for moderate bid–ask spreads. Hence historical correlations are more often used, even though as with all historical estimates, they are hard to hedge

and can change with macro- and microeconomic shifts. When all assets belong to the same sector, a single correlation number is commonly used. This high amount of asset interdependence makes cross-gammas important, adding further to the hedging complexity.

Risk Profile

Payoffs for Himalayan options can be surprising. For simplicity, we look at a homogeneous portfolio with identical pairwise correlations given by a single number in the simple lognormal model. We look at the effects of removing the best performing assets compared to removing assets in a predetermined order. For high correlations, since basket assets move together, the effects of the “greediness” is small—removing the best performing assets leaves the portfolio with similarly performing assets. So compared to removal by some predetermined order, its effects are small. Increasing the volatility increases the dispersion, and as we next see for the case of low-correlation baskets, it may adversely impact the value of the option.

Assets with low correlation, on the other hand, become more disperse as time passes. Since the expected sum of assets at termination of the option is independent of correlation, removing the best

performing asset would in most scenarios leave the worse performing assets, and since no asset is given a chance to grow—it would get removed in the next cut—we are left with increasingly worse performing assets. Thus greediness would actually be a bad strategy compared to some predetermined asset-removal order. Increasing volatility increases this dispersion, further reducing the payoff of the Himalayan. We remind the reader that this simple analysis applies to a homogeneous basket. Individual volatilities, dividends, and a nonconstant correlation can affect the payoff in ways not always easily explained.

References

- [1] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*, Springer, New York.
- [2] Mountain Range Options, document downloadable from global-derivatives.com.

Related Articles

Altiplano Option; Atlas Option; Basket Options; Correlation Risk.

REZA K. GHARAVI

Altiplano Option

In the late 1990s, Societe Generale introduced a series of options on baskets of assets that are now commonly referred to as *Mountain Range* options [2]. They were introduced in part to replicate certain portfolio strategies and in part to extend single-name options to portfolios. What these options share is a strong dependence on the correlation structure of the assets, brought about by their nonlinear and path-dependent payoffs. But beyond this similarity, each type has its own distinct payoff tailored to its own risk profile and use, making each deserving of its own study. In this last of three articles on Mountain Range options, we look at the Altiplano option, which can be thought of as an extension of barrier options to baskets.

This article is organized as follows. We first provide a description of the Altiplano option and discuss the financial motivations for and strategies of its usage. We then discuss modeling, valuation, and risk issues that Altiplano options share with all Mountain Range options, and conclude with a brief analysis of the risk profile that is unique to it. We remark that although the following discussion holds for a wide class of assets, such as foreign exchange (FX) and commodities, these options are traded mostly on baskets of stocks.

Contract Description

Like single-name barrier options, Altiplano options have two components—a vanilla-type payoff if a barrier event occurs, and a coupon payoff if it does not. Usually the barrier event is to have at least one stock reach a predetermined barrier.

More precisely, given a portfolio, or a basket of n stocks, let $S_i(t)$ be the price of the stock $i = 1, \dots, n$ at time t , with 0 being the start of the option and T its maturity. The payoff is

$$\max \left(\sum_{i=1}^n \frac{S_i(T)}{S_i(0)} - K, 0 \right) I_B + (1 - I_B)C \quad (1)$$

where I_B is 1 if a barrier event in relation to barrier level B occurs, and is 0 if it does not, in which case the option would pay a coupon C . Here we have used a call option, but, in principle, we could

use any option. As in single-name barriers, there are many possible wrinkles for the barrier event, such as the Parisian type (*see Parisian Option*). But unlike a simple extension from the single-name case where the barrier is triggered by the sum of the portfolio, individual assets can trigger the barriers by themselves. In the example above, all it takes is for one asset to activate a barrier, independently of the level of other assets at the time. This makes the Altiplano sensitive to individual asset moves, rather than the collective sum.

As in single-name barriers, since I_B is always less than 1, the payoff, and thus the risk, are lower than for standard options on baskets, which makes their premiums lower as well (assuming C is small or zero). The lower premium makes it more attractive when used as a hedge, for example.

Modeling

In the simplest of implementations, as in single-name options, the asset price processes are modeled as lognormal processes but with a correlation matrix. In more advanced implementations, to account for the volatility smile, some versions of stochastic volatility models are often used. One may even include some form of default component. However, in these more complex models, the modeling of correlation and its estimation become more complex as well.

Valuation and Risk

The number of assets in Mountain Range options generally ranges from a low of 4 or 5 to a high of about 20. Owing to their complex payoff and path-dependency, idiosyncratic characteristics of each asset need to be taken into account. Hence one cannot assume homogeneity of assets for either small or large baskets, making any closed-form approximation (especially in light of path dependence) intractable. As a result, Mountain Range options, specially the path-dependent varieties such as the Altiplano, are calculated using Monte Carlo simulation [1]. Monte Carlo methods, especially for high-dimensional payoffs with large number of assets and time points, are slow to converge, and usually one or more variance-reduction techniques are employed. Additionally, since the barrier event is binary, the number of simulation paths needed is even greater than those with

continuous payoffs, making first- and second- order Greeks calculations even more noisy. This makes use of variance-reduction methods even more critical.

The other challenge posed by these options is the correlation. Even in the simple lognormal model, the sheer size of the correlation matrix can become a challenge. Since for n assets, there can be $n(n - 1)/2$ distinct correlations, even for a modest basket of 10 assets, 45 distinct correlations are possible. Moreover, it is not clear how one can obtain the pairwise correlations themselves. If, theoretically speaking, there existed $n(n - 1)/2$ traded spread options on each pair, their implied correlations could be used with the spread options as hedges. However, it is unlikely that every pair of assets in a basket would have a traded spread option. Even if they did, their sheer number would make transaction costs prohibitive, even for moderate bid–ask spreads. Hence historical correlations are more often used, even though as with all historical estimates, they are hard to hedge and can change with macro- and microeconomic shifts. When all assets belong to the same sector, a single correlation number is commonly used. This high amount of asset interdependence makes cross-gammas important, adding further to the hedging complexity.

Risk Profile

As in single-asset barrier options, the payoff of an Altiplano option is determined by two competing

terms—the option term and the barrier term. Again, for ease of analysis, we look at a homogeneous portfolio with identical pairwise correlations given by a single number in the simple lognormal model with no coupon payout. For a simple call option on a basket, it is known that high correlation and volatility increase its price. For the barrier event, it depends on the type. If it takes only one asset to hit a barrier, then low correlation and high volatility increase the probability of hitting it. Moreover, depending on the barrier, the paths that lead to higher option prices may make the barrier even more or less likely, leading to possible nonmonotonous behavior. So even in this simple homogeneous case with a single correlation, the behavior is rather complex.

References

- [1] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*, Springer, New York.
- [2] *Mountain Range Options*, document downloadable from global-derivatives.com

Related Articles

Atlas Option; Basket Options; Correlation Risk; Himalayan Option.

REZA K. GHARAVI

Constant Proportion Portfolio Insurance

Portfolio insurance is a dynamic management technique that aims at giving the investor the ability to limit the downside risk while allowing some participation in the upside market. Option-based portfolio insurance combines a position in the risky asset with a put option on the asset to achieve this goal. In many cases, options on a portfolio may not be available. *Constant proportion portfolio insurance* (CPPI) is an alternative to that approach. CPPI was first introduced by Perold [4] for fixed-income instruments and by Black and Jones [2] for equity instruments.

CPPI utilizes a rule-based strategy to allocate assets dynamically over time. It involves maintaining a dynamic mix of a “riskless” asset (usually treasury bills or liquid money market instruments) and the risky asset, usually a market index. In the case of having more than one risky asset, an index is formed, which would be treated as a single risky asset. The weights in the index do not change during the life of the trade.

The strategy is based on the notion of *cushion*, which is the difference between the current portfolio value and the guaranteed level, called *bond floor*. Obviously, the initial floor F_0 is less than the initial portfolio value V_0 . Define C_t to be t -time value of the cushion, that is

$$C_t = V_t - F_t \quad (1)$$

The final payoff at maturity is the maximum of these two quantities: (i) the value of the portfolio at maturity and (ii) the guaranteed level. In a nutshell, CPPI is a path-dependent, self-financing capital guarantee structured product that has final payoff linked to the performance of a pool of assets.

Throughout the existence of the contract, an amount of wealth is invested into the risky asset. This amount called *exposure* is proportional to the cushion and is calculated by multiplying the cushion by a predetermined *multiple*.

$$e_t = mC_t \quad (2)$$

The remainder of wealth is allocated into the riskless asset. Trivially, the higher the multiple, the more the holder will participate in rising markets and

the faster the portfolio value approaches the bond floor in downturn markets.

Both the bond floor and multiple are specified in the contract and indicate the investor appetite for risk. Assuming Black–Scholes framework for risky asset S_t

$$dS_t = \mu S_t dt + \sigma S_t dW_t \quad (3)$$

and assuming floor evolves as follows:

$$dF_t = r F_t dt \quad (4)$$

the value of the portfolio at time t , as shown in [5] and [3], is

$$V_t(S_t, m) = F_0 \exp(rt) + \alpha_t S_t^m \quad (5)$$

where $\alpha_t = \frac{C_0}{S_0^m} \exp(\beta t)$ and β is given by

$$\beta = r - m \left(r - \frac{1}{2} \sigma^2 \right) - m^2 \frac{\sigma^2}{2} \quad (6)$$

In periods of negative performance, a specified amount of the risky asset, according to a predetermined asset allocation formula, is liquidated and used to purchase riskless assets. On the other hand, when market goes up, a specified amount of riskless assets, according to the formula, is liquidated and proceeds are used to purchase the risky asset. The provider undertakes the risk of managing of the pool of assets (both risk-free and risky assets).

For risk managing of the pool in CPPI, the provider receives an annual fee. The fee is specified in the contract in one of the following ways: (i) as a fixed percentage of the initial notional per annum, (ii) as a fixed percentage of the value of the portfolio per annum (path dependent), and (iii) as a fixed percentage of the value of the equity held in the pool per annum (path dependent). The first two are more common than the third one.

The provider also receives the potential value of dividends on the equity. This amount is also path dependent.

If there is more than one risky asset in the pool, they would be treated like a basket in a basket option, so that in the event of rebalancing the collection of risky assets is treated like a single underlying.

Rebalancing Procedure

The asset allocation formula is a part of the contract. The terms in the formula are negotiated between the

provider and the counterparty before entering into the transaction.

A feature is built into the allocation formula in order to avoid constant rebalancing. Rebalancing occurs where the difference between theoretical equity exposure from the formula and the actual equity allocation is greater than a predefined number of percentage points specified in the contract.

In the case that it is triggered, the provider sells/purchases an amount of equity and purchases/sells risk-free assets to make the actual ratio equal to the number generated by the formula. The timetable for rebalancing is up to the investor, with monthly or quarterly rebalancing being often cited.

Gap Risk

The risk that the value of the portfolio is less or equal to the bond floor is called the *gap risk*.

If there is no drastic jump in the value of the risky asset for the life of the trade, then there is no need for injection of money for rebalancing of the pool. Therefore downfall is lower than the gap risk and the value of the pool would be above the bond floor. In that case, hedging the CPPI trade would be risk free.

There are cases that the value of the pool may go under the bond floor. Either the provider cannot liquidate the risky asset due to illiquidity or the value of the equity asset has dropped so much that proceeds are not sufficient to maintain the value of the portfolio above the floor—the market simply drops by more than the gap risk before a rebalancing can be undertaken. In either case, the provider must make up the shortfalls. Whenever the portfolio value reaches a given floor, the investor receives a given amount. At this point, the entire pool comprises of 100% exposure to riskless asset. The gap risk is presented as basis points per annum.

Modeling Gap Risk

Modeling the gap risk is the main concern in CPPI. It is analogous to a string of one-day out-of-the-money put options.

Using standard Black–Scholes formula would not be ideal for the following two reasons: (i) lack of volatility information on such deep out-the-money

options and (ii) fat-tailed behaviors observed in stock returns distributions.

In [1], the authors apply extreme value theory to determine the multiple. A quantile hedging approach is introduced, which provides an upper bound on multiple. This bound is statistically estimated from the behavior of extreme variations in rates of asset returns. The authors also introduce the distributions of interarrival times of these extreme movements and show their impacts on CPPI.

In [5] the authors analyze the cost of the guarantee and the performance of portfolio based on such a strategy. They provide two extensions. One is based on Levy processes that allow jumps into the dynamics of the underlying asset. Second, they deal with insurance against all hitting times of modified floor.

In [3], Cont and Tankov study the behavior of CPPI strategies in models where the price of the underlying portfolio may experience downward jumps. That allows them to quantify the gap risk while maintaining the analytical tractability of the continuous-time framework. With respect to the work done in [5], in [3] the authors consider various risk measures for the loss and provide an analytical method to compute them.

CPPI techniques have also been applied to credit portfolios (*see Credit Portfolio Insurance*).

References

- [1] Bertrand, P. & Prigent, J.-L. (2002). Portfolio insurance: the extreme value approach to the CPPI method, *Finance* **23**(2), 69–86.
- [2] Black, F. & Jones, R. (1987). Simplifying portfolio insurance, *Journal of Portfolio Management* **14**(1), 48–51.
- [3] Cont, R. & Tankov, P. (2009). Constant proportion portfolio insurance in presence of jumps in asset prices, *Mathematical Finance* **19**(3), 379–401.
- [4] Perold, A.R. (1986). *Constant Proportion Portfolio Insurance*, Harvard Business School, Working Paper.
- [5] Prigent, J.-L. & Tahar, F. (2005). *CPPI with Cushion Insurance*, University of Cergy-Pontoise, working paper.

Related Articles

Credit Portfolio Insurance.

ALI HIRSA

Equity Default Swaps

Equity default swaps (EDS) are equity derivatives that are structured as far out-of-the-money American-style binary puts with periodic swap payments rather than an up-front premium. The structure of EDS is similar to **credit default swaps** (CDS) except that the default event is defined in terms of a decline in the share price of the reference entity rather than a credit event experienced by the reference entity. Thus, similar to a CDS, an EDS can be seen in terms of the protection buyer and protection seller counterparties. In the case of an EDS, protection buyers are hedging themselves against a large decline in the share price.

More specifically, the protection buyer in an EDS makes periodic fixed payments to the protection seller. The size of the periodic payments is called the *EDS spread*. In return for the periodic payments, a default payment from the protection seller to the protection buyer is made if the share price of the reference entity declines a prespecified amount from the share price at initiation of the EDS contract. If this equity default event occurs and the default payment is made, then the contract terminates with the protection buyer ceasing to make further payments. If the equity default event never occurs before the maturity of the EDS contract, then no payment is ever made from the protection seller to the protection buyer.

Typically, the prespecified fall in the share price is a 70% decline, so that the equity default event is defined as the first time that the shares of the reference entity trade at 30% of the share price when the EDS contract is entered. The amount of the default payment is fixed when the EDS contract is initiated and computed as

$$N \times (1 - R) \quad (1)$$

where N is the notional value of the contract and R is the recovery rate, which is predetermined and typically set to be 50%. The formulation of the default payment in terms of a recovery rate is to further the analogy to CDS. EDS are usually medium-term contracts with maturities of five years being the most common.

The first EDS were issued in 2004 as a means to allow protection sellers to receive a higher return

than from CDS. A credit event almost certainly implies that an equity default event will occur, but, conversely, an equity default event can occur without a credit event having occurred. This implies that for the same reference entity, the EDS spread must be greater than the CDS spread. Besides, since protection sellers receiving a higher return for EDS, the default event for EDS is easier to define than for CDS. Whether or not the share price has reached a predetermined level is unambiguous, but the various credit events can sometimes cause confusion for counterparties and have led to legal proceedings.

One difference between EDS and most equity derivatives is the swap feature. Not only is there no up-front premium but also upon an equity default event, no further swap payments are made. Gil-Bazo [4] finds the fraction of the EDS spread that is due to this swap feature under the Black–Scholes model, given plausible parameter values.

Applications of EDS

Besides the obvious application of EDS as portfolio protection on long positions in shares, EDS are often used in ways that exploit their similarity to CDS. Two examples are given here: relative value trades in EDS–CDS carry trades and as yield-enhancing replacements for CDS in **collateralized debt obligations** (CDO).

In EDS–CDS carry trades, an investor sells protection with an EDS and buys protection with a CDS. The larger EDS spread is received while the smaller CDS is paid, so that the investor simply collects the positive carry if neither a credit nor an equity default event occurs. In the case that both default events occur, the investor is partially hedged with the effectiveness of the hedge depending on the relative timing of the default events as well as the recovery rate on the credit event. There is a risk of large losses if an equity default occurs without a credit event occurring.

Some CDO were constructed with the reference portfolio including some EDS in addition to CDS. While this increases the risk to the CDO investors, there can be significant yield enhancement. There have even been CDO where the reference portfolio was exclusively composed of EDS.

These are termed *equity collateralized obligations (ECO)*.

EDS as Credit–Equity Hybrids

Since a large fall in the equity price is needed to trigger the payout of an EDS, there are often accompanying credit implications to the firm leading many to refer to EDS as a *credit–equity hybrid instrument*, despite the default event being defined exclusively in terms of the share price.

There is empirical evidence that EDS should not be considered simply as an equity derivative. de Servigny and Jobst [8] assess the relative weighting of debt and equity factors for equity default probabilities. They find that for the typical definition of equity default as 30% of the initial share price, debt factors are more important than equity factors. In addition, Jobst and de Servigny [5] study EDS correlation and EDS–CDS correlation and find that multivariate analyses commonly used for credit are the most appropriate.

EDS Pricing

Consider the problem of finding the fair spread for an EDS that is to be initiated now at time $t = 0$, with the current stock price as S_0 . The default payment in an EDS is made the first time that the share price trades at the prespecified level L where $L < S_0$. In addition to this, the swap payments are made contingent on the share price not being traded at or below the prespecified level. Thus, to price an EDS, the probability distribution of the first passage time of the level L is required. Here, the first passage time of L is defined as

$$\tau_L = \inf\{t > 0; S_t \leq L\} \quad (2)$$

Then, if an EDS with \$1 notional requires a periodic payment of C at times $T_1, \dots, T_n$ the price of the EDS for a protection buyer is

$$\begin{aligned} EDS(S_0; C, R, L) = & -C \sum_{i=1}^n D^{T_i} P(\tau_L > T_i | S_0) \\ & + (1 - R) E[D^{\tau_L} 1_{\{\tau_L \leq T_n\}} | S_0] \end{aligned} \quad (3)$$

where R is the recovery rate, D^t is the discount factor to time t and P and E are probabilities and expectations under the risk-neutral measure. Then, the fair spread for an EDS is the value of C that makes $EDS(S_0; C, R, L) = 0$.

The problem of deriving the first passage time distribution has been solved in several special cases. Albanese and Chen [1] compute the EDS spread under the assumption that the stock price follows a constant elasticity of variance (CEV) process (see **Constant Elasticity of Variance (CEV) Diffusion Model**) and Asmussen *et al.* [2] use the Wiener–Hopf factorization (see **Wiener–Hopf Decomposition**) to compute the EDS spread under the assumption that the stock price follows a Carr–Geman–Madan–Yorr (CGMY) Lévy process (see **Tempered Stable Process**). For models where credit considerations are more explicitly addressed, EDS spreads have been priced using different methods: Albanese and Chen [1] use a credit barrier model with a credit to equity mapping; Campi *et al.* [3] extend the CEV case to include jump to default; Medova and Smith [6] use a structural model of credit risk (see **Structural Default Risk Models**) where the firm’s asset value follows a geometric Brownian motion; and Sepp [7] makes use of an extended structural model where the firm’s asset value can have stochastic volatility (see **Heston Model**) or be a double exponential jump diffusion, with the default barrier being deterministic or stochastic.

References

- [1] Albanese, C. & Chen, O. (2005). Pricing equity default swaps, *Risk* **18**, 83–87.
- [2] Asmussen, S., Madan, D. & Pistorius, M.R. (2008). Pricing equity default swaps under an approximation to the CGMY Lévy Model, *Journal of Computational Finance* **11**, 79–93.
- [3] Campi, L., Polbennikov, S. & Sbuelz, A. (2009). Systematic equity-based credit risk: a CEV model with jump to default, *Journal of Economic Dynamics and Control* **33**, 93–108.
- [4] Gil-Bazo, J. (2006). The value of the ‘swap’ feature in equity default swaps, *Quantitative Finance* **6**, 67–74.
- [5] Jobst, N. & de Servigny, A. (2006). An empirical analysis of equity default swaps (II): multivariate insights, *Risk* **19**, 97–103.

-
- [6] Medova, E. & Smith, R. (2006). A structural approach to EDS pricing, *Risk* **19**, 84–88.
 - [7] Sepp, A. (2006). Extended CreditGrades Model with stochastic volatility and jumps, *Wilmott Magazine* September, 50–62.
 - [8] de Servigny, A. & Jobst, N. (2005). An empirical analysis of equity default swaps (I): univariate insights, *Risk* **18**, 84–89.

Related Articles

Constant Proportion Portfolio Insurance; Credit Default Swaps; Total Return Swap; Variance Swap.

OLIVER CHEN

Exchange-traded Funds (ETFs)

Exchange-traded funds (ETFs) are indexed products and trade like equities on the main exchanges. Technically, an ETF holder possesses certificates that state legal right of ownership over a portion of a basket of individual stock certificates. It may be tempting to think of ETFs as mutual funds. However, ETFs differ from mutual funds in several ways. One key distinction is that mutual funds only trade at the end of each day at their calculated NAVs while ETFs trade throughout the day at ever-changing prices. Another is that ETFs, unlike mutual funds, can be sold short.

A number of financial entities are required to create and maintain ETFs. Initially, the fund manager submits a detailed plan to SEC as to the ETF's constitution and how it will function. Once the plan is approved, the fund manager enters into an agreement with a market maker or specialist known as the *authorized participant*. The authorized participant, typically by borrowing, begins to assemble a basket of instruments that compose the index the ETF is meant to replicate. Once assembled, the instruments are placed in trust at a custodial bank that, in turn, uses them to form what are known as *creation units*. Each creation unit represents a subset of the basket of instruments. The custodial bank then divides the creation units into (typically) 10 000 to 600 000 ETF shares [5], which are legal claims on the aforementioned instruments, and forwards the shares to the authorized participant. It should be mentioned that this is an in-kind trade (i.e., ETF shares are exchanged for the basket of financial instruments) with no tax implications. The authorized participant subsequently sells the shares in the open market just like shares of stock. The ETF shares then continue to be sold and resold by investors. The instruments underlying the creation units, and therefore the ETF shares, remain in trust with the custodian who is responsible for paying any cash flows (e.g., dividends and coupons) from the instruments to the ETF holders and providing administrative oversight of the creation units themselves.

A long position in an ETF can be unwound two ways. The first is simply to sell the share in the open

market. The second is to purchase enough ETF shares to form a creation unit and then exchange the creation unit for the securities that comprise it. As with the creation of the ETF shares, this second option has no tax implications but is generally only available to large institutional investors.

The financial firms mentioned above are motivated to participate in the ETF space by different profit opportunities. Fund managers and custodial banks each collect a small portion of the fund's annual assets. Investors who loan instruments that compose the aforementioned baskets receive interest fees, while market makers seek to earn both arbitrage (i.e., the difference in price between the ETF and the basket of instruments) and bid/ask-spread profits.

According to a July 2008 Morningstar article, the average ETF charged 54 bps in annual fees [3]. This number was up from 41 bps a year earlier [3]. The average has been raised owing to recently formed exotic, narrowly focused ETFs. However, there are still many broad-market ETFs with annual fees on the order of 10 bps. Furthermore, an ETF's management fee is usually lower than the (approximately) 80 bps charged by the typical mutual fund [2]. Finally, over 90% of US ETFs have bid/ask spreads of fewer than 50 bps (with over half having spreads of fewer than 20 bps) [1], while the expense ratio (which includes management fees, administrative costs, 12b-1 distribution fees, and other operating expenses) for an average mutual fund is 150 bps [4]. Assuming one must transact at the bid (offer) when selling (buying), even ETFs with higher bid/ask spreads have, on average, substantially lower frictional costs than mutual funds (104 bps vs. 150 bps).

There are close to 2000 ETFs trading today, tracking numerous broad-market composites as well as a multitude of sector and geographic indexes. These products cover approximately 40 different investment categories (e.g., Utilities Sector Equity, High Yield Fixed Income, US Real Estate Equity and European Mid/Small Cap Equity) and fall into several investment styles (e.g. equity, fixed income, and alternatives). Furthermore, these funds are offered by a large number of investment banks and investment management companies. Consequently, the ETF market gives investors numerous choices both in terms of the nature of investment and the fund manager.

In conclusion, a number of financial entities are necessary to create and maintain ETFs, indexed products that trade like stocks on major bourses. Furthermore, ETFs differ from mutual funds in that they trade throughout the day, can be shorted, and generally carry lower fees. Finally, the close to 2000 ETFs trading today cover a plethora of different investment categories, offering investors an inexpensive means to construct well-diversified portfolios.

References

- [1] Amery, P. (2008). *European ETF Secondary Market Dealing Spreads*, Index Universe, Retrieved on July 25, 2008 from <http://www.indexuniverse.com/sections/features/12/4294-european-etf-secondary-market-dealing-spreads.html>
- [2] Kinnel, R. (2007). *Fund Fees are Coming Down*, Morningstar, Retrieved on July 25, 2008 from <http://ibd.morningstar.com/article/article.asp?CN=aol828&id=194298>
- [3] Marquardt, K. (2008). *Surprise: ETF Fees are Going Up*, U.S. News, Retrieved on July 25, 2008 from <http://www.usnews.com/blogs/new-money/2008/7/9/surprise-etf-fees-are-going-up.html>
- [4] McKeever, C. (2007). *A Cost Comparison—The Real Cost of Mutual Funds v ETF's*, Chance Favors, Retrieved on July 28, 2008 from <http://chancefavors.com/2007/10/cost-comparison-mutual-funds-vs-etfs/>
- [5] McWhinney, J. (2005). *An Inside Look at ETF Construction*, Investopedia, Retrieved on July 25, 2008 from <http://www.investopedia.com/articles/mutualfund/05/062705.asp>

MICHAEL J. TARI

Volume-weighted Average Price (VWAP)

The volume-weighted average price (VWAP) and its close cousin, the time-weighted average price (TWAP), are commonly used measures of the average price of a security over a period of time. VWAP and TWAP are used by traders and other investment professionals as reference prices, an indication of the average transaction price over an interval of time. So, for example, if the TWAP of a security is \$10 on a given day and a trader had bought a sizeable block of shares at \$9.50, we might conclude that the trader had added value in that he or she obtained a better than a naïve program that mechanically sends out orders in the market at a steady rate throughout the day.

Mathematical Definition

More formally, the VWAP of a security over a specified trading horizon (e.g., from market open to close) is defined as the ratio of the total transaction value in that security (i.e., the sum, over all trades in the specified horizon, of the product of each trade's share volume and the corresponding price) to the total volume of shares traded (i.e., the sum of all shares traded in the trading horizon). When the trading horizon is typically a trading day, intraday or multiday VWAP measures are also computed. A related concept is the TWAP, defined as the average price over a particular time interval with no explicit volume weighting. Traders use TWAP over VWAP for securities where the temporal pattern of volume exhibits considerable variation, for example, in less-active securities.

Formally, given N trades in the relevant interval, let $S_1, \dots, S_N$ be the shares transacted with corresponding prices $P_1, \dots, P_N$. Then, we have

$$VWAP = \frac{\sum_{i=1}^N P_i S_i}{\sum_{i=1}^N S_i} \quad (1)$$

$$TWAP = \frac{\sum_{i=1}^N P_i}{N} \quad (2)$$

Subtleties in the computation of VWAP/TWAP include (i) the choice of volume definition (e.g., primary market volume or composite volume), (ii) the treatment of certain trades (e.g., block trades that might be negotiated off market), and (c) the decision whether to include volumes at the open and close of the market.

Uses

VWAP is commonly used as an approximation to the price that could be realized by a trader who passively participates in trading activity. As such, the performance of traders can be measured by their ability to execute orders at prices better than the VWAP benchmark prevailing over the trading horizon.

The computational simplicity of the VWAP is a major factor in its popularity in measuring trade execution, especially in markets where detailed trade level data is difficult or expensive to obtain. VWAP can be misleading as a benchmark in certain situations where the trader's objective is to control the slippage from a given strike or decision price, or where the strategy is not passive. In such cases, for example, if the trader has short-term alpha, the mechanical application of a VWAP strategy (i.e., trading in parallel to historical volume patterns) can lead to significant opportunity costs in terms of slippage. VWAP is not appropriate when the trader's executions are large relative to market volumes. In this case, VWAP might conceal a large price impact because the trader's own trades constitute the bulk of the reported volume. Finally, if traders have discretion over whether to execute or not, the VWAP benchmark can be gamed by selectively timing executions.

An important application is to so-called VWAP strategies, typically algorithmic trading strategies that automatically break up an order and send trades to the market to match the historical volume pattern or profile (see, e.g., [1]) of a security. See, for example, [2] for a discussion of the uses of VWAP in trading strategies and algorithms. The goal of a VWAP strategy is to obtain an execution price close to the VWAP for the day. Some brokers also guarantee VWAP execution, essentially taking on the execution risk for a fee.

References

- [1] Hobson, D. (2006). VWAP and volume profiles, *Journal of Trading*, 1(2), Spring, 38–42.
- [2] Madhavan, A. (2002). VWAP Strategies, in *Transaction Performance: The Changing Face of Trading*, Handbook Series in Finance, B. Bruce, ed, Institutional Investor Inc.

Related Articles

Automated Trading; Execution Costs; Price Impact.

ANANTH N. MADHAVAN

Equity Swaps

A swap contract is a bilateral agreement between two parties, known as counterparties, to exchange cash flows at regularly scheduled dates in the future. In an equity swap, some of the cash flows are determined by the return on a stock or an equity index. Typically, one of the parties pays to the other the total return of a stock or an equity index. In exchange, he or she receives from the other a cash flow determined by a fixed or floating rate or the return of another stock or equity index. Equity swaps are also known as *equity-linked swaps* and *equity-indexed swaps*.

Equity swaps are not traded on an exchange but are privately negotiated. They are referred to as over-the-counter (OTC) contracts. One of the first-known equity swap agreements was offered by the Bankers Trust in 1989. Since then, the market for equity swaps and other equity-linked derivatives has grown rapidly. There are no exact figures on the size of the market. However, the Bank for International Settlements (BIS) provides market size estimates. According to BIS, the estimate of the worldwide total notional amounts outstanding of equity swaps and equity forwards was over \$300 trillion as of December 2007.

Equity swaps provide means to get exposure to the underlying stock or index without making a direct investment. Because equity swaps are OTC-contracts, they can be tailor-made to specific needs. The contracts have been used to circumvent barriers for direct investments in particular markets, bypass various taxes, and minimize transaction costs.

Defining Equity Swaps

Let $\{T_0, T_1, \dots, T_M\}$ be a sequence of dates. This is the *tenor structure* and we denote it by $\mathcal{T}$. For a given day-count convention, we specify a sequence of year fractions^a $\{\delta_1, \delta_2, \dots, \delta_M\}$ to $\mathcal{T}$. We denote the counterparties by A and B .

Definition 1 A generic equity swap

An equity swap with tenor structure $\mathcal{T}$ is a contract that starts at time T_0 and has payment dates $T_1, T_2, \dots, T_M$. At each payment date T_i for $i = 1, 2, \dots, M$, the two counterparties A and B exchange payments. At least one of the payments will be based

on the return of a stock or an equity index over the period $[T_{i-1}, T_i]$.

In general, the cash flows are specified in such a way that the initial value, at time T_0 , of the swap equals zero. Usually the equity swap pays out the total return of the underlying stock or equity index including dividends. However, there are also variants where the dividend is excluded.

A swap contract has a notional principal.^b It is a currency amount specified in the swap contract that determines the size of the payments expressed in currency units. While the notional principal of a bond, for instance, is paid out at maturity, the notional principal of a swap contract is, in general, never exchanged. Equity swaps can be classified into two categories depending on whether the notional principal is constant or varies over the lifetime of the swap. We focus on the former case, which is considered in the next section.

Contracts with Fixed Notional Principal

Let N denote the fixed notional principal. Let $\{Z(t)\}$ denote the price process of a stock or an equity index. Define the *period return* $R(T_i, T_{i+1})$ over the interval $[T_i, T_{i+1}]$ for asset Z by

$$R(T_i, T_{i+1}) = \frac{Z(T_{i+1})}{Z(T_i)} - 1 \quad (1)$$

Definition 2 A generic equity-for-fixed-rate swap

An equity-for-fixed-rate swap, with tenor structure $\mathcal{T}$, which is written on the equity Z will have a predetermined swap rate K and will give rise to the following payments between the counterparties A and B at each payment date T_i :

- A pays to B the amount: $NR(T_{i-1}, T_i)$.
- B pays to A the amount: $N\delta_i K$.

In general, the swap rate is chosen such that the initial value of the swap at time T_0 equals zero.

In its most simple form, this contract is referred to as a *plain vanilla equity swap*. The period return is then determined by a domestic asset or index and the nominal amount is expressed in units of the domestic currency. Examples can be found in [5] and [8].

Some equity swaps are structured so that instead of a fixed swap rate they pay a floating interest rate,

usually a LIBOR rate. Let $L(T_i, T_{i+1})$ denote the simple spot rate over the period $[T_i, T_{i+1}]$.

Definition 3 A generic equity-for-floating-rate swap

An equity-for-floating-rate swap, with tenor structure T , which is written on the equity Z will give rise to the following payments between the counterparties A and B at each payment date T_i :

- *A pays to B the amount: $NR(T_{i-1}, T_i)$.*
- *B pays to A the amount: $N\delta_i (L(T_{i-1}, T_i) + s)$.*

where s is a constant rate such that the initial value of the swap at time T_0 equals zero.

An equity-for-floating swap can be decomposed into an equity-for-fixed swap and a suitably chosen interest rate swap (see **LIBOR Rate** and [2]).

Let R_1 and R_2 denote the return of assets Z_1 and Z_2 , respectively.

Definition 4 A generic equity-for-equity swap

An equity-for-equity swap, with tenor structure T , which is written on the equities Z_1 and Z_2 , will give rise to the following payments between the counterparties A and B at each payment date T_i :

- *A pays to B the amount: $NR_1(T_{i-1}, T_i)$.*
- *B pays to A the amount: $N(R_2(T_{i-1}, T_i) + s\delta_i)$.*

where s is the constant rate such that the initial value of the swap at time T_0 equals zero.

The equity-for-equity swap is also referred to as a *two-way equity swap*. The simplest contract of this type is a domestic equity-for-equity swap where both returns are based on domestic indices or assets.

So far, we have only considered domestic equity indices and assets. However, all of the three equity swaps mentioned above have versions where one or both cash flows are based on a foreign equity return or interest rate. They are so called *cross-currency swaps*.

To illustrate a cross-currency equity swap, suppose that the United States is the domestic market. Let the notional principal be expressed in US dollars. Let Z_1 be a foreign equity index such as, for instance, the NIKKEI, while Z_2 is a domestic equity index such as the S&P 500. The period return R_1 is based on a foreign equity index, while the nominal amount is in domestic units. There is a currency mismatch in the cash flow that A pays (but none in the cash

flow that B pays). This type of contract is referred to as a *quanto swap* (see also **Quanto Options**). Quanto swaps are more complicated to price than other swaps. Quanto contracts have been considered in [3, 4] and [7].

From a pricing and hedging perspective, the simplest cross-currency swaps are the ones that are currency adjusted. Consider a cross-currency equity-for-equity swap with currency-adjusted returns. Let Z_1 be a foreign equity, while Z_2 is a domestic equity. Let $X(t)$ denote the exchange rate expressed as the number of domestic currency units per foreign currency unit. Then the currency-adjusted period return over the interval $[T_i, T_{i+1}]$ for the asset Z_1 is

$$R_1(T_i, T_{i+1}) = \frac{X(T_{i+1})Z_1(T_{i+1})}{X(T_i)Z_1(T_i)} - 1 \quad (2)$$

While the unit of $Z_1(t)$ is foreign currency, the unit of $Z_1(t)X(t)$ is domestic currency. Regarding the underlying index as the foreign asset times the exchange rate, R_1 can be treated as the return on a domestic index. A cross-currency equity-for-equity swap that is currency adjusted is, from a valuation point of view, equal to a domestic equity-for-equity swap.

Contracts with Variable Notional Principal

Some equity swaps are constructed with a variable notional principal. A variable notional principal changes over time according to changes in the referenced equity index.

Consider an equity-for-fixed-rate swap. It can essentially be regarded as a leveraged position in the underlying equity. If the notional principal is constant, the realized returns from the equity index are withdrawn in each period, resulting in a position that is rebalanced periodically. If the notional principal is variable, the realized returns in each period are reinvested.

Let N_i denote the variable notional principal, which determines the size of the payments at time T_i for $i = 1, 2, \dots, M$. Let $N_1 = 1$ and $N_i = Z(T_{i-1})/Z(T_0)$ for $i = 2, 3, \dots, M$. Thus, for instance, at the third payment date T_3

- *A pays to B the amount: $\frac{Z(T_2)}{Z(T_0)} R(T_2, T_3)$.*
- *B pays to A the amount: $\frac{Z(T_2)}{Z(T_0)} \delta_3 K$.*

Equity swaps with variable notional principals are treated in [2, 6] and [9].

More Equity Swaps & Strategies

There can be many variations of the equity swaps listed so far. For instance, there can be more than one tenor structure, that is, the payments made by A and B can have different periodicity. It is also possible to make forward agreements to enter into a swap contract in the future. Such contracts are known as *forward swaps* or *deferred swaps*. There are also equity swaps with option features like *capped equity swaps* and *barrier equity swaps*. Further examples are *blended index swaps* and *outperformance swaps* (see [2, 6] and [8]).

We conclude by providing an example of how equity swaps were used in the United States during the 1990s to circumvent taxes. The executive equity swap strategies were developed for large single stock shareholders, for instance, a founder of a company. The swap was constructed so that the shareholder made payments based on the return of the stock to the swap contractor. In exchange, the shareholder received either a fixed interest rate or the return of a large equity index such as S&P 500. By entering into such a contract, the stockholder could keep the stocks and the voting rights, but still reduce the risk of the total portfolio and avoid capital gains taxes. As a result, the tax regulation was changed. The new regulation states that taxpayers should recognize that transactions that are essentially equivalent to a sale should be treated as such and thus be taxed. For a more detailed description on this topic, see [1] and [8].

End Notes

^aFor instance, if the convention “Actual/365” is used, δ_1 is equal to the number of days between the dates T_0 and T_1 , divided by 365.

^bSometimes called *face value*.

References

- [1] Bolster, P., Chance, D. & Rich, D. (1996). Executive equity swaps and corporate insider holdings, *Financial Management* **25**(2), 14–24.
- [2] Chance, D. & Rich, D. (1998). The pricing of equity swaps and swaptions, *The Journal of Derivatives* **5**, 19–31.
- [3] Chung, S. & Yang, H. (2005). Pricing quanto equity swaps in a stochastic interest rate economy, *Applied Mathematical Finance* **12**(2), 121–146.
- [4] Hinnerich, M. (2007). *Derivatives Pricing and Term Structure Modeling*. PhD Thesis, Stockholm School of Economics, EFI, The Economic Research Institute, Stockholm.
- [5] Jarrow, R. & Turnbull, S. (1996). *Derivative Securities*, South-Western Publishing, Cincinnati.
- [6] Kijima, M. & Muromachi, Y. (2001). Pricing equity swaps in a stochastic interest rate economy, *The Journal of Derivatives* **8**, 19–35.
- [7] Liao, M. & Wang, M. (2003). Pricing models of equity swaps, *The Journal of Futures Markets* **23**(8), 121–146, 751–772.
- [8] Marshall, J. & Yuyuenyongwatana, R. (2000). Equity swaps: structures, uses, and pricing, In *Handbook of Equity Derivatives*, Jack C. Francis, William W. Toy, & J.G. Whittaker, eds, Wiley, New York.
- [9] Wu, T. & Chen, S. (2007). Equity swaps in a labor market model, *The Journal of Futures Markets* **27**(9), 893–920.

Further Reading

Chance, D. (2004). Equity swaps and equity investing, *The Journal of Alternative Investing* **7**, 75–97.

Related Articles

Equity Default Swaps; Forwards and Futures; LIBOR Rate; Quanto Options; Total Return Swap.

MIA HINNERICH

Volatility Index Options

Volatility index options are options on a volatility index. Volatility index options enable one to take a pure volatility exposure without the need to take positions in the index itself and without the need to delta-hedge. These features make these options very interesting for pure volatility trades and bets. They also allow one to trade the spread between realized and implied volatility, or to hedge the volatility exposure of other positions or businesses, without being contaminated by the index price dependence like in the standard index options. This explains their popularity among traders and hedge funds.

Originally, volatility index options were traded over the counter (OTC), and a large part is still traded OTC, through volatility and variance swaps (see **Variance Swap**). In February 2006, the Chicago board option exchange (CBOE) realized the interest in the financial community for exchanged standardized volatility index options and started offering standardized options first on the VIX (which is the CBOE volatility index) and later on the Russel 2000 volatility index. In Europe, although the major exchanges have already developed volatility on the major indexes (and futures on these) like the VDAX, FTSE 100 volatility index, VSTOXX, VCAC, or VSML, they still do not offer any options on these indexes.

Volatility Index

The underlying volatility index is computed from option prices to capture information from the options market by some means. The overall idea of index volatility is to provide a good estimate of the so-called implied volatility extracted from the options market, as opposed to the historical volatility. More precisely, the aim is to estimate the risk-neutral market's expectation of the future volatility. In the specific case of the VIX, the formula is public and weights the various options as follows:

$$\sigma^2 = \sum_i \frac{2\Delta K_i}{T_i K_i^2} e^{RT} Q(K_i) - \frac{1}{T_i} \left(\frac{F}{K_0} - 1 \right)^2 \quad (1)$$

where σ is the $VIX/100$, or equivalently, $VIX = 100\sigma$, T_i the time to expiration of the i th option, F

the forward index level derived from option prices, K_0 the highest strike just below the forward index level, F , K_i the strike price of the i th option: a call if $K_i > K_0$, a put if $K_i < K_0$ and both call and put for $K_i = K_0$, ΔK_i the interval between strike prices, R the risk-free rate to expiration, and $Q(K_i)$ the midpoint of the bid-ask spread for each options with strike K_i .

The CBOE calculates and publishes minute-to-minute the VIX using real-time bid-ask market quotes of options on S & P500 index (SPX) with nearby and second nearby maturities and applying the multiplier of \$100. Overall, the VIX reflects market's view of the future short term volatility. A high value of the index indicates a more volatile market, while a low value indicates a less volatile environment. Often referred to as the *fear index*, it represents one measure of the market's expectation of volatility over the next 30-day period. In short term, bias can explain some key differences between the VIX and the overall market sentiment. This is particularly true at times when the most liquid options are in the range of 2–6 months to expiration. Therefore, VIX tries to quantify the market volatility, mainly focusing on short term, being unable to completely explain the market volatility, which is a complex concept.

It is worth noting that VIX is computed to be the square root of the par variance swap rate (see **Variance Swap**), and not the volatility swap rate (see **Volatility Swaps**). This is because variance swap can be perfectly, statically replicated through vanilla puts and calls, whereas volatility swap requires dynamic hedging [1]. We will discuss this point later in the section on pricing.

VIX options are European call and put options on the VIX index, with strikes ranging from 10 to 65 (with interval of 1 and 2.5 points for liquid and 5 points for less liquid points), while maturities are up to 6 months. Like many other listed options, VIX options are quoted with a multiplier of \$100. The expiration date is roughly speaking the third Wednesday of the expiry month. More details can be found on the CBOE website [3].

Pricing Models

In terms of pricing, roughly speaking there are two methods to price index volatility options:

- a model-dependent approach that assumes a model for the index volatility diffusion and provides a closed-form formula of call and put options;
- a model-free approach that computes the cost of the static hedge to replicate the volatility index option.

Historically, the model-dependent approach has been the first to emerge. Successively, among others, Whaley [11], Grunbichler and Longstaff [6], Howison *et al.* [7], and Elliott *et al.* [5], and lately Sepp [9, 10], presented the model-dependent approaches to price volatility index options. They all assume an underlying stochastic process for the index volatility (or the index volatility futures) and explicitly compute the price of the call and put options.

The first model-dependent approach presented by Whaley [11] assumed a log normal diffusion for the VIX cash index and the VIX futures leading to a standard Black–Scholes formula for the VIX call options as follows:

$$C(T, F, K) = D_T [F N(d_+) - K N(d_-)] \quad (2)$$

$$d_{\pm} = \frac{\ln \frac{F}{K} \pm \frac{1}{2} T \sigma^2}{\sqrt{T} \sigma} \quad (3)$$

where $C(T, F, K)$ stands for the value of the call option with expiry time T , strike K , and forward

price F , and where σ is the volatility of the futures price, D_T is the discount factor expiring at time T , and $N(x)$ is the cumulative normal distribution up to x . A straightforward drawback of the Whaley approach was the strong log normal underlying assumption. This motivated further research and led to many works. Grunbichler and Longstaff [6] proposed a mean-reverting square root process for the volatility process. Following the popularity of stochastic volatility, Howison [7] and Elliott [5] suggested to use stochastic volatility model for the index volatility to capture the risk of volatility for the index volatility. Moreover, because it is well known that index volatility is upward sloping, the stochastic volatility approach, which can cope with this important feature, was an appealing modeling choice. Figure 1, for instance, gives the VIX smile for 27 July 2009.

Lately, Sepp [9, 10] argued in favor of adding jumps to the stochastic volatility model to get a more realistic diffusion for the index volatility. This was supported by the econometric works that confirmed the evidence of jumps for the volatility index.

To sum up, the model-dependent approach aims at modeling the index volatility evolution as accurately as possible and providing a very consistent framework for pricing any type of option on index volatility. The strength is the flexibility in terms of pricing as there is no limitation for the types of options. The

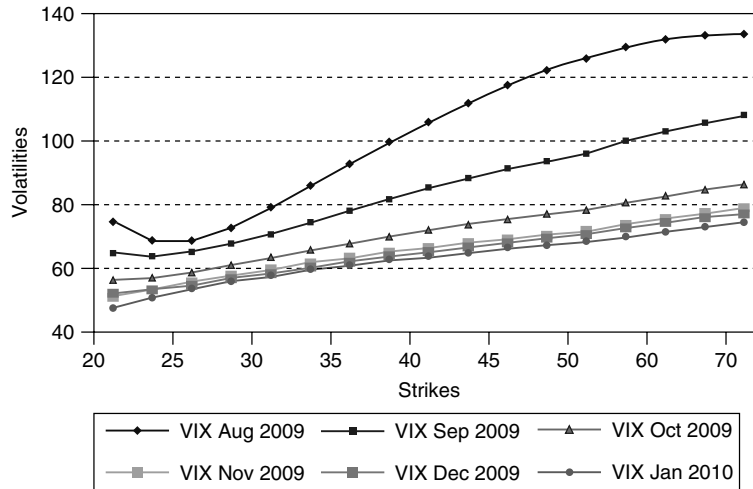


Figure 1 VIX Smile

weakness is the strong assumption of a specific model and distribution for the index volatility.

Another approach initiated by Neuberger [8], Demeter *et al.* [4], Carr and Lee [1], and Carr and Wu [2] is to exhibit a static hedge and compute in a model free way the price of this hedge using call and put options on the index itself. In a very insightful paper, Demeter *et al.* [4] showed that a static portfolio of call and puts options on the volatility index can replicate a variance swap. Lately, Carr and Lee [1] and Carr and Wu [2] extended the closed form formula to the case of both the variance and the volatility swap. The starting point is to assume a pure diffusion given as follows:

$$dS_t = \mu_t S_t dt + \sigma(t, \dots) S_t dW_t \quad (4)$$

where μ is the drift term and $\sigma(t, \dots)$ is a very general volatility function (that can be assumed to be a local volatility for the clarity of the explanation). A trivial application of the Itô lemma on $\log(S_t)$ provides the main intuition and leads to the fact that one can relate the volatility to a log contract as follows:

$$\frac{dS_t}{S_t} - d\log(S_t) = \frac{1}{2}\sigma^2(t, \dots) dt \quad (5)$$

Like any function of the underlying asset, the log contract can be replicated by a series of call and put options. This leads, in particular, to the Par swap rate of a variance swap as follows (referred in the literature as the replication variance swap price):

$$\begin{aligned} K_{\text{var}} = & \frac{2}{T} \left[rT - \left(\frac{S_0}{S_*} e^{rT} - 1 \right) - \log \left(\frac{S_*}{S_0} \right) \right. \\ & + e^{rT} \int_0^{S_*} \frac{1}{K^2} P(K) dK \\ & \left. + e^{rT} \int_{S_*}^{\infty} \frac{1}{K^2} C(K) dK \right] \quad (6) \end{aligned}$$

where $P(K)$ and $C(K)$, respectively, denote the current fair value of a put and call option of strike K , r is the risk free rate, T is the maturity of the variance swap, S_0 is the initial spot value of the underlying asset, and S_* is an arbitrary point to do the split between the liquid call and put options. It is often chosen to be the forward value.

Unlike the previous model-dependent approaches, the model-free approach replicates the variance and

the volatility swaps using market prices of volatility options as inputs. The resulting pricing consists in numerically computing the cost of the hedging strategy with the series of options as shown by the integrals of equation (6).

The obvious strength of this approach is to avoid any assumption on the underlying distribution of the index volatility. Primarily, the weaknesses are that the replication methods do not work for very specific index volatility options and that the discretization bias due to the lack of reliable liquid quotes for call and put options at any strikes can be of the same order of the magnitude as the misspecification of the index volatility distribution.

References

- [1] Carr, P. & Lee, R. (2007). Realized volatility and variance: Options via swaps, *Risk* May, 76–83.
- [2] Carr, P. & Wu, L. (2009). Variance risk premiums, *Review of Financial Studies* **22**, 1311–1341.
- [3] CBOE (2009). *Vix Options CBOE*. www.cboe.com.
- [4] Demeterfi, K., Derman, E., Kamal, M. & Zou, J. (1999). *More than You Ever wanted to know about Volatility Swaps*, Goldman Sachs Quantitative Strategies, March 1999.
- [5] Elliott, R., Siu, T. & Chan, L. (2004). Pricing volatility swaps under heston's stochastic volatility model with regime switching, *Applied Mathematical Finance* **14**(1), 41–62.
- [6] Grunbichler, A. & Longstaff, F. (1996). Valuing futures and options on volatility, *Journal of Banking and Finance* **20**, 985–1001.
- [7] Howison, S., Rafailidis, A. & Rasmussen, H. (2004). On the pricing and hedging of volatility derivatives, *Applied Mathematical Finance* **11**(4), 317–346.
- [8] Neuberger, A. (1994). The log contract: a new instrument to hedge volatility, *Journal of Portfolio Management* **20**(2), 74–80.
- [9] Sepp, A. (2008). Pricing options on realized variance in the heston model with jumps in returns and volatility, *Journal of Computational Finance* **11**(4), 33–70.
- [10] Sepp, A. (2008). Vix option pricing in a jump-diffusion model, *Risk* April, 84–89.
- [11] Whaley, R. (1993). Derivatives on market volatility: hedging tools long overdue, *Journal of Derivatives* **1**, 71–84.

Further Reading

Bergomi, L. (2008). Dynamic properties of smile models, in R. Cont ed. *Frontiers in Quantitative Finance: Volatility and Credit Risk Modeling*, Wiley, Chapter 3.

Cont, R. and Kokholm, T. (2009). *A Consistent Pricing Model for Index Options and Volatility Derivatives*. Available at SSRN: <http://ssrn.com/abstract=1474691>.

**Realized Volatility and Multipower Variation;
Realized Volatility Options; Stochastic Volatility
Models; Variance Swap; Weighted Variance Swap.**

Related Articles

ERIC BENHAMOU & MARIAN CIUCA

**Call Options; Corridor Variance Swap; Gamma
Swap; Heston Model; Implied Volatility Surface;**

Realized Volatility Options

Let the underlying process Y be a positive semi-martingale, and let $X_t := \log(Y_t/Y_0)$.

Define realized variance to be $[X]$, where $[\cdot]$ denotes the quadratic variation (but see the section “Contract Specifications in Practice”).

Define a *realized variance option* on Y with variance strike Q and expiry T to pay

$$\begin{aligned} ([X]_T - Q)^+ & \text{ for a realized variance call} \\ (Q - [X]_T)^+ & \text{ for a realized variance put} \end{aligned}$$

and define a *realized volatility option* on Y with volatility strike $Q^{1/2}$ and expiry T to pay

$$\begin{aligned} ([X]_T^{1/2} - Q^{1/2})^+ & \text{ for a realized volatility call} \\ (Q^{1/2} - [X]_T^{1/2})^+ & \text{ for a realized volatility put} \end{aligned}$$

In some places, we restrict attention to puts. Call prices follow by put–call parity: for realized variance options, a long-call short-put combination pays $[X]_T - Q$, equal to a Q -strike variance swap, and for realized volatility options, a long-call short-put combination pays $[X]_T^{1/2} - Q^{1/2}$, equal to a $Q^{1/2}$ -strike volatility swap.

Unlike variance swaps (*see* **Variance Swap; Weighted Variance Swap**), which admit exact model-free (assuming only continuity of Y) hedging and pricing in terms of Europeans, variance, and volatility options have a range of values, consistent with the given prices of Europeans. With no further assumptions, there exist sub/superreplication strategies and lower/upper pricing bounds (in the section “Pricing Bounds by Model-free Use of Europeans”). Under an independence condition, there exist exact pricing formulas in terms of Europeans (in the section “Pricing by Use of Europeans, Under an Independence Condition”). Under specific models, there exist exact pricing formulas in terms of model parameters (in the section “Pricing by Modeling the Underlying Process”).

Unless otherwise noted, all prices are denominated in units of a T -maturity discount bond. The results apply to dollar-denominated prices, provided that

interest rates vary deterministically, because if Y' is a dollar-denominated share price and Y is that share’s bond-denominated price, then $\log Y - \log Y'$ has finite variation; hence, $[\log Y] = [\log Y']$.

Expectations $\mathbb{E}$ will be with respect to martingale measure $\mathbb{P}$.

Transform Analysis

Some of the methods surveyed here (in the sections “Pricing by Modeling the Underlying Process” and “Pricing via Transform”) will price variance/volatility options by integrating prices of payoffs of the form $e^{z[X]_T}$. Transform analysis relates the former to the latter, by the following pricing formulas, proved in [5].

Assume that the continuous payoff function $h : \mathbb{R} \rightarrow \mathbb{R}$ satisfies

$$\int_{-\infty}^{\infty} e^{-\alpha q} h(q) dq < \infty \quad (1)$$

for some $\alpha \in \mathbb{R}$. For all $z \in \alpha + \mathbb{R}i := \{z \in \mathbb{C} : \operatorname{Re} z = \alpha\}$, define the bilateral Laplace transform

$$H(z) := \int_{-\infty}^{\infty} e^{-zq} h(q) dq \quad (2)$$

If $|H|$ is integrable along $\alpha + \mathbb{R}i$ for some $\alpha \leq 0$, then by Bromwich and Fubini, the $h([X]_T)$ payoff has price

$$\mathbb{E}h([X]_T) = \frac{1}{2\pi i} \int_{\alpha - \infty i}^{\alpha + \infty i} H(z) \mathbb{E}e^{z[X]_T} dz \quad (3)$$

For a *variance put*, let $h(q) = (Q - q)^+$. Then for all $\alpha < 0$, formula (3) holds with

$$H(z) = \frac{e^{-Qz}}{z^2} \quad (4)$$

For a *volatility put*, let $h(q) = (\sqrt{Q} - \sqrt{q^+})^+$. Then for all $\alpha < 0$, formula (3) holds with

$$H(z) = -\frac{\sqrt{\pi} \operatorname{Erf}(\sqrt{zQ})}{2z^{3/2}} \quad (5)$$

To price variance and volatility calls by put–call parity, we have the variance swap value

$$\mathbb{E}[X]_T = \frac{\partial}{\partial z} \bigg|_{z=0} \mathbb{E}e^{z[X]_T} \quad (6)$$

and the volatility swap value

$$\mathbb{E}[X]_T^{1/2} = \frac{1}{2\sqrt{\pi}} \int_0^\infty \frac{1 - \mathbb{E}e^{-z[X]_T}}{z^{3/2}} dz \quad (7)$$

if $\mathbb{E}e^{z[X]_T}$ is analytic in a neighborhood of $z = 0$.

Pricing by Modeling the Underlying Process

Under Heston and under Lévy models, we give formulas for the transform $\mathbb{E}e^{z[X]_T}$, where $\text{Re } z \leq 0$. Hence, formula (3) prices the variance put and volatility put, using equations (4) and (5), respectively.

Example: Heston Dynamics

Under the Heston model for instantaneous variance (see **Heston Model**),

$$dV_t = (a - \kappa V_t) dt + \beta \sqrt{V_t} dW_t \quad (8)$$

and the transform of $[X]_T = \int_0^T V_t dt$ is

$$\mathbb{E}e^{z[X]_T} = e^{A(z) + B(z)V_0} \quad (9)$$

where

$$A(z) := \frac{a}{\beta^2} \left[(\kappa - \gamma)T - 2 \log \left(1 + \frac{\kappa - \gamma}{2\gamma} (1 - e^{-\gamma T}) \right) \right] \quad (10)$$

$$B(z) := \frac{2z(e^{\gamma T} - 1)}{2\gamma + (\gamma + \kappa)(e^{\gamma T} - 1)},$$

$$\gamma := \sqrt{\kappa^2 - 2\beta^2 z} \quad (11)$$

by [6]. Other affine models also have explicit formulas for $\mathbb{E}e^{z[X]_T}$.

Example: Lévy Dynamics

If X is a Lévy process (see **Lévy Processes**) with Gaussian variance σ^2 and Lévy measure ν , then $[X]$ has transform

$$\mathbb{E}e^{z[X]_T} = \exp \left(T \frac{\sigma^2 z^2}{2} + T \int_{\mathbb{R}} (e^{zx^2} - 1) \nu(dx) \right) \quad (12)$$

For variance option pricing under pure-jump processes with independent increments, but without assuming stationary increments, see [2].

Pricing by Use of Europeans, Under an Independence Condition

In this section, let Y be a share price that follows general stochastic volatility dynamics

$$dY_t = \sigma_t Y_t dW_t \quad (13)$$

where σ and the Brownian motion W are independent. Although all three subsections use this assumption, the schemes in the sections “Pricing via Transform” and “Pricing and Hedging via Uniform or L^2 Payoff Approximation” are immunized, to first order, against violations of the independence condition.

Pricing via Transform

The transform of $[X]_T = \int_0^T \sigma_t^2 dt$ satisfies [5]

$$\mathbb{E}e^{z[X]_T} = \mathbb{E} \left(\theta_+ (Y_T/Y_0)^{1/2 + \sqrt{(1/4) + 2z}} + \theta_- (Y_T/Y_0)^{1/2 - \sqrt{(1/4) + 2z}} \right) \quad (14)$$

provided that the expectations are finite. Here, $\theta_{\pm} := (1 \mp 1/\sqrt{1 + 8z})/2$. The right-hand side (RHS) of equation (14) is in principle observable from T -expiry Europeans, which allows variance/volatility put option pricing by the formulas (3–5). In this context, equation (6) can be replaced by the log-contract value $-2\mathbb{E}X_T$, and equation (7) can be replaced by the synthetic volatility swap value (see **Volatility Swaps**).

Moreover, source [5] shows that equation (14) still holds approximately in the presence of correlation between σ and W , in the sense that the RHS is constructed to have zero sensitivity to first-order correlation effects.

Pricing and Hedging via Uniform or L^2 Payoff Approximation

For continuous payoffs, $h : [0, \infty) \rightarrow \mathbb{R}$ with finite limit at ∞ , such as the variance put or volatility put,

consider an n th-order approximation to $h(q)$

$$A_n(q) := a_{n,n}e^{-cnq} + a_{n,n-1}e^{-c(n-1)q} + \dots + a_{n,0} \quad (15)$$

where $c > 0$ is an arbitrary constant.

To choose A by uniform approximation, $a_{n,k}$ may be determined as the coefficients of the n th Bernstein polynomial approximation to the function $x \mapsto h(-(1/c) \log x)$ on $[0, 1]$.

Then source [5] shows that

$$\begin{aligned} \mathbb{E}h([X]_T) &= \lim_{n \rightarrow \infty} \mathbb{E} \sum_{k=0}^n a_{n,k} \left(\theta_+(Y_T/Y_0)^{1/2 + \sqrt{1/4 - 2ck}} \right. \\ &\quad \left. + \theta_-(Y_T/Y_0)^{1/2 - \sqrt{1/4 - 2ck}} \right) \end{aligned} \quad (16)$$

where $\theta_{\pm} := (1 \mp 1/\sqrt{1 - 8ck})/2$. The RHS of equation (16) is, in principle, observable from T -expiry Europeans and is moreover designed to have zero sensitivity to first-order correlation effects.

Alternatively, to choose A by L^2 approximation, the $a_{n,k}$ may be determined by $L^2(\mu)$ projection of h onto $\text{span}\{1, e^{-cq}, \dots, e^{-cnq}\}$, where the “prior” μ is a finite measure on $[0, \infty)$. In practice, $a_{n,k}$ may be computed by weighted least squares regression of $h(q)$ on the regressors $\{q \mapsto e^{-ckq} : k = 0, \dots, n\}$, with weights given by μ . Then source [5] shows that equation (16) still holds, regardless of the choice of the prior μ , provided that $dP/d\mu$ exists in $L^2(\mu)$, where P denotes the $\mathbb{P}$ -distribution of $[X]_T$.

For *hedging* purposes, the summation in the RHS of equation (16) provides a European-style payoff that, in conjunction with share trading, replicates the volatility payoff $h([X]_T)$ to arbitrary accuracy.

Pricing via Variance Distribution Inference

Given the prices $\mathbf{c} \in \mathbb{R}^{N \times 1}$ of vanilla options at strikes $K_1, \dots, K_N$, a scheme in [8] discretizes into $\{v_1, \dots, v_J\}$ the possible values of $[X]_T$, and proposes to infer the discretized variance distribution $\mathbf{p} \in \mathbb{R}^{J \times 1}$ where $p_j := \mathbb{P}([X]_T = v_j)$, by solving approximately for $\mathbf{p}$ in

$$\mathbf{B}\mathbf{p} = \mathbf{c} \quad (17)$$

where $\mathbf{B} \in \mathbb{R}^{N \times J}$ is given by $B_{nj} := C^{BS}(K_n, v_j)$, the Black–Scholes formula for strike K_n and squared unannualized volatility v_j . The approximate solution is chosen to minimize $\|\mathbf{B}\mathbf{p} - \mathbf{c}\|^2$ plus a convex penalty term. The contact paying $h([X]_T)$ is then priced as $\sum p_j h(v_j)$.

Pricing by Use of Variance or Volatility Swaps

With sufficient liquidity, variance and/or volatility swap quotes can be taken as inputs. For example, an approximation in [8] prices variance options by fitting a lognormal variance distribution to variance and volatility swaps of the same expiry. An approximation in [4] prices and hedges variance and volatility options by fitting a displaced lognormal, to variance and volatility swaps.

The variance curve models in [1] apply a different approach to using variance swaps; they take as inputs the variance swap quotes at multiple expiries, and they model the dynamics of the term structure of forward variance. Applications include pricing and hedging of realized variance options.

Pricing Bounds by Model-free Use of Europeans

In this section, consider variance options on, more generally, any continuous share price Y .

Given European options of the same expiry T , there exist model-free sub/superreplication strategies, and hence lower/upper pricing bounds, for the variance options. Here *model-free* means that, aside from continuity and positivity, we make *no* assumptions on Y .

Subreplication and Lower Bounds

The following subreplication strategy is due to [7]; this exposition also draws from [3].

Let $\lambda : (0, \infty) \rightarrow \mathbb{R}$ be convex, let λ_y denote its left-hand derivative, and assume that its second derivative in the distributional sense has a density, denoted λ_{yy} , which satisfies for all $y \in \mathbb{R}_+$

$$\lambda_{yy}(y) \leq 2/y^2 \quad (18)$$

Define for $y > 0$ and $v > 0$

$$BS(y, v; \lambda) := \int_{-\infty}^{\infty} \lambda(ye^z) \frac{1}{\sqrt{2\pi}v} e^{-(z+v/2)^2/(2v)} dz \quad (19)$$

and define $BS(y, 0; \lambda) := \lambda(y)$, and let BS_y denote its y -derivative. Let $\tau_Q := \inf\{t \geq 0 : [X]_t \geq Q\}$. Then the following trading strategy subreplicates the variance call payoff: hold statically a claim that pays at time T

$$\lambda(Y_T) - BS(Y_0, Q; \lambda) \quad (20)$$

and trade shares dynamically, holding at each time $t \in (0, T)$

$$\begin{aligned} -BS_y(Y_t, Q - [X]_t; \lambda) & \quad \text{shares if } t \leq \tau_Q \\ -\lambda_y(Y_t) & \quad \text{shares if } t > \tau_Q \end{aligned} \quad (21)$$

and a bond position that finances the shares and accumulates the trading gains or losses. Therefore, the time-0 value of the contract paying (20) provides a lower bound on the variance call value.

The lower bound from equation (20) is optimized by λ consisting of $2/K^2 dK$ out-of-the-money vanilla payoffs at all K where $I_0(K, T)$, the squared unannualized Black–Scholes implied volatility, exceeds Q :

$$\lambda(y) = \int_{\{K: I_0(K, T) > Q\}} \frac{2}{K^2} \text{van}_K(y) dK \quad (22)$$

See [3] for generalization to forward-starting variance options.

Superreplication and Upper Bounds

The following superreplication strategy is due to [3]. Choose any $b_d \in (0, Y_0]$ and $b_u \in [Y_0, \infty)$. Let

$$\begin{aligned} BP(y, q) & \\ := \int_{-\infty - \alpha i}^{\infty - \alpha i} & \left[\sqrt{y/b_u} \sinh(\log(b_d/y) f(z)) \right. \\ & \left. - \sqrt{y/b_d} \sinh(\log(b_u/y) f(z)) \right] \\ & \left/ \left[2\pi z^2 e^{i(Q-q)z} \sinh(\log(b_u/b_d) f(z)) \right] \right. dz \end{aligned} \quad (23)$$

where $f(z) := \sqrt{1/4 - 2iz}$ and where $\alpha > 0$ is arbitrary. For $y > 0$ and $b_d \neq b_u$, define

$$\begin{aligned} L(y; b_d, b_u) & \\ := -2 \log(y/b_u) + 2 \frac{\log(b_u/b_d)}{b_u - b_d} (y - b_u) \end{aligned} \quad (24)$$

and define $L(y; Y_0, Y_0) := -2 \log(y/Y_0) + 2y/Y_0 - 2$. Let

$$L^*(y) := \begin{cases} L(y) & \text{if } y \notin (b_d, b_u) \\ -BP(y, 0) & \text{if } y \in (b_d, b_u) \end{cases} \quad (25)$$

Let BP_y and L_y denote the y -derivatives, and let $\tau_b := \inf\{t \geq 0 : Y_t \notin (b_d, b_u)\}$.

Then, the following strategy superreplicates the variance call payoff $([X]_T - Q)^+$. Hold statically a claim that pays at time T

$$L^*(Y_T) - L^*(Y_0) \quad (26)$$

and trade shares dynamically, holding at each time $t \in (0, T)$

$$\begin{aligned} BP_y(Y_t, [X]_t - [X]_0) & \quad \text{shares if } 0 \leq t \leq \tau_b \\ -L_y(Y_t) & \quad \text{shares if } t > \tau_b \end{aligned} \quad (27)$$

and a bond position that finances the shares and accumulates the trading gains or losses.

Therefore, the time-0 value of the contract paying (26) provides an upper bound on the variance call value. Given T -expiry European options data, the upper bound from equation (26) may be optimized over all choices of (b_d, b_u) .

Connection to the Skorokhod Problem

Whereas the sections “Subreplication and Lower Bounds” and “Superreplication and Upper Bounds” presented explicit hedging strategies, which imply pricing bounds, this section presents (a logarithmic version of) the result in [7], which showed that stopping-time analysis also implies pricing bounds.

Denote by ν the $\mathbb{P}$ -distribution of Y_T , which is revealed by the prices of T -expiry options on Y .

Suppose that $\tilde{Y}$ is a continuous $\mathcal{F}$ -martingale with $\tilde{Y}_T \sim \nu$, and $[\tilde{X}]_T$ has finite expectation, where $\tilde{X} := \log \tilde{Y}$. Then Dambis–Dubins–Schwartz implies that

$\tilde{Y}_t = G_{[\tilde{X}]_t}$, where G is a driftless unit-volatility geometric $\mathcal{G}$ -Brownian motion (on an enlarged probability space if needed) with $G_0 = Y_0$, and $[\tilde{X}]_t$ are $\mathcal{G}$ -stopping times, where $\mathcal{G}_s := \mathcal{F}_{\inf\{t: [\tilde{X}]_t > s\}}$. Thus $G_{[\tilde{X}]_T} \sim \nu$; and hence $[\tilde{X}]_T$ solves a *Skorokhod problem* (see **Skorokhod Embedding**): it is a finite-expectation stopping time that embeds the distribution ν in G . Conversely, if some finite-expectation τ embeds ν in a driftless unit-volatility geometric Brownian motion G , then $\tilde{Y}_t := G_{\tau \wedge (t/(T-t))}$ defines a continuous martingale with $\tilde{Y}_T \sim \nu$ and $[\log \tilde{Y}]_T = \tau$.

Therefore, distributions of stopping times solving the Skorokhod problem are identical to distributions of realized variance consistent with the given price distribution ν . Skorokhod solutions that have optimality properties, therefore, imply bounds on prices of variance/volatility options. In particular, *Root's* solution is known [9] to minimize the expectations of convex functions of the stopping time; the minimized expectation is, in that sense, a sharp lower bound on the price of a variance option (see also **Skorokhod Embedding**).

Contract Specifications in Practice

In practice, the realized variance in the payoff specification is defined by replacing quadratic variation $[X]_T$ with an annualized discretization that monitors Y , typically daily, for N periods, resulting in a specification

$$\text{Annualization} \times \sum_{n=1}^N \left(\log \frac{Y_n}{Y_{n-1}} \right)^2 \quad (28)$$

If the contract adjusts for dividends (as typical for single-stock dividends but not index dividends) then

the term inside the parentheses becomes $\log((Y_n + D_n)/Y_{n-1})$, where D_n denotes the discrete dividend payment, if any, of the n th period.

References

- [1] Buehler, H. (2006). Consistent variance curve models, *Finance and Stochastics* **10**(2), 178–203.
- [2] Carr, P., Geman, H., Madan, D. & Yor, M. (2005). Pricing options on realized variance, *Finance and Stochastics* **9**(4), 453–475.
- [3] Carr, P. & Lee, R. Hedging variance options on continuous semimartingales, *Finance and Stochastics*, forthcoming.
- [4] Carr, P. & Lee, R. (2007). Realized volatility and variance: options via swaps, *Risk* **20**(5), 76–83.
- [5] Carr, P. & Lee, R. (2008). *Robust Replication of Volatility Derivatives*, Bloomberg LP and University of Chicago.
- [6] Cox, J., Ingersoll, J. & Ross, S. (1985). A theory of the term structure of interest rates, *Econometrica* **53**(2), 385–407.
- [7] Dupire, B. (2005). *Volatility Derivatives Modeling*, Bloomberg LP.
- [8] Friz, P. & Gatheral, J. (2005). Valuation of volatility derivatives as an inverse problem, *Quantitative Finance* **5**(6), 531–542.
- [9] Rost, H. (1976). Skorokhod stopping times of minimal variance, *Séminaire de Probabilités (Strasbourg)*, Springer-Verlag, Vol. 10, pp. 194–208.

Related Articles

Exponential Lévy Models; Heston Model; Lévy Processes; Skorokhod Embedding; Variance Swap; Volatility Swaps; Volatility Index Options; Weighted Variance Swap.

ROGER LEE

Put–Call Parity

Put–call parity means that one may switch between call and put positions by selling or buying the underlying forward: “long call, short put is long forward contract” or $\mathbf{c} - \mathbf{p} \equiv \mathbf{f}$. In other words, one may replicate a put contract by buying a call of identical characteristics (underlying asset, strike, maturity) and selling the underlying asset forward ($\mathbf{p} \equiv \mathbf{c} - \mathbf{f}$), and one may replicate a call by buying a put and the underlying forward ($\mathbf{c} \equiv \mathbf{p} + \mathbf{f}$). This is shown in the three payoff diagrams (Figures 1–3).

A logical proof of the third instance ($\mathbf{c} \equiv \mathbf{p} + \mathbf{f}$) is as follows: a rational investor will exercise a call option whenever the asset price S at maturity is above the strike K ; this is equivalent to promising to buy the asset at K and having the option to sell it at that level, which a rational investor will exercise whenever S falls below K .

Put–call parity is often referred to as *option synthetics* by practitioners and holds only for European options.^a It does not require any assumption other than the ability to buy or sell the asset forward, but it is worth noting that this may not always be the case: to sell forward, either a futures market must exist or one must be able to short-sell the asset.

Put–call parity must not be confused with “put–call symmetry” (see **Foreign Exchange Symmetries**) in foreign exchange, which states that a call struck at K on a given exchange rate S (e.g., dollars per 1 euro) is identical to a put struck at $1/K$ on the reverse rate $1/S$ (euros per 1 dollar), after the *ad hoc* numeraire conversions: $\mathbf{c}(S, K)/S \equiv K \mathbf{p}(1/S, 1/K)$.

Price Relationship

Assuming no arbitrage, the synthetic relationship immediately translates into the well-known price relationship: “call minus put equals forward” or $c_t - p_t = f_t$. Note that here f_t denotes the price of a forward contract struck at K , that is, the present value (p.v.) of the gap between the forward price F_t and the strike price K (see **Forwards and Futures**). Denoting the price of the zero-coupon bond maturing

at T by B_t and rearranging terms, we have

$$\begin{aligned} & \text{call} + \text{p.v. of strike price} \\ &= \text{put} + \text{p.v. of forward price} \end{aligned} \quad (1)$$

or

$$c_t + K \cdot B_t = p_t + F_t \cdot B_t \quad (2)$$

For all investment assets where short selling is feasible, the forward price can be further expressed as a function of the spot price S_t and the revenue or cost of carry until maturity T (see **Forwards and Futures**). For example, the forward price of a stock with continuous dividend rate q satisfies $F_t = S_t/B_t \cdot \exp(-q(T-t))$, and put–call parity simplifies to

$$c_t + K \cdot B_t = p_t + S_t \cdot e^{-q(T-t)} \quad (3)$$

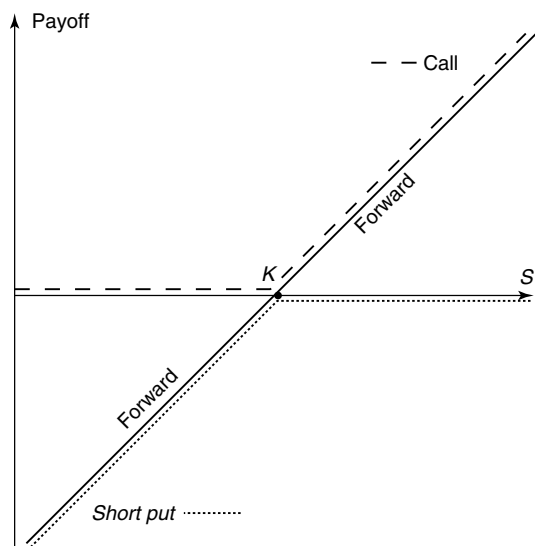
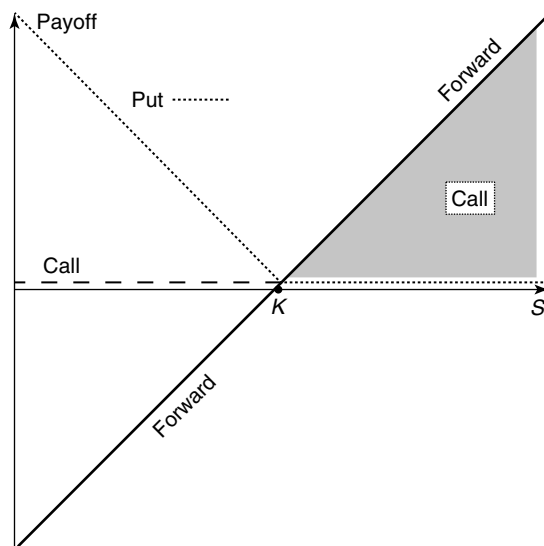
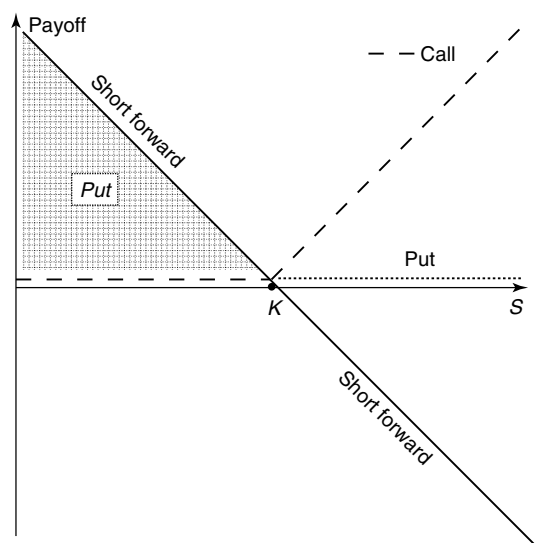
In practice, Kamara and Miller [5] give empirical evidence that while put–call parity has many small violations, almost half of the arbitrages would result in a loss when execution delays are accounted for.

Basic Implications

- For trading purposes, puts and calls are identical instruments (up to a directional position in the underlying asset).
- At-the-money-forward calls and puts must have the same value. (An at-the-money-forward option has its strike set at the forward price of the underlying asset.)
- In the absence of revenue or cost of carry, the deltas (see **Black–Scholes Formula; Delta Hedging**) of a call and put must add up to 1 (in absolute value).
- Puts and calls must have the same gamma and vega (see **Black–Scholes Formula; Gamma Hedging**).

In volatility modeling, put–call parity implies that calls and puts of identical characteristics must have the same **implied volatility**.

In exotic option pricing, Carr and Lee [1] put forward the idea of a generalized American option that may be indefinitely exercised until maturity to lock-in the intrinsic value and switch between call and put styles. The authors show that this option may be replicated by holding onto a European

Figure 1 $c - p \equiv f$ Figure 3 $c \equiv p + f$ Figure 2 $p \equiv c - f$

vanilla call and subsequently selling and buying the forward contract at every exercise. This strategy is a straightforward illustration of how put–call parity may be exploited to alternate between call and put positions by only trading in the forward contract.

History

Haug [3] traces put–call parity as far back as the seventeenth century, but its formulation was then “diffuse”. According to the author, an early formulation of put–call parity “as we know it” can be found in the work by Higgins [4], who wrote in 1902:

It can be shown that the adroit dealer in options can convert a ‘put’ into a ‘call’, a ‘call’ into a ‘put’ [...] by dealing against it in the stock.

Derman and Taleb [2] argue that the Black–Scholes–Merton formulas could have been established earlier than 1973 *via* put–call parity instead of the dynamic replication argument. Specifically, the authors cite similar formulas published in the 1960s, all of which “involved unknown risk premiums that would have been determined to be zero had [...] the put–call replication argument” been used.

Put–call parity can fail when there are restrictions on short selling, when the underlying asset is hard to borrow or illiquid, or in the case of corporate events such as leveraged buyouts.

End Notes

^aThe reason put–call parity fails with American options is best seen in the first instance ($c - p \equiv f$), whereby an

agent attempts to replicate a forward contract by buying a call and selling a put. If the put is American, it may be exercised against the agent before maturity, thus breaking the replication strategy.

References

- [1] Carr, P. & Lee, R. (2002). *Hyper Options*. Working paper, Courant Institute and Stanford University, December 2002.
- [2] Derman, E. & Taleb, N.N. (2005). The illusions of dynamic replication, *Quantitative Finance* **5**(4), 323–326.
- [3] Haug, E. (2007). *Derivatives: Models on Models*. Wiley.
- [4] Higgins, L.R. (1902). *The Put-and-Call*. E. Wilson, London.
- [5] Kamara, A. & Miller, T.W. (1995). Daily and intradaily tests of European put-call parity, *Journal of Financial and Quantitative Analysis* **30**, 519–539.

Related Articles

Black–Scholes Formula; Call Options; Forwards and Futures; Option Pricing; General Principles; Options: Basic Definitions.

SÉBASTIEN BOSSU

Discretely Monitored Options

Traditional pricing models for path-dependent options rely on continuously monitoring the underlying, often resulting in closed-form or analytic formulas. References include [14, 19, 20, 21] for barrier options, [6, 12, 13] for look-back options, and [11, 18] for Asian or average options. However, in practice, monitoring is performed over discrete dates (e.g., monthly, weekly, or daily), while the underlying is still assumed to follow a continuous model. In contrast to continuous monitoring, discrete monitoring rarely, if ever, leads to similarly tractable solutions and using continuous monitoring as approximation for discrete monitoring often leads to significant mispricing (cf. [5, 15, 16].) As a consequence, various approaches have been followed to arrive at practically useful computational schemes.

For illustration, we focus on a down-and-out call option, where a standard call option with strike K is canceled if the underlying falls below a barrier prior to expiry T . We first assume the traditional Black–Scholes–Merton setup with the price $\{S_t\}$ of the underlying following a geometric Brownian motion

$$S_t = S_0 e^{B_t} \quad (1)$$

where $\{B_t\}$ is a Brownian motion with drift $r - \sigma^2/2$ and standard deviation σ . Here the parameters r and σ represent the prevailing risk-free rate and the return volatility of the underlying asset, respectively. Let $H > 0$ be a given constant (barrier) and assume $H < S_0$. With monitoring effected over a set of m dates $n\Delta t$ ($n = 1, \dots, m$) such that $\Delta t = T/m$, let $U_n = X_1 + X_2 + \dots + X_n$, where the X_i s are independent normal random variables with mean $\mu = (r - \sigma^2/2) \Delta t$ and standard deviation $\tilde{\sigma} = \sigma \sqrt{\Delta t}$. Then the call is knocked-out the first (random) time $\tau \in \{1, 2, \dots, m\}$ such that $H \geq S_\tau$ and the time-0 price of such a call is

$$V_m(H) = e^{-rT} E \left(S_0 e^{U_m} - K \right)^+ 1_{\{\tau > m\}} \quad (2)$$

where $\tau = \inf\{n : U_n \leq \log(H/S_0)\}$. The main source of the evaluation of the above expectation

is the computational complexity associated with an m -variate normal distribution for even moderate values of m . For example, Monte Carlo or tree-based algorithms may take several hours or even days for common values of m [4]. In their paper, Broadie *et al.* [3] (see also [17]), opt to circumvent this hurdle by linking $V_m(H)$ to the price of a continuously monitored option with a barrier shifted away from the original. More precisely, they show that

$$V_m(H) = V \left(H e^{\pm \beta \sigma \sqrt{\Delta t}} \right) + o(1/\sqrt{m}) \quad (3)$$

where $V(\tilde{H})$ is the price of a continuously monitored barrier option with threshold $\tilde{H}$ and $\beta \approx 0.5826$, with $+$ for an up option and $-$ for a down option.

Although this approach works very well, it appears to be inaccurate when the barrier is near the initial price of the underlying. Under such circumstances, one can opt to use the recursive method of Ait-Sahlia and Lai [1], which consists in reducing an m -dimensional integration problem to successively evaluating m one-dimensional integrals. Specifically, they show that

$$V_m(H) = \int_{\log(H/S_0)}^{\infty} (S_0 e^x - K)^+ f_m(x) dx \quad (4)$$

where, for $1 \leq n \leq m$, $f_n(x) dx = P\{\tau > n, U_n \in dx\}$ for $x > \log(H/S_0)$, with f_n defined recursively for each n according to the following:

$$\begin{aligned} f_1(x) &= \psi(x) \\ f_n(x) &= \int_{\log(H/S_0)}^{\infty} f_{n-1}(y) \psi(x - y) dy \\ &\text{for } 2 \leq n \leq m \end{aligned} \quad (5)$$

Here $\psi(x) = \tilde{\sigma}^{-1} \phi = ((x - \mu)/\tilde{\sigma})$, with ϕ being the density of a standard normal distribution, and $f_n(x) = 0$, for $x \leq \log(H/S_0)$ and $1 \leq n \leq m$. This approach is very accurate and efficient for generally moderate values of m , using as little as 20 integration points.

Both the continuity correction and recursive integration methods can also be similarly applied to discretely monitored look-back options (cf. [2] and [4].) Alternatives to the above abound as well. Fusai *et al.* [10] use a Wiener–Hopf machinery to also compute hedge parameters. In the

context of a GARCH model, Duan *et al.* [7] propose a Markov chain technique that can also handle American-style exercise. Partial differential equations are used in [9, 22, 23, 24] to price average and barrier options, including when volatility is stochastic and exercise is of American style. Finally, [8] contains an approach that ultimately relies on Hilbert and Fourier transform techniques to address the situation when the underlying follows a Lévy process.

References

- [1] AitSahlia, F. & Lai, T. (1997). Valuation of discrete barrier and hindsight options, *Journal of Financial Engineering* **6**, 169–177.
- [2] AitSahlia, F. & Lai, T. (1998). Random walk duality and the valuation of discrete lookback options, *Applied Mathematical Finance* **5**, 227–240.
- [3] Broadie, M., Glasserman, P. & Kou, S. (1997). A continuity correction for discrete barrier options, *Mathematical Finance* **7**, 325–349.
- [4] Broadie, M., Glasserman, P. & Kou, S. (1999). Connecting discrete and continuous path-dependent options, *Finance and Stochastics* **3**, 55–82.
- [5] Chance, D. (1994). The pricing and hedging of limited exercise caps and spreads, *Journal of Financial Research* **17**, 561–583.
- [6] Conze, A. & Viswanathan, R. (1991). Path dependent options: the case of lookback options, *Journal of Finance* **46**, 1893–1907.
- [7] Duan, J.C., Dudley, E., Gauthier, G. & Simonato, J.G. (2003). Pricing discretely monitored barrier options by a Markov Chain, *Journal of Derivatives* **10**, 9–31.
- [8] Feng, L. & Linetsky, V. (2008). Pricing discretely monitored barrier options and defaultable bonds in Lévy process models: a fast Hilbert transform approach, *Mathematical Finance* **18**, 337–384.
- [9] Forsyth, P.A., Vetzal, K. & Zvan, R. (1999). A finite element approach to the pricing of discrete lookbacks with stochastic volatility, *Applied Mathematical Finance* **6**, 87–106.
- [10] Fusai, G., Abrahams, D. & Sgarra, C. (2006). An exact analytical solution for discrete barrier options, *Finance and Stochastics* **10**, 1–26.
- [11] Geman, H. & Yor, M. (1993). Bessel processes, Asian options and perpetuities, *Mathematical Finance* **3**, 349–375.
- [12] Goldman, M., Sosin, H. & Gatto, M. (1979). Path dependent options: ‘Buy at the low, sell at the high’, *Journal of Finance* **34**, 1111–1127.
- [13] Goldman, M., Sosin, H. & Shepp, L. (1979). On contingent claims that insure ex-post optimal stock market timing, *Journal of Finance* **34**, 401–414.
- [14] Heynen, R.C. & Kat, H.M. (1994). Partial barrier options, *Journal of Financial Engineering* **3**, 253–274.
- [15] Heynen, R.C. & Kat, H.M. (1994). Lookback options with discrete and partial monitoring of the underlying price, *Applied Mathematical Finance* **2**, 273–284.
- [16] Kat, H. & Verdonk, L. (1995). Tree surgery, *Risk* **8**, 53–56.
- [17] Kou, S. (2003). On pricing of discrete barrier options, *Statistica Sinica* **13**, 955–964.
- [18] Linetsky, V. (2004). Spectral expansions for Asian (average price) options, *Operations Research* **52**, 856–867.
- [19] Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.
- [20] Rich, D. (1994). The mathematical foundations of barrier option pricing theory, *Advances in Futures and Options Research* **7**, 267–312.
- [21] Rubinstein, M. & Reiner, E. (1991). Breaking down the barriers, *Risk* **4**, 28–35.
- [22] Vetzal, K. & Forsyth, P.A. (1999). Discrete Parisian and delayed barrier options: a general numerical approach, *Advanced Futures Options Research* **10**, 1–16.
- [23] Zvan, R., Forsyth, P.A. & Vetzal, K. (1999). Discrete Asian barrier options, *Journal of Computational Finance* **3**, 41–68.
- [24] Zvan, R., Vetzal, K. & Forsyth, P.A. (2000). PDE methods for pricing barrier options, *Journal of Economic Dynamics and Control* **24**, 1563–1590.

FARID AITSAHLIA

Weighted Variance Swap

Let the underlying process Y be a semimartingale taking values in an interval I . Let $\varphi : I \rightarrow \mathbb{R}$ be a difference of convex functions, and let $X := \varphi(Y)$. A typical application takes Y to be a positive price process and $\varphi(y) = \log y$ for $y \in I = (0, \infty)$.

Then (the floating leg of) a forward-starting *weighted variance swap* or *generalized variance swap* on $\varphi(Y)$ (shortened to “on Y ” if the φ is understood), with *weight* process w_t , forward-start time θ , and expiry T , is defined to pay, at a fixed time $T_{\text{pay}} \geq T > \theta \geq 0$,

$$\int_{\theta}^T w_t d[X]_t \quad (1)$$

where $[\cdot]$ denotes quadratic variation. In the case that $\theta = 0$, the trade date, we have a spot-starting weighted variance swap. The basic cases of weights take the form $w_t = w(Y_t)$, for a measurable function $w : I \rightarrow [0, \infty)$, such as the following.

1. The weight $w(y) = 1$ defines a *variance swap* (see **Variance Swap**).
2. The weight $w(y) = \mathbb{1}_{y \in C}$, the indicator function of some interval C , defines a *corridor variance swap* (see **Corridor Variance Swap**) with corridor C . For example, a corridor of the form $C = (0, H)$ produces a *down* variance swap.
3. The weight $w(y) = y/Y_0$ defines a *gamma swap* (see **Gamma Swap**).

Model-free Replication and Valuation

Assuming a deterministic interest rate r_t , let Z_t be the time- t price of a bond that pays 1 at time T_{pay} . Assume that Y is the continuous price process of a share that pays continuously a deterministic proportional dividend q_t . Let

$$Z_t = \exp\left(-\int_t^{T_{\text{pay}}} r_u du\right) \quad \text{and} \quad Q_t := \exp\left(\int_0^t q_u du\right) \quad (2)$$

so the share price with reinvested dividends is $Y_t Q_t$. Then the payoff

$$\int_{\theta}^T w(Y_t) d[X]_t \quad (3)$$

admits a model-independent replication strategy, which holds European options statically and trades the underlying shares dynamically. Indeed, let $\lambda : I \rightarrow \mathbb{R}$ be a difference of convex functions, let λ_y denote its left-hand derivative, and assume that its second derivative in the distributional sense has a signed density, denoted λ_{yy} , which satisfies for all $y \in I$

$$\lambda_{yy}(y) = 2\varphi_y^2(y)w(y) \quad (4)$$

where φ_y denotes the left-hand derivative of φ . Then

$$\begin{aligned} \int_{\theta}^T w(Y_t) d[X]_t &= \lambda(Y_T) - \lambda(Y_{\theta}) - \int_{\theta}^T \lambda_y(Y_t) dY_t \\ &= \lambda(Y_T) - \lambda(Y_{\theta}) + \int_{\theta}^T (q_t - r_t) \lambda_y(Y_t) Y_t dt \\ &\quad - \int_{\theta}^T \lambda_y(Y_t) \frac{Z_t}{Q_t} d(Y_t Q_t / Z_t) \end{aligned} \quad (5)$$

where equation (5) is by a proposition in [2] that slightly extends [1], and equation (6) is by Ito's rule. So the following self-financing strategy replicates (and hence prices) the payoff (3). Hold statically a claim that pays at time T_{pay}

$$\lambda(Y_T) - \lambda(Y_{\theta}) + \int_{\theta}^T (q_{\tau} - r_{\tau}) \lambda_y(Y_{\tau}) Y_{\tau} d\tau \quad (7a)$$

and trade shares dynamically, holding at each time $t \in (\theta, T)$

$$-\lambda_y(Y_t) Z_t \quad \text{shares} \quad (7b)$$

and a bond position that finances the shares and accumulates the trading gains or losses. Hence, the payoff (3) has time-0 value equal to that of the replicating claim (7a), which is synthesizable from Europeans with expiries in $[\theta, T]$. Indeed, for a put/call separator κ (such as $\kappa = Y_0$), if $\lambda(\kappa) = \lambda_y(\kappa) = 0$, then

each λ claim decomposes into puts/calls at all strikes K , with quantities $2\varphi_y^2(K)w(K) dK$:

$$\lambda(y) = \int_I 2\varphi_y^2(K)w(K)\text{Van}(y, K) dK \quad (8)$$

where $\text{Van}(y, K) := (K - y)^+ \mathbb{1}_{K < \kappa} + (y - K)^+ \mathbb{1}_{K > \kappa}$ denotes the vanilla put or call payoff. For put/call decompositions of general European payoffs, see [1].

Futures-dependent Weights

In equation (3), the weight is a function of spot Y_t . The alternative payoff specification

$$\int_{\theta}^T w(Y_t Q_t / Z_t) d[X]_t \quad (9)$$

makes w_t a function of the *futures* price (a constant times $Y_t Q_t / Z_t$).

In the case $\varphi = \log$, we have $[X] = [\log Y] = [\log(YQ/Z)]$; hence

$$\begin{aligned} & \int_{\theta}^T w(Y_t Q_t / Z_t) d[X]_t \\ &= \lambda\left(\frac{Y_T Q_T}{Z_T}\right) - \lambda\left(\frac{Y_{\theta} Q_{\theta}}{Z_{\theta}}\right) \\ & \quad - \int_{\theta}^T \lambda_y(Y_t Q_t / Z_t) d(Y_t Q_t / Z_t) \end{aligned}$$

for λ satisfying equation (4). So the alternative payoff (9) admits replication as follows: hold statically a claim that pays at time T_{pay}

$$\lambda(Y_T Q_T / Z_T) - \lambda(Y_{\theta} Q_{\theta} / Z_{\theta}) \quad (10a)$$

and trade shares dynamically, holding at each time $t \in (\theta, T)$

$$- \lambda_y(Y_t Q_t / Z_t) Q_t \text{ shares} \quad (10b)$$

and a bond position that finances the shares and accumulates the trading gains or losses. Thus, the payoffs (9) and (10a) have equal values at time 0.

In special cases (such as $w = 1$ or $r = q = 0$), the spot-dependent (3) and futures-dependent (9) weight specifications are equivalent. In general, the spot-dependent weighting is harder to replicate, as it requires a continuum of expiries in equation (7a),

unlike equation (10a). The spot-dependent weighting is, however, the more common specification and is assumed in remainder of this article.

Examples

Returning to the previously specified examples of weights $w(Y_t)$, we express the replication payoff λ in a compact formula, and also expanded in terms of vanilla payoffs according to equation (8). We take $\varphi(y) = \log y$ unless otherwise stated.

- *Variance swap*: Equation (4) has solution

$$\begin{aligned} \lambda(y) &= -2 \log(y/\kappa) + 2y/\kappa - 2 \\ &= \int_0^{\infty} \frac{2}{K^2} \text{Van}(y, K) dK \end{aligned} \quad (11)$$

- *Arithmetic variance swap*: For $\varphi(y) = y$, equation (4) has solution

$$\lambda(y) = (y - \kappa)^2 = \int_0^{\infty} 2 \text{Van}(y, K) dK \quad (12)$$

- *Corridor variance swap*: Equation (4) has solution

$$\lambda(y) = \int_{K \in C} \frac{2}{K^2} \text{Van}(y, K) dK \quad (13)$$

- *Gamma swap*: Equation (4) has solution

$$\begin{aligned} \lambda(y) &= \frac{2}{Y_0} \left[y \log(y/\kappa) - y + \kappa \right] \\ &= \int_0^{\infty} \frac{2}{Y_0 K} \text{Van}(y, K) dK \end{aligned} \quad (14)$$

In all cases, the strategy (7) replicates the desired contract. In the case of a variance swap, the strategy (10) also replicates it, because $w(Y) = 1 = w(YQ/Z)$.

Discrete Dividends

Assume that at the fixed times t_m where $\theta = t_0 < t_1 < \dots < t_M = T$, the share price jumps to $Y_{t_m} = Y_{t_m-} - \delta_m(Y_{t_m-})$, where each discrete dividend is given by a function δ_m of prejump price. In this case, the dividend-adjusted weighted variance swap can be defined to pay at time T_{pay}

$$\sum_{m=1}^M \int_{t_{m-1}+}^{t_m-} w(Y_t) d[X]_t \quad (15)$$

If the function $y \mapsto y - \delta_m(y)$ has an inverse $f_m : I \rightarrow I$, and if Y is still continuous on each $[t_{m-1}, t_m)$, then each term in equation (15) can be constructed via equation (7), together with the relation $\lambda(Y_{t_m-}) = \lambda(f_m(Y_{t_m}))$. Specifically, the m th term admits replication by holding statically a claim that pays at time T_{pay}

$$\begin{aligned} & \lambda(f_m(Y_{t_m})) - \lambda(Y_{t_{m-1}}) \\ & + \int_{t_{m-1}}^{t_m} (q_\tau - r_\tau) \lambda_y(Y_\tau) Y_\tau \, d\tau \end{aligned} \quad (16)$$

and holding dynamically $-\lambda_y(Y_t)Z_t$ shares, at each time $t \in (t_{m-1}, t_m)$.

Contract Specifications in Practice

In practice, weighted variance swap transactions are forward settled; no payment occurs at time 0, and at time T_{pay} the party long the swap receives the total payment

$$\text{Notional} \times (\text{Floating} - \text{Fixed}) \quad (17)$$

where “fixed” (also known as the *strike*), expressed in units of annualized variance, is the price contracted at time 0 for time- T_{pay} delivery of “floating”, an annualized discretization of equation (15) that monitors Y , typically daily, for N periods. In the usual case of $\varphi = \log$, this results in a specification

Floating := Annualization

$$\times \sum_{n=1}^N w(Y_n) \left(\log \frac{Y_n + D_n}{Y_{n-1}} \right)^2 \quad (18)$$

where D_n denotes the discrete dividend payment, if any, of the n th period. Both here and in the theoretical form (15), no adjustment is made for any dividends deemed to be continuous (for example, index variance contracts typically do not adjust for index dividends; see [3]).

In some contracts—for example, single-stock (down-)variance—the risk to the variance seller that Y crashes is limited by imposing a cap on the payoff. Hence,

$$\text{Notional} \times \left(\min(\text{Floating}, \text{Cap} \times \text{Fixed}) - \text{Fixed} \right) \quad (19)$$

replaces equation (17), where “Cap” is an agreed constant, such as the square of 2.5.

References

- [1] Carr, P. & Madan, D. (1998). Towards a theory of volatility trading, in *Volatility*, R. Jarrow, ed, Risk Publications, pp. 417–427.
- [2] Carr, P. & Lee, R. (2009). *Hedging Variance Options on Continuous Semimartingales*, Forthcoming in Finance and Stochastics.
- [3] Overhaus, M., Bermúdez, A., Buehler, H., Ferraris, A., Jorindson, C. & Lamnouar, A. (2007). *Equity Hybrid Derivatives*, John Wiley & Sons.

Related Articles

Corridor Variance Swap; Gamma Swap; Variance Swap.

ROGER LEE

Model Calibration

The fundamental theorem of asset pricing (*see **Fundamental Theorem of Asset Pricing***) shows that, in an arbitrage-free market, market prices can be represented as (conditional) expectations with respect to a martingale measure $\mathbb{Q}$: a probability measure $\mathbb{Q}$ on the set Ω of possible trajectories $(S_t)_{t \in [0, T]}$ of the underlying asset such that the asset price S_t/N_t discounted by the numeraire N_t is a martingale. The value $V_t(H_T)$ of a (discounted) terminal payoff H_T at T is then given by

$$V_t(H_T) = E^{\mathbb{Q}}[B(t, T)H_T | \mathcal{F}_t] \quad (1)$$

where $B(t, T) = N_t/N_T$ is the discount factor. For example, the value under the pricing rule $\mathbb{Q}$ of a call option with strike K and maturity T is given by $E^{\mathbb{Q}}[B(t, T)(S_T - K)^+ | \mathcal{F}_t]$. However, this result does not say how to construct the pricing measure $\mathbb{Q}$. Given that data sets of option prices have become increasingly available, a common approach for selecting a pricing model $\mathbb{Q}$ is to choose, given a set of liquidly traded derivatives with (discounted) terminal payoffs $(H^i)_{i \in I}$ and market prices $(C_i)_{i \in I}$, a pricing measure $\mathbb{Q}$ compatible with the observed market prices:

Problem 1 [Calibration Problem] *Given market prices $(C_i)_{i \in I}$ (say at date $t = 0$) for a set of options with discounted terminal payoffs $(H_i)_{i \in I}$, construct a probability measure $\mathbb{Q}$ on Ω such that*

- *the (discounted) asset price $(S_t)_{t \in [0, T]}$ is a martingale under $\mathbb{Q}$*

$$T \geq t \geq u \geq 0 \Rightarrow E^{\mathbb{Q}}[S_t | \mathcal{F}_u] = S_u \quad (2)$$

- *the pricing rule implied by $\mathbb{Q}$ is consistent with market prices*

$$\forall i \in I, \quad E^{\mathbb{Q}}[H_i] = C_i \quad (3)$$

where, for ease of notation, we have set discount factors to 1 (prices are discounted) and $E[\cdot]$ denotes the conditional expectation given initial information $\mathcal{F}_0$. Thus, a pricing rule $\mathbb{Q}$ is said to be calibrated to the *benchmark instruments* H_i if the value of these instruments, computed in the model, correspond to their market prices C_i .

Option prices being evaluated as expectations, this inverse problem can also be interpreted as a (generalized) moment problem for the law $\mathbb{Q}$ of risk-neutral process given a finite number of option prices, it is typically an *ill-posed* problem and can have many solutions. However, the number of observed options can be large ($\simeq 100 - 200$ for index options) and finding even a single solution is not obvious and requires efficient numerical algorithms.

In the Black–Scholes model (*see **Black–Scholes Formula***), calibration amounts to picking the volatility parameter to be equal to the implied volatility of a traded option. However, if more than one option is traded, the Black–Scholes model cannot be calibrated to market prices, since in most options markets implied volatility varies across strikes and maturities; this is the *volatility smile* phenomenon. Therefore, to solve the calibration problem, we need more flexible models, some examples of which are given here.

Example 1 [Diffusion Model (*see **Local Volatility Model***)] If an asset price is modeled as a diffusion process

$$dS_t = S_t[\mu dt + \sigma(t, S_t) dW_t] \quad (4)$$

parameterized by a local volatility function

$$\sigma : (t, S) \rightarrow \sigma(t, S) \quad (5)$$

then the values of call options can be computed by solving the Dupire equation (*see **Implied Volatility Surface***)

$$\begin{aligned} \frac{\partial C_0}{\partial T} + Kr \frac{\partial C_0}{\partial K} - \frac{K^2 \sigma^2(T, K)}{2} \frac{\partial^2 C_0}{\partial K^2} &= 0 \\ \forall K \geq 0, \quad C_0(T = 0, K) &= (S - K)^+ \end{aligned} \quad (6)$$

The corresponding inverse problem is to find a (smooth) volatility function $\sigma : [0, T] \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $C^\sigma(T_i, K_i) = C^*(T_i, K_i)$ where C^σ is the solution of equation (6) and $C^*(T_i, K_i)$ are the market prices of call options.

Example 2 In an exponential–Lévy model $S_t = \exp X_t$, where X_t is a Lévy process (*see **Exponential Lévy Models***) with diffusion coefficient $\sigma > 0$ and Lévy measure ν , call prices $C^{\sigma, \nu}(t_0, S_0; T_i, K_i)$ are easily computed using Fourier-based methods (*see*

Fourier Methods in Options Pricing). The calibration problem is to find σ, ν such that

$$\forall i \in I, C^{\sigma, \nu}(t_0, S_0; T_i, K_i) = C^*(T_i, K_i) \quad (7)$$

This is an example of a nonlinear inverse problem where the parameter lies in a space of measures.

Example 3 In the LIBOR market model, a set of N interest rates (LIBOR rates) is modeled as a diffusion process $L_t = (L_t^i)_{i=1..N}$ with constant covariance matrix $\Sigma = {}^t \sigma \cdot \sigma \in \text{Sym}^+(n \times n)$:

$$dL_t^i = \mu_t^i dt + L_t^i \sigma_i \cdot dW_t^j \quad (8)$$

This model can then be used to analytically price caps, floors, and swaptions (using a lognormal approximation), whose prices depend on the entries of the covariance matrix Σ . The calibration problem is to find a symmetric semidefinite positive matrix $\Sigma \in \text{Sym}^+(n \times n)$ such that the model prices C^Σ match market prices

$$\forall i \in I, C^\Sigma(T_i, K_i) = C^*(T_i, K_i) \quad (9)$$

This problem can be recast as a semi-definite programming problem [2].

Other examples include the construction of yield curves from bond prices (*see Bond Options*) calibration of term structure models (*see Term Structure Models*) to bond prices, recovering the distribution of volatility from option prices [28] calibration to American options in diffusion models [1] and recovery of portfolio default rates from market quotes of credit derivatives [16, 18].

These problems are typically ill-posed in the sense that, either solutions may not exist (model class is too narrow to reproduce observations) or solutions are not unique (if data is finite or sparse). In practice, existence of a solution is restored by formulating the problem as an optimization problem

$$\inf_{\theta \in E} F(C^\theta - C) \quad (10)$$

where E is the parameter space and F is a loss function applied to the discrepancy $C^\theta - C$ between market prices and model prices. An algorithm is then used to retrieve *one* solution and the main issue is the *stability* of this reconstructed solution as a function of inputs (market prices).

Inversion formulas

In the theoretical situation where prices of European options are available for all strikes and maturities, the calibration problem can sometimes be explicitly solved using an inversion formula.

For the diffusion model in Example 1, the Dupire formula [25] (*see Dupire Equation*):

$$\sigma(T, K) = \sqrt{\frac{\frac{\partial C_0}{\partial T} + Kr \frac{\partial C_0}{\partial K}}{\frac{K^2}{2} \frac{\partial^2 C_0}{\partial K^2}}} \quad (11)$$

allows to invert the volatility function from call option prices. Similar formulas can be obtained in credit derivative pricing models, for inverting portfolio default rates from collateralized debt obligation (CDO) tranche spreads [16] and pure jump models with state-dependent jump intensity (“local Levy” model) [12]. No such inversion formula is available in the case of American options (*see American Options*). The Dupire formula (11) has been widely used by practitioners for recovering the local volatility function from call/put option prices by interpolating in strike and maturity and applying equation (11). However, since equation (11) involves differentiating the inputs, it suffers from instability and sensitivity to small changes in inputs, as shown in Figure 1. This instability deters one from using inversion formulas such as equation (6) even in the rare cases where they exist.

Least-squares Formulation

Typically, if the model is misspecified, the observed option prices may not lie within the range of prices attainable by the model. Also, option prices are defined up to a bid–ask spread: a model may generate prices compatible with the market but may not exactly fit the mid-market prices for any given $\theta \in E$. For these reasons, one often reformulates calibration as a least-squares problem

$$\inf_{\theta \in E} J_0(\theta), \quad J_0(\theta) = \sum_{i=1}^I w_i |C_i(\theta) - C_i|^2 \quad (12)$$

where C_i are mid-market quotes and $w_i > 0$ are a set of weights, often chosen inversely proportional to the (squared) bid–ask spread of C_i .

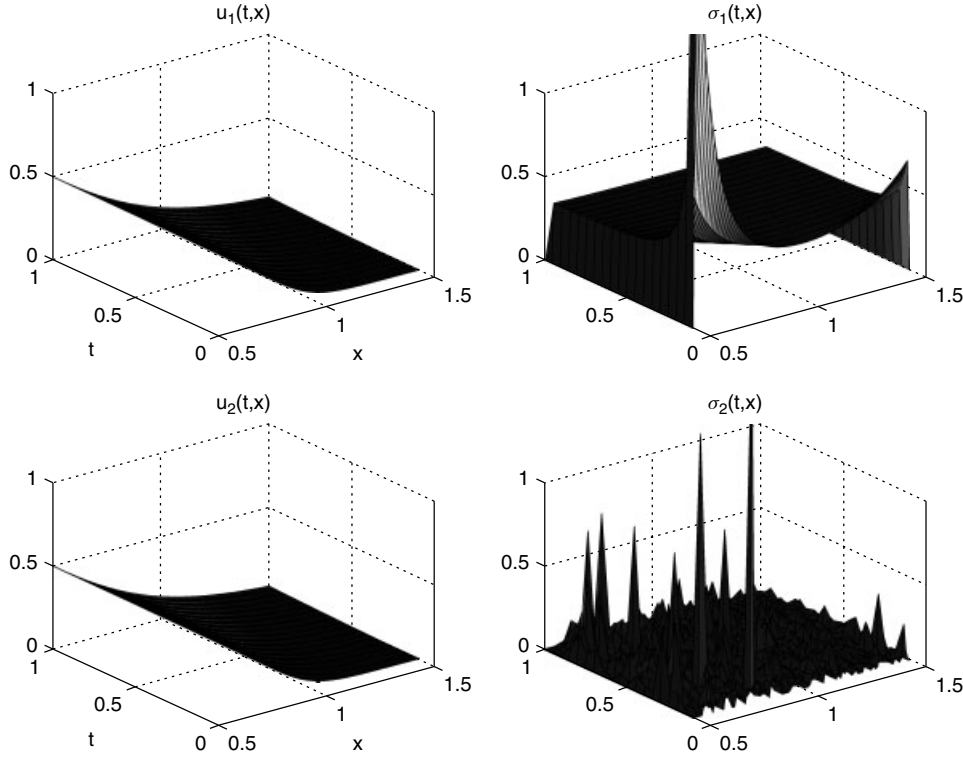


Figure 1 Extreme sensitivity of Dupire formula to noise in the data. Two examples of call price function (left) and their corresponding local volatilities (right). The prices differ through IID noise $\sim UNIF(0, 0.001)$, representing a bid–ask spread

In most models, the call prices are computed numerically *via* Fourier transform (see **Fourier Methods in Options Pricing**) or by solving a partial differential equation (PDE) (see **Partial Differential Equations**). However, in many situations (short or long maturity, small vol–vol, etc.) approximation formulae for implied volatilities $\Sigma(T_i, K_i)$ of call options are available [5, 10, 11, 30] in terms of model parameters (see **Implied Volatility in Stochastic Volatility Models; Implied Volatility: Volvol Expansion; Implied Volatility: Long Maturity Behavior; SABR Model**). In these situations, parameters are calibrated by a least-squares fit to the approximate formula:

$$\inf_{\theta \in E} \sum_{i=1}^I w_i |\Sigma(T_i, K_i; \theta) - \Sigma^*(T_i, K_i)|^2 \quad (13)$$

An example is the SABR model (see **SABR Model**), whose popularity is almost entirely due

to its ease of calibration using the Hagan formula [30].

In most cases, option prices $C_i(\theta)$ depend continuously on θ and E is a subset of a finite dimensional space (i.e., there are a finite number of bounded parameters), so the least-squares formulation always admits a solution. However, the solution of equation (12) need not be unique: J_0 may, in fact, have several *global* minima, when the observed option prices do not uniquely identify the model. Figures 2 and 3 show examples of the function J_0 for some popular parametric option pricing models, computed using a data set of DAX index options prices on May 11, 2001. The pricing error in the Heston stochastic volatility model (see **Heston Model**), shown in figure as a function of the “volatility of volatility” and the mean reversion rate, displays a line of local minima. The pricing error for the variance gamma model (see **Variance-gamma Model**) in Figure 3 displays a non-convex profile, with two distinct minima in the range

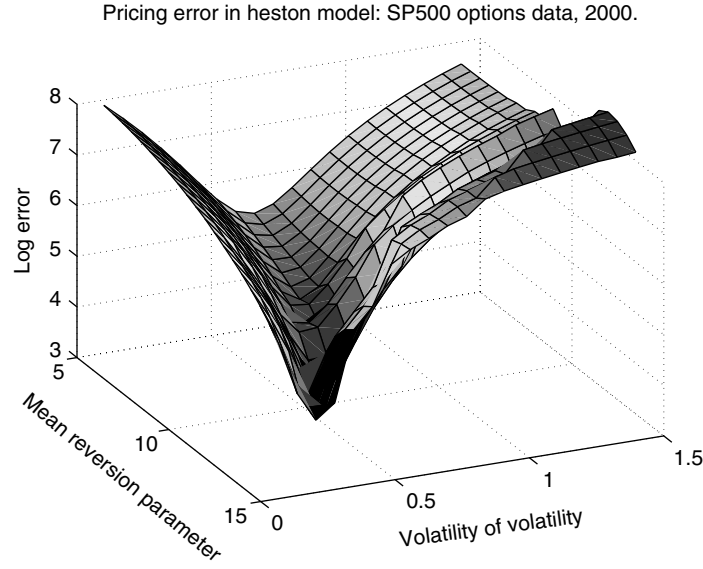


Figure 2 Error surface for the Heston stochastic volatility model, DAX options

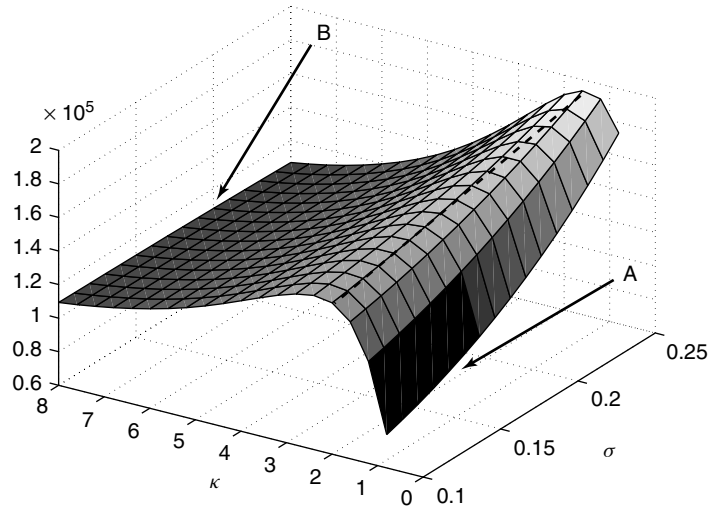


Figure 3 Error surface for variance gamma (pure jump) model, DAX options

of observed values. These examples show that, even if the number of observations (option prices) is much higher than the number of parameters, this does not imply identifiability of parameters.

Regularization methods can be used to overcome this problem [27]. A common method is to have a convex penalty term R , called the *regularization*

term, to the pricing error and solve the auxiliary problem:

$$\inf_{\theta \in E} J_{\alpha}(\theta) \quad (14)$$

where

$$J_{\alpha}(\theta) = J_0(\theta) + \alpha R(\theta) \quad (15)$$

The functional (16) consists of two parts: the regularization term $\alpha R(\theta)$ which is convex in its argument and the quadratic pricing error which measures the precision of calibration. The coefficient α , called *regularization parameter*, defines the relative importance of the two terms: it characterizes the trade-off between prior knowledge and the information contained in option prices. $J_\alpha(\cdot)$ is usually minimized by gradient-based methods, where the crux of the algorithm is an efficient computation of the gradient $\nabla_\theta J$.

When parameter is a function (such as the local volatility function), the regularization term is often chosen to be a smoothness (e.g., Sobolev) norm. This method, called *Tikhonov regularization* (see **Tikhonov Regularization**) has been applied to diffusion models [1, 2, 13, 23, 26] and to exponential-Lévy models [19].

Another popular choice of regularization term is the relative entropy (see **Entropy-based Estimation**) $R(\theta) = H(\mathbb{Q}^\theta | \mathbb{P})$ with respect to a prior probability measure $\mathbb{P}$. In continuous-time models, relative entropy can be used as regularization criterion only if the prior possesses a nonempty class of equivalent martingale measures, that is, it corresponds to an incomplete market model (see **Complete Markets**). From a calibration perspective, market incompleteness (i.e., the nonuniqueness of equivalent martingale measure) is therefore an *advantage*: it allows to conciliate compatibility with option prices and equivalence with respect to a reference probability measure. Examples are provided by jump processes (see **Jump Processes; Exponential Lévy Models**) or reduced-form credit risk models (see **Reduced Form Credit Risk Models**): one can modify the jump size distribution (Lévy measure) or the default intensity while preserving equivalence (see **Equivalence of Probability Measures**) of measures [18, 20]. For Lévy processes (see **Exponential Lévy Models**), the relative entropy term $H(\nu)$ is computable in terms of the Lévy measure ν [21]. The calibration problem then takes the following form:

Problem 2 *Given a prior Lévy process with law $\mathbb{Q}_0$ and characteristics (σ_0, ν_0) , find a Lévy measure ν which minimizes*

$$J_\alpha(\nu) = \alpha H(\nu) + \sum_{i=1}^N w_i (C_0^\nu(T_i, K_i) - C_0(T_i, K_i))^2 \quad (16)$$

This regularized formulation has the advantage that its solution exhibits continuous dependence on market prices and with respect to the choice of the prior model [21, 22].

Simpler regularization methods can be used in settings where prices are computed using analytical transform methods. Belomestny & Reiss [8] propose a spectral regularization method for calibrating exponential-Lévy models. Aspremont [3] formulates the calibration of LIBOR market models (Example 3) as semidefinite programming problems under constraints.

Different regularization terms select *different* solutions: Tikhonov regularization approximates the least-squares solution with smallest norm [27] while entropy-based regularization selects the minimum-entropy least-squares solution [22].

Entropy Minimization Under Calibration Constraints

An alternative approach to regularization is to select a pricing model $\mathbb{Q}$ by minimizing the relative entropy (see **Entropy-based Estimation**) of the probability measure $\mathbb{Q}$ with respect to a prior, under calibration constraints

$$\inf_{\mathbb{Q} \sim \mathbb{P}} H(\mathbb{Q} | \mathbb{P}) \quad \text{under} \quad C_i = E^{\mathbb{Q}}[H_i] \quad \text{for } i \in I \quad (17)$$

Relative entropy being strictly convex, any solution of equation (17) is unique and can be computed in a stable manner using Lagrange multiplier (dual) methods [24] (see **Convex Duality**).

Application of these ideas to a set of scenarios leads to the *weighted Monte Carlo algorithm* (see **Weighted Monte Carlo**) [6]: one first simulates N sample paths $\Omega_N = \{\omega_1, \dots, \omega_N\}$ from a *prior* model $\mathbb{P}$ and then solves the above problem (AV) using as prior the uniform distribution on Ω_N . The idea is to weight the paths in order to verify the calibration constraints. The weights $(\mathbb{Q}_N(\omega_i), i = 1..N)$ are constructed by minimizing relative entropy under calibration constraints

$$\inf_{\mathbb{Q}_N \in \mathcal{P}(\Omega_N)} \sum_{i=1}^N \mathbb{Q}_N(\omega_i) \ln \frac{\mathbb{Q}_N(\omega_i)}{\mathbb{P}_N(\omega_i)} \quad \text{under}$$

$$\sum_{i=1}^N \mathbb{Q}_N(\omega_i) G_j(\omega_i) = C_j \quad (18)$$

This constrained optimization problem is solved by duality [6, 24]: the dual has an explicit solution, in the form of a Gibbs–Boltzmann measure [4, 6] (see **Entropy-based Estimation**). A (discounted) payoff X is then priced using the same set of simulated paths *via*

$$\begin{aligned} E^{\mathbb{Q}_N}[X] &= \sum_{i=1}^N \mathbb{Q}_N(\omega_i) X(\omega_i) \\ &= \frac{1}{N} \sum_{i=1}^N \frac{\mathbb{Q}_N(\omega_i)}{\mathbb{P}_N(\omega_i)} X(\omega_i) \end{aligned} \quad (19)$$

The benchmark payoffs (calibration instruments) play the role of *biased* control variates, leading to variance reduction [29]:

$$E^{\mathbb{Q}_N}[X] = E^{\mathbb{Q}_N} \left[X - \sum_{i=1}^I \alpha_i H_i \right] + \sum_{i=1}^I \alpha_i C_i \quad (20)$$

This method yields as a by-product, a static hedge portfolio α_i^* , which minimizes the variance in equation (20) [3, 6, 17].

A drawback is that the martingale property is lost in this process since it would correspond to an infinite number of constraints. As a result, derivative prices computed with the weighted Monte Carlo algorithm may fail to verify arbitrage relations across maturities (e.g. calendar spread relations), especially when applied to forward-starting contracts.

These arbitrage constraints can be restored by representing $\mathbb{Q}$ as a random mixture of martingales the law of random mixture being chosen *via* relative entropy minimization under calibration constraints [17]. This results in an arbitrage-free version of the weighted Monte Carlo approach, which is applied to recovering covariance matrices implied by index options in [15].

Stochastic Control Methods

In certain continuous-time models, the relative entropy minimization approach can be mapped, *via* a duality argument, into a stochastic control problem, which can then be solved using dynamic

programming techniques. Consider a Markovian model where the state variable S_t (asset price, interest rate,...) follows a stochastic differential equation

$$\begin{aligned} dX_t &= \mu_\theta(t) dt + \sigma_\theta(t, S_t) dW_t \\ &\quad + \int \gamma_\theta(t, X_{t-}) \mu(dt dz) \end{aligned} \quad (21)$$

where W is a Wiener process and μ a compensated Poisson random measure with intensity $\nu_\theta(dz) \lambda_\theta(t) dt$. The coefficients of the model are parameterized by some parameter $\theta \in E$; in a non-parametric setting, θ is just the coefficient itself and E is a functional space. Denote the law of the solution by $\mathbb{Q}^\theta$. Consider now the case where the calibration criterion $J(\cdot)$ can be expressed as an expected value $J(\theta) = E^\theta[\int_0^T \phi(X_t) dt]$ with a strictly convex function $\phi(\cdot)$. A classical approach to solve the calibration problem

$$\inf_{\theta \in E} J(\theta), \quad \text{under} \quad E^\theta[H_i] = C_i \quad (22)$$

is to introduce the *Lagrangian* functional

$$\begin{aligned} \mathcal{L}(\theta, \lambda) &= J(\theta) - \sum_{i \in I} \lambda_i (E^\theta[H_i] - C_i) \\ &= E^\theta \left[\int_0^T \phi(X_t) dt - \sum_{i \in I} \lambda_i (H_i - C_i) \right] \end{aligned} \quad (23)$$

where λ_i is the Lagrange multiplier associated to the calibration constraint for payoff H_i . The *dual* problem associated to the constrained minimization problem (22) is given by

$$\begin{aligned} \inf_{\theta \in E} \mathcal{L}(\theta, \lambda) &= \inf_{\theta \in E} E^\theta \int_0^T \phi(X_t) dt \\ &\quad - \underbrace{\sum_{i \in I} \lambda_i (H_i - C_i)}_{\Phi} \end{aligned} \quad (24)$$

It can be viewed as a *stochastic control problem* (see **Stochastic Control**) with running cost $\phi(t, X_t)$ and *terminal cost* Φ .

This original formulation of the calibration problem was first presented by Avellaneda *et al.* [7] in the

context of diffusion model with unknown volatility

$$dS_t = S_t \sigma(t, S_t) dW_t \quad (25)$$

The calibration criterion in [7] was chosen to be

$$J(\sigma) = E^\sigma \left[\int_0^T dt \eta(\sigma^2(t, X_t^\sigma)) \right] \quad (26)$$

where η is a strictly convex function. Duality between (22) and (24) is not obvious in this case since the Lagrangian is not convex with respect to its argument [31]. The stochastic control approach can also be applied in the context of model calibration by relative entropy minimization for classes of models where absolute continuity is preserved under a change of parameters, such as models with jumps. Cont and Minca [18] use this approach for retrieving the default rate in a portfolio from CDO tranche spreads indexed on the portfolio.

Stochastic Algorithms

Objective functions used in calibration (with the exception of entropy-based methods) are typically nonconvex, even after regularization, leading to multiple minima and lack of convergence in gradient-based methods. Stochastic algorithms known as evolutionary algorithms, which contain simulated annealing as a special case, have been widely used for global nonconvex optimization and are natural candidate for solving such problems [9].

Suppose, for instance, we want to minimize the pricing error

$$J_0(\theta) = \sum_{i=1}^I w_i |C_i^\theta - C_i|, \quad \theta \in E \quad (27)$$

where C_i^θ are model prices and C_i are observed (transaction or mid-market) prices for the benchmark options. Now define the *a priori* error level δ as

$$\delta = \sum_{i=1}^I w_i |C_i^{\text{bid}} - C_i^{\text{ask}}| \quad (28)$$

Given the uncertainty on option values due to bid–ask spreads, one cannot meaningfully distinguish a “perfect” fit $J_0(\theta) = 0$ from any other fit with $J_0(\theta) \leq \delta$. Therefore, all parameter values in the level set $G_\delta = \{\theta \in E, J_0(\theta) \leq \delta\}$ correspond to

models that are compatible with the market data $(C_i^{\text{bid}}, C_i^{\text{ask}})_{i=1..I}$. An evolutionary algorithm simulates an inhomogeneous Markov chain $(X_n)_{n \geq 1}$ in E^N which undergoes mutation–selection cycles [9] designed such that as the number of iterations n grows, the components $(\theta_1^N, \dots, \theta_n^N)$ of X_n converge to the G_δ , yielding a population of points (θ_k) which converges to a sample of model parameters compatible with the market data $(C_i^{\text{bid}}, C_i^{\text{ask}})_{i=1..I}$ in the sense that $J_0(\theta_k) \leq \delta$. We thus obtain a population of N model parameters calibrated to market data, which can be different especially if the initial problem has multiple solutions.

Figure 4 shows a sample of local volatility functions obtained using this approach [9]. These examples illustrate that precise reconstruction of local volatility from call option prices is at best illusory; the parameter uncertainty is too important to be ignored, especially for short maturities where it does not affect the prices very much; short-term volatility hovers anywhere between 15% and 30%. These observations cast a doubt on the volatility content of very short-term options in terms of volatility and questions whether one can solely rely on short maturity asymptotics (see **SABR Model**) in model calibration.

Parameter Uncertainty

Model calibration is usually the first step in a procedure whose ultimate purpose is the pricing and hedging of (exotic) options. Once the model parameter θ is calibrated to market prices, it is used to compute a model-dependent quantity $f(\theta)$ —price of an exotic option or a hedge ratio—using a numerical procedure. Given the ill-posedness of the calibration problem and the resulting uncertainty on the solution θ , one question is the impact of this uncertainty on such model-dependent quantities. This aspect is often neglected in practice and many users of pricing models view the calibrated parameter as fixed, equating calibration with a curve-fitting exercise.

Particle methods yield, as a by-product, a way to analyze model uncertainty. While calibration algorithms based on deterministic optimization yield a *point estimate* for model parameters, particle methods yield a *population* $\mathcal{Q} = \{\mathbb{Q}_{\theta_1}, \dots, \mathbb{Q}_{\theta_k}\}$ of pricing models, all of which price the benchmark options with equivalent precision $E^{\mathbb{Q}}(H_i) \in [C_i^{\text{bid}}, C_i^{\text{ask}}]$. The

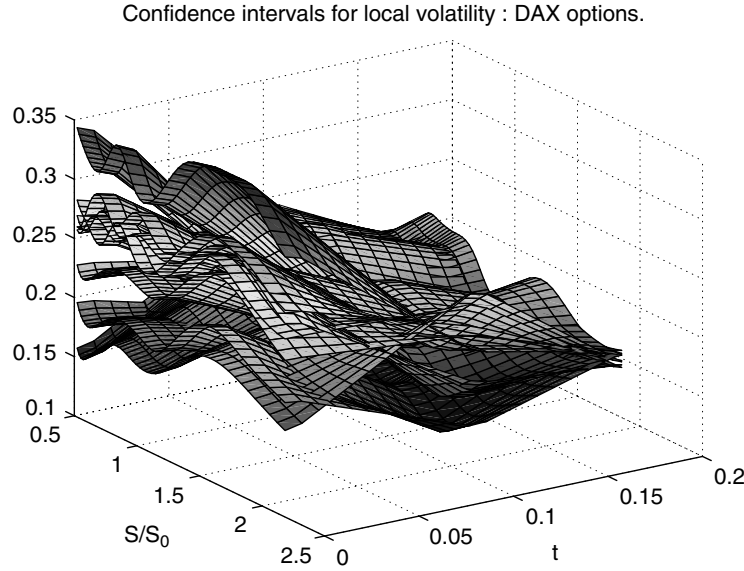


Figure 4 A sample of local volatility surfaces calibrated to DAX options

heterogeneity of this population reflects the uncertainty in model parameters, which are left undetermined by the benchmark options. This idea can be exploited to produce a quantitative measure of model uncertainty compatible with observed market prices of benchmark instruments [14], by considering the interval of prices

$$\left[\inf_{Q \in \mathcal{Q}} E^Q[X], \sup_{Q \in \mathcal{Q}} E^Q[X] \right] \quad (29)$$

for a payoff X in the various calibrated models. Another approach is to calibrate several different models to the same data and compare the value of the exotic option across models [14, 32]. Model uncertainty in derivative pricing is further discussed in [14].

Relation with Pricing and Hedging

Calibrating a model to market prices simply ensures that model prices of benchmark instruments reflect current “mark-to-market” values. It also ensures that the cost of a static hedge (see **Static Hedging**) using these benchmark instruments is correctly reflected in model prices: if a payoff H can be statically hedged

with a portfolio containing α_i units of benchmark instrument H_i ,

$$H = \alpha_0 + \sum_{i \in I} \alpha_i H_i \quad (30)$$

the cost $\alpha_0 + \sum \alpha_i C_i$ of setting up the hedge is automatically equal to the model price $E^Q[H]$.

Calibration does not entail that prices, hedge ratios, or risk parameters generated by the model are “correct” in any sense. This requires a correct model specification with realistic dynamics for risk factors. Indeed, many different models may calibrate the same prices of, say, a set of call options but lead to very different prices of hedge ratios for exotics [14, 32]. For example, any equity volatility smile can be reproduced by a one-factor diffusion model (see Example 1) *via* an appropriate specification of the local volatility surface, but there is ample evidence that volatility itself should be modeled as a risk factor (see **Stochastic Volatility Models**) and a one-factor diffusion may lead to an underestimation of volatility risk and unrealistic dynamics [30].

However, a model that is *not calibrated* to market prices of liquidly traded derivatives is typically not easy to use. For example, even if a payoff can be statically hedged with traded derivatives using an initial capital V_0 , the model price will not be

equal to V_0 . Thus, model prices will, in general, be inconsistent with hedging costs if the model is not calibrated. Thus, calibration seems a necessary but not sufficient condition for choosing a model for pricing and hedging.

References

- [1] Achdou, Y. (2005). An inverse problem for a parabolic variational inequality arising in volatility calibration with American options, *SIAM Journal on Control and Optimization* **43**, 1583–1615.
- [2] Achdou, Y. & Pironneau, O. (2002). Volatility smile by multilevel least square, *International Journal of Theoretical and Applied Finance* **5**(2), 619–643.
- [3] d’Aspremont, A. (2005). Risk-management methods for the Libor market model using semidefinite programming, *Journal of Computational Finance* **8**(4), 77–99.
- [4] Avellaneda, M. (1998). The minimum-entropy algorithm and related methods for calibrating asset-pricing models, *Proceedings of the International Congress of Mathematicians*, Documenta Mathematica, Berlin, Vol. III, pp. 545–563.
- [5] Avellaneda, M., Boyer-Olson, D., Busca, J. & Friz, P. (2002). Reconstructing the smile, *Risk Magazine* October.
- [6] Avellaneda, M., Buff, R., Friedman, C., Grandchamp, N., Kruk, L. & Newman, J. (2001). Weighted Monte Carlo: a new technique for calibrating asset-pricing models, *International Journal of Theoretical and Applied Finance* **4**, 91–119.
- [7] Avellaneda, M., Friedman, C., Holmes, R. & Samperi, D. (1997). Calibrating volatility surfaces via relative entropy minimization, *Applied Mathematical Finance* **4**, 37–64.
- [8] Belomestny, D. & Reiss, M. (2006). Spectral calibration of exponential Lévy Models, *Finance and Stochastics* **10**(4), 449–474.
- [9] Ben Hamida, S. & Cont, R. (2004). Recovering volatility from option prices by evolutionary optimization, *Journal of Computational Finance* **8**(3), 43–76.
- [10] Berestycki, H., Busca, J. & Florent, I. (2004). Computing the implied volatility in stochastic volatility models, *Communications on Pure and Applied Mathematics* **57**(10), 1352–1373.
- [11] Bouchouev, I., Isakov, V. & Valdivia, N. (2002). Recovering a volatility coefficient by linearization, *Quantitative Finance* **2**, 257–263.
- [12] Carr P., Geman H., Madan D.B. & Yor M. (2004). From local volatility to local Lévy models, *Quantitative Finance* **4**(5), 581–588.
- [13] Coleman, T., Li, Y. & Verma, A. (1999). Reconstructing the unknown volatility function, *Journal of Computational Finance* **2**(3), 77–102.
- [14] Cont, R. (2006). Model uncertainty and its impact on the pricing of derivative instruments, *Mathematical Finance* **16**(3), 519–547.
- [15] Cont, R. & Deguest, R. (2009). *What do index options imply about the dependence among stock returns?* Columbia University Financial Engineering Report 2009-06, www.ssrn.com.
- [16] Cont, R., Deguest, R. & Kan, Y.H. (2009). *Default Intensities Implied by CDO Spreads: Inversion Formula and Model Calibration*. Columbia University Financial Engineering Report 2009-04, www.ssrn.com.
- [17] Cont, R. & Léonard, Ch. (2008). *A Probabilistic Approach to Inverse Problems in Option Pricing*. Working Paper.
- [18] Cont, R. & Minca, A. (2008). *Recovering Portfolio Default Intensities Implied by CDO Tranches*. Financial Engineering Report 2008-01, Columbia University.
- [19] Cont, R. & Rouis, M. (2006). *Recovering Lévy Processes from Option Prices by Tikhonov Regularization*. Working Paper.
- [20] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, Chapman and Hall/CRC Press, Boca Raton.
- [21] Cont, R. & Tankov, P. (2004). Nonparametric calibration of jump-diffusion option pricing models, *Journal of Computational Finance* **7**(3), 1–49.
- [22] Cont, R. & Tankov, P. (2005). Recovering Lévy processes from option prices: regularization of an ill-posed inverse problem, *SIAM Journal on Control and Optimization* **45**(1), 1–25.
- [23] Crépey, S. (2003). Calibration of the local volatility in a trinomial tree using Tikhonov regularization, *Inverse Problems* **19**, 91–127.
- [24] Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems, *The Annals of Probability* **3**, 146–158.
- [25] Dupire, B. (1994). Pricing with a smile, *Risk* **7**, 18–20.
- [26] Engl, H. & Egger, H. (2005). Tikhonov regularization applied to the inverse problem of option pricing: convergence analysis and rates, *Inverse Problems* **21**, 1027–1045.
- [27] Engl, H.W., Hanke, M. & Neubauer, A. (1996). *Regularization of Inverse Problems*, Mathematics and its Applications, Kluwer Academic Publishers, Dordrecht, The Netherlands, Vol. 375.
- [28] Friz, P. & Gatheral, J. (2005). *Valuing Volatility Derivatives as an Inverse Problem*, *Quantitative Finance*, December 2005.
- [29] Glasserman, P. & Yu, B. (2005). Large sample properties of weighted Monte Carlo estimators, *Operations Research* **53**(2), 298–312.
- [30] Hagan, P., Kumar, D., Lesniewski, A.S. & Woodward, D.E. Managing smile risk, *Wilmott Magazine* September, 84–108.
- [31] Samperi, D. (2002). Calibrating a diffusion model with uncertain volatility, *Mathematical Finance* **12**, 71–87.

- [32] Schoutens, W., Simons, E. & Tistaert, J. (2004). A perfect calibration! Now what? *Wilmott Magazine* March.

Further Reading

- Biagini, S. & Cont, R. (2006). Model-free representation of pricing rules as conditional expectations, in *Stochastic Processes and Applications to Mathematical Finance*, J. Aka-hori, S. Ogawa and S. Watanabe, eds, World Scientific, Singapore, pp. 53–66.
- Harrison, J.M. & Pliska, S.R. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and their Applications* **11**, 215–260.

Related Articles

Black–Scholes Formula; Convex Duality; Dupire Equation; Entropy-based Estimation; Exponential Lévy Models; Implied Volatility in Stochastic Volatility Models; Implied Volatility: Large Strike Asymptotics; Jump Processes; Local Volatility Model; Markov Functional Models; SABR Model; Stochastic Volatility Models; Weighted Monte Carlo; Yield Curve Construction.

RAMA CONT

Dupire Equation

The Dupire equation is a partial differential equation (PDE) that links the contemporaneous prices of European call options of all strikes and maturities to the instantaneous volatility of the price process, assumed to be a function of price and time only. The main application of the equation is to compute (i.e., invert) local volatilities from market option prices to build a local volatility model, which many major banks currently use for option pricing.

If we assume that the price process S follows the stochastic differential equation,

$$\frac{dS_t}{S_t} = \mu_t dt + \sigma(S_t, t) dW_t \quad (1)$$

Then if $C(S, t, K, T)$ denotes the price at time t for an underlying price of S of the European call of strike K and maturity T that pays $(S_T - K)^+$ at time T , C satisfies, for a fixed (S, t) .

For a fixed (S, t) , the Dupire equation,

$$\frac{\partial C}{\partial T} = \frac{\sigma^2(K, T)}{2} K^2 \frac{\partial^2 C}{\partial K^2} - (r - q)K \frac{\partial C}{\partial K} - qC \quad (2)$$

where r is the interest rate, q is the dividend yield (or foreign interest rate in the case of a currency), and the boundary conditions are given by $C(S, t, K, T) = (S - K)^+$.

This can be established by a variety of methods, including double integration of the Fokker–Plank equation, Tanaka formula, and replication strategy. It is commonly named the *forward equation*, as it indicates how current call prices are affected by an increase in maturity. This can be contrasted with the classical backward Black–Scholes PDE that applies to a European call of fixed strike and maturity:

$$\frac{\partial C}{\partial t} = -\frac{\sigma^2(S, t)}{2} S^2 \frac{\partial^2 C}{\partial S^2} - (r - q)S \frac{\partial C}{\partial S} + rC \quad (3)$$

Interpretation

The backward Black–Scholes equation applies to a given call option and relates its time derivative to its convexity. It is a heat equation that defines the price at a given time as the discounted expectation of the

call price an instant later and the Jensen convexity bias θ depends on.

According to the forward Dupire equation, the cost of extending the maturity of a call depends on the probability of being at the strike at maturity and on the level of volatility there. It can be seen as relating the price of a calendar spread to the price of a butterfly spread.

Uses

The equation

$$\frac{\partial C}{\partial T} = \frac{\sigma^2(K, T)}{2} K^2 \frac{\partial^2 C}{\partial K^2} - (r - q)K \frac{\partial C}{\partial K} - qC \quad (4)$$

can be used in the following two ways:

1. If the local volatility $\sigma(S, t)$ is known, the PDE can be used to compute the price today of all call options in a single sweep, starting from the boundary condition $C(S, t, K, T) = (S - K)^+$. In contrast, the Black–Scholes backward equation requires one PDE for each strike and maturity.

In the case of calibrating a parametric form of $\sigma(S, t)$ to a set of market option prices, one needs to compute the model price of all these options and the forward equation can accelerate the computation to a factor 100.

2. If the call prices are known today, one can compute their derivatives and extract the local volatility by the following formula:

$$\sigma(K, T) = \sqrt{\frac{\frac{\partial C}{\partial T} + (r - q)K \frac{\partial C}{\partial K} + qC}{K^2 \frac{\partial^2 C}{\partial K^2}}} \quad (5)$$

This equation is also known as the *stripping formula*.

Starting from a finite set of listed option prices, a good interpolation in strike and maturities provides a continuum of option prices and we can apply the stripping formula to get the local volatilities. Here is an example on the NASDAQ, where interpolation/extrapolation is performed by first fitting a stochastic volatility Heston model to the listed option

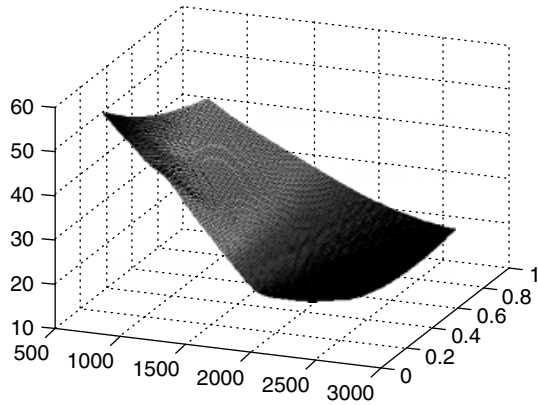


Figure 1 Implied volatility surface of the NASDAQ

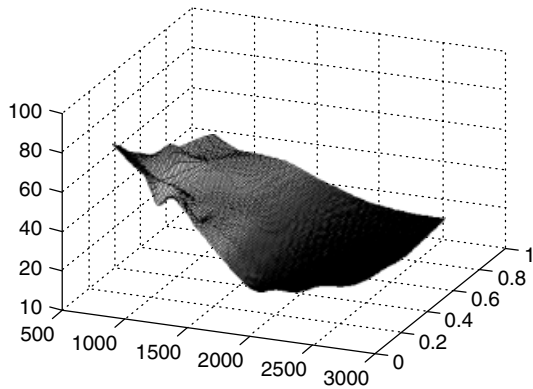


Figure 2 Local volatility surface of the NASDAQ

prices and then applying a nonparametric interpolation to the residuals.

Figure 1 displays the implied volatility surface of NASDAQ, and the associated local volatility surface is shown in Figure 2.

Once the local volatilities are obtained, one can price nonvanilla instruments with this calibrated local volatility model (Figure 3).

Properly accounting for the market skew can have a massive impact on the price of exotics. For instance, an up-and-out call option has a positive gamma close to the strike and a negative gamma close to the barrier. A typical equity negative skew corresponds to high local volatilities close to the strike, which adds value to the option due to the positive gamma and low local volatilities close to the barrier, which is also beneficial to the option holder as the gamma is negative there.

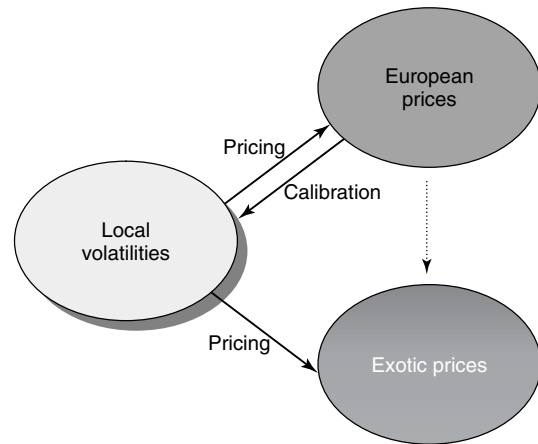


Figure 3 Local volatilities give a way to price exotic options from European options

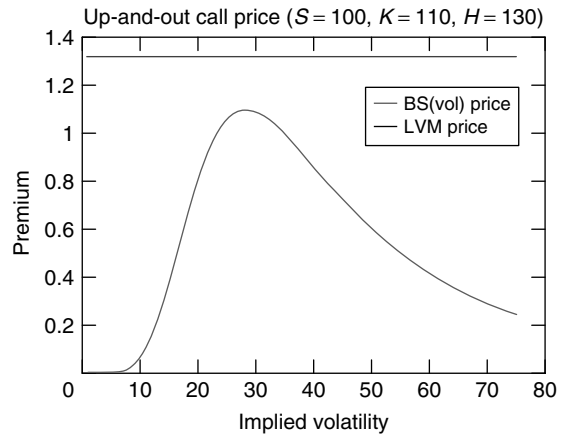


Figure 4 Comparison of an up-and-out call option in the local volatility model and in Black–Scholes model with various volatilities

The combined effect is that the up-and-out call local volatility price may exceed the price of any Black–Scholes model, irrespective of the volatility input used (Figure 4).

Local Volatilities as Forward Volatilities

The most common interpretation of local volatility is that it is the instantaneous volatility as a certain function of spot price and time that fits market prices. It gives the simplest model calibrated to the market but assumes a deterministic behavior of instantaneous

volatility, a fact crudely belied by the market. As such, the local volatility model is an important step away from the Black–Scholes model, which assumes constant volatility, though it may not necessarily provide the most realistic dynamics for the price.

The second interpretation, as forward volatilities, is far more patent. More precisely, the square of the local volatility, the local variance, is the instantaneous forward variance conditional to the spot price being equal to the strike at maturity:

$$\sigma^2(K, T) = E[\sigma_t^2 | S_T = K] \quad (6)$$

This means that in a frictionless market where all strikes and maturities are available, it is possible to combine options into a portfolio that will lock these forward values. In other words, the local variance is not only a function calibrated to the market that allows to retrieve market prices but it is also the fair value of the fixed leg of a swap with a floating leg equal to the instantaneous variance at time T , with the exchange taking place only if the price at maturity is K . It can be seen as an infinitesimal forward corridor variance swap.

By way of consequence, if one disagrees with the forward variance, one can put on a trade (in essence calendar spread against butterfly spread) aligned with this view. Conversely, if one has no view but finds someone who disagrees with the forward view and accepts to trade at a different level, one can lock the difference.

Another important consequence of this relationship is that a stochastic volatility model (with no jumps) will be calibrated to the market if and only if the conditional expectation of the instantaneous variance is the local variance computed from the market prices. In essence, it means that a calibrated stochastic volatility model is a noisy version of the local volatility model, which is centered on it. In this sense, the local volatility model plays a central role.

Beyond the fit to the current market prices, these results have dynamic consequences. For example, they imply that in the absence of jumps, the at-the-money (ATM) implied volatility converges to the instantaneous volatility when the maturity shrinks to 0. The same relation indicates that for any stochastic volatility model calibrated to the market, the average level of the short-term ATM implied variance at any

time in the future, conditioned on a price level, has to equal the local variance, which is dictated by the current market prices of calls and puts. Fitting to today's market strongly constrains future dynamics and, for instance, the backbone, defined as the behavior of the at-the-money volatility as a function of the underlying price, cannot be independently specified.

Once we get a perfect fit of option prices using equation (5), we can perturb the volatility surface, recalibrate, and conduct a sensitivity analysis. This provides a decomposition of the volatility risk of any structured product (or portfolio of) across strikes and maturities, because seeing the price as a function of the whole volatility surface provides through perturbation analysis the sensitivity to all volatilities.

Extensions

There are numerous extensions of the forward PDE, with stochastic rates and dividends, stochastic volatility, jumps, to the Greeks (sensitivities) and to other products than European options, such as barrier options, compound options, Asian options, and basket options. However, until now, there is no satisfactory counterpart for American options.

Further Reading

- Derman, E. & Kani, I. (1994). Riding on a smile, *Risk* 7(2), 32–39, 139–145.
- Dupire, B. (1993). Model art, *Risk* 6(9), 118–124.
- Dupire, B. (1994). Pricing with a smile, *Risk* 7, 18–20.
- Dupire, B. (1997). Pricing and hedging with smiles, in *Mathematics of Derivative Securities*, M.A.H. Dempster & S.R. Pliska, eds, Cambridge University Press.
- Dupire, B. (2004). A unified theory of volatility, working paper Paribas capital markets 1996, reprinted in *Derivatives Pricing: The Classic Collection*, P. Carr, ed., Risk Books, London.

Related Articles

Implied Volatility Surface; Local Times; Local Volatility Model; Markov Processes; Model Calibration.

BRUNO DUPIRE

Implied Volatility Surface

The widespread practice of quoting option prices in terms of their Black–Scholes implied volatilities (IVs) in no way implies that market participants believe underlying returns to be lognormal. On the contrary, the variation of IVs across option strike and term to maturity, which is widely referred to as the *volatility surface*, can be substantial. In this article, we highlight some empirical observations that are most relevant for the construction and validation of realistic models of the volatility surface for equity indices.

The Shape of the Volatility Surface

Ever since the 1987 stock market crash, volatility surfaces for global indices have been characterized by the *volatility skew*: For a given expiration date, IVs increase as strike price decreases for strikes below the current stock price (spot) or current forward price. This tendency can be seen clearly in the S&P500 volatility surface shown in Figure 1. For short-dated expirations, the cross section of IVs as a function of strike is roughly V-shaped, but has a rounded vertex and is slightly tilted. Generally, this V-shape softens and becomes flatter for longer dated expirations, but the vertex itself may rise or fall depending on whether the term structure (TS) of (ATM) At-the-money volatility is upward or downward sloping.

Conventional explanations for the volatility skew include the following:

- *The leverage effect*: Stocks tend to be more volatile at lower prices than at higher prices.
- Volatility moves and spot moves are anticorrelated.
- Big jumps in spot tend to be downward rather than upward.
- *The risk of default*: There is a nonzero probability for the price of a stock to collapse if the issuer defaults.
- *Supply and demand*: Investors are net long of stock and so tend to be net buyers of downside puts and sellers of upside calls.

The volatility skew probably reflects all of these factors.

Conventional stochastic volatility (SV) models imply a relationship between the assumed dynamics of the instantaneous volatility and the volatility skew (see Chapter 8 of [8]). Empirically, volatility is well known to be roughly lognormally distributed [1, 4] and in this case, the derivative of IV with respect to log-strike in an SV model is approximately independent of volatility [8]. This motivates a simple measure of skew: For a given term to expiration, the “95–105” skew is simply the difference between the IVs at strikes of 95% and 105% of the forward price. Figure 2 shows the historical variation of this measure as a function of term to expiration as calculated from end-of-day SPX volatility surfaces generated from listed options prices between January 2, 2001 to February 6, 2009. To fairly compare across different dates and over all volatility levels, all volatilities for a given date are scaled uniformly to ensure that the one-year at-the-money-forward (ATMF) volatility equals its historical median value over this period (18.80%). The skews for all listed expirations are binned by their term to expiration; the median value for each five-day bin is plotted along with fits to both $1/\sqrt{T}$ and the best-fitting power-law dependence on T .

The important conclusion to draw here is that the TS of skew is approximately consistent with square-root (or at least power-law) decay. Moreover, this rough relationship continues to hold for longer expirations that are typically traded (OTC) Over-the-counter.

Significantly, this empirically observed TS of the volatility skew is inconsistent with the $1/T$ dependence for longer expirations typical of popular one-factor SV models (see Chapter 7 of [8] for example): Jumps affect only short-term volatility skews, so adding jumps does not resolve this disagreement between theory and observation. Introducing more volatility factors with different timescales [3] does help but does not entirely eliminate the problem. Market models of IV (*see Implied Volatility: Market Models*) obviously fit the TS of skew by construction, but such models are, in general, time inhomogeneous and, in any case, have not so far proven to be tractable. In summary, fitting the TS of skew remains an important and elusive benchmark by which to gauge models of the volatility surface.

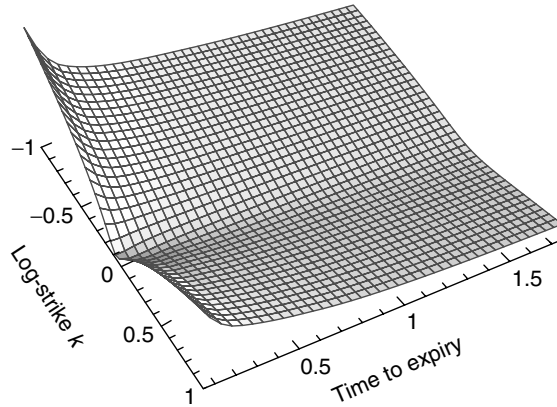


Figure 1 Graph of the S&P500-implied volatility surface as of the close on September 15, 2005, the day before triple witching

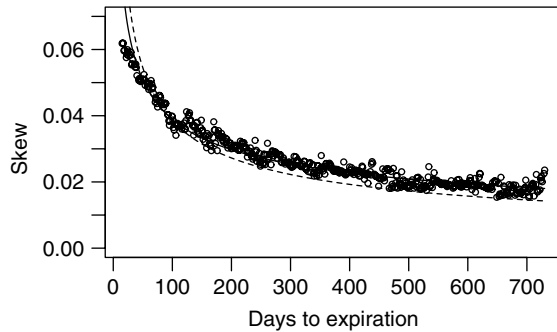


Figure 2 Decay of S&P500 95–105% skew with respect to term to expiration. Dots show the median value for each five-day bin and lines represent best fit to $1/\sqrt{T}$ (dashed) and to power-law behavior: $T^{-0.39}$ (solid)

Volatility Surface Dynamics

Volatility surfaces cannot have arbitrary shape; they are constrained by no-arbitrage conditions (such as the convexity of price with respect to strike). In practice, these restrictions are not onerous and generally are met provided there are no large gradients anywhere on the surface. This observation, together with the fact that index options in most markets trade actively over a wide range of strikes and expiration dates, might lead one to expect the dynamics of the volatility surface to be quite complicated. On the contrary, several principal component analysis (PCA) studies have found that an overwhelming fraction of the total daily variation in volatility surfaces is explained by just a few factors, typically three.

Table 1 makes clear that a level mode, one where the entire volatility surface shifts upward or downward in tandem, accounts for the vast majority of variation; this result holds across different markets, historical periods, sampling frequencies, and statistical methodologies. Although the details vary, generally this mode is not quite flat; short-term volatilities tend to move more than longer ones, as evidenced by the slightly upward tilt in Figure 3(a).

In most of the studies, a TS mode is the next most important mode: Here, short-term volatilities move in the opposite direction from longer term ones, with little variation across strikes. In the Merrill Lynch data, which are sampled at 30, 91, 182, 273, 365, and 547 days to expiration, the pivot point is close to the 91-day term (see Figure 3(b)). In all studies where TS is the second most important mode, the third mode is always a skew mode: one where strikes below the spot (or forward) move in the opposite direction

Table 1 PCA studies of the volatility surface. GS, Goldman Sachs study [9] and ML, Merrill Lynch proprietary data

Source	Market	Top 3 modes	Var. explained by		Correlation of 3 modes with spot
			First mode (%)	Top 3 (%)	
GS	S&P500, weekly, 1994–1997	Level, TS, skew	81.6	90.7	−0.61, −0.07, 0.07
GS	Nikkei, daily, 1994–1997	Level, TS, skew	85.6	95.9	−0.67, −0.05, 0.04
Cont <i>et al.</i>	S&P500, daily, 1900–2001	Level, skew, curvature	94	97.8	−0.66, ~0, 0.27
Cont <i>et al.</i>	FTSE100, daily, 1999–2001	Level, skew, curvature	96	98.8	−0.70, 0.08, 0.7
Daglish <i>et al.</i>	S&P500, monthly, 1998–2002	Level, TS, skew	92.6	99.3	n.a.
ML	S&P500, daily, 1901–2009	Level, TS, skew	95.3	98.2	−0.87, −0.11, ~0

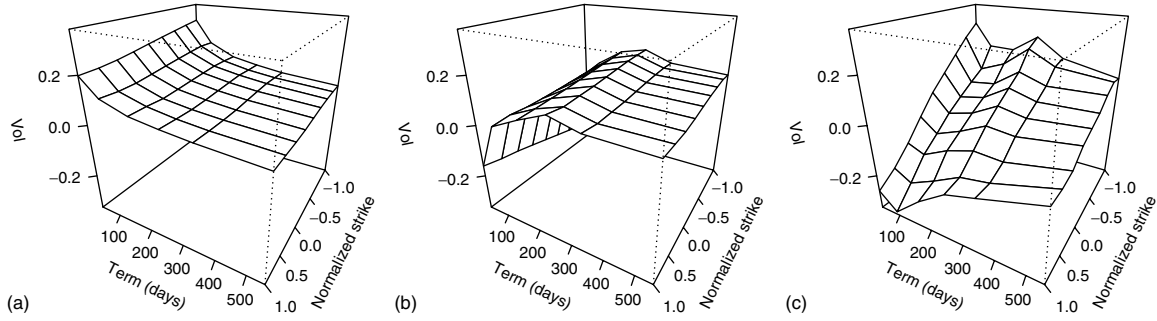


Figure 3 PCA modes for Merrill Lynch S&P500 volatility surfaces: (a) level; (b) term structure; and (c) skew

from those above and where the overall magnitude is attenuated as term increases (Figure 3c).

It is also worth noting that the two studies [5, 9] that looked at two different markets during comparable periods found very similar patterns of variation; the modes and their relative importance were very similar, suggesting strong global correlation across index volatility markets.

In the study by Cont and da Fonseca [5], a TS mode does not figure in the top three modes; instead a skew mode and another strike-related mode related to the curvature emerge as number two and three. This likely reflects the atypically low variation in TS over the historical sample period and is not due to any methodological differences with the other studies. As in the other studies, the patterns of variation were very similar across markets (S&P500 and FTSE100).

Changes in Spot and Volatility are Negatively Correlated

Perhaps the sturdiest empirical observation of all is simply that changes in spot and changes in volatility (by pretty much any measure) are negatively and strongly correlated. From the results that we have surveyed here, this can be inferred from the high R^2 obtained in the regressions of daily ATMF volatility changes shown in Table 1, as well as directly from the correlations between spot return and PCA modes shown in Table 2. It is striking that the correlation between the level mode and spot return is consistently high across studies, ranging from -0.66 to -0.87 . Correlation between the spot return and the other modes is significantly weaker and less stable.

This high correlation between spot returns and changes in volatility persists even in the most extreme

Table 2 Historical estimates of β_T

T	β_T standard error	R^2
30	1.55 (0.02)	0.774
91	1.50 (0.02)	0.825
182	1.48 (0.02)	0.818
365	1.49 (0.02)	0.791

market conditions. For example, during the turbulent period following the collapse of Lehman Brothers in September 2008, which was characterized by both high volatility and high volatility of volatility, spot-volatility correlation remained at historically high levels: -0.92 for daily changes between September 15, 2008 and end December 31, 2008. On the other hand, the skew mode, which is essentially uncorrelated with spot return in the full historical period (see Table 1), did exhibit stronger correlation in this period (-0.55), while the TS mode did not. These observations underscore the robustness of the level-spot correlation as well as the time-varying nature of correlations between spot returns and the other modes of fluctuation of the volatility surface.

Other studies have also commented on the robustness of the spot-volatility correlation. For example, using a maximum-likelihood technique, Aït-Sahalia and Kimmel [1] carefully estimated the parameters of The Heston, CEV, and GARCH models from S&P500 and VIX data between January 2, 1990 and September 30, 2003; the correlation between spot and volatility changes varied little between these models and various estimation techniques, and all estimates were around -0.76 for the period studied.

A related question that was studied by Bouchaud *et al.* [4] is whether spot changes drive *realized*

volatility changes or *vice versa*. By computing the correlations of leading and lagging returns and squared-returns, they find that for both stocks and stock indices, price changes lead to volatility changes. In particular, there is no volatility feedback effect, whereby changes in volatility affect future stock prices. Moreover, unlike the decay of the IV correlation function itself, which is power-law with an exponent of around 0.3 for SPX, the decay of the spot-volatility correlation function is exponential with a short half-life of a few days. Supposing the general level of IV (the variation of which accounts for most of the variation of the volatility surface) to be highly correlated with realized volatility, these results also apply to the dynamics of the IV surface. Under diffusion assumptions, the relationship between implied and realized volatility is even more direct: Instantaneous volatility is given by the IV of the ATM option with zero time to expiration.

Skew Relates Statics to Dynamics

Volatility changes are related to changes in spot: as mentioned earlier, volatility and spot tend to move in opposite directions, and large moves in volatility tend to follow large moves in the spot.

It is reasonable to expect skew to play a role in relating the magnitudes of these changes. For example, if all the variation in ATM volatility were explained simply by movement along a surface that is unchanged as a function of strike when spot changes, then we would expect

$$\Delta\sigma_{\text{ATMF}}(T) = \beta_T \frac{d\sigma}{d(\log K)} \frac{\Delta S}{S} \quad (1)$$

with $\beta_T = 1$ for all terms to expiration (T).

The empirical estimates of β_T shown in Table 2 are based on the daily changes in S&P500 ATM volatility from January 2, 2001 to February 6, 2009 (volatilities tied to fixed expiration dates are interpolated to arrive at volatilities for a fixed number of days to expiration.) Two important conclusions may be drawn: (i) β is not 1.0, rather it is closer to 1.5 and (ii) remarkably β_T does not change appreciably by expiration. In other words, although the volatility skew systematically underestimates the daily change in volatility, it does so by roughly the same factor for all maturities. It is also worth noting that the hypothesis $\beta_T = 1$ would be rejected even if the

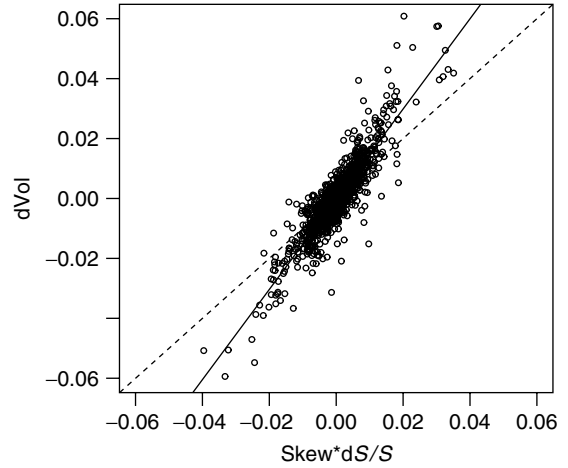


Figure 4 Regression of 91-day volatility changes *versus* spot returns. A zero-intercept least squares fit to model (1) leads to $\beta_{91} = 1.50$ (solid lines). The $\beta = 1$ (“sticky-strike”) prediction (dashed line) clearly does not fit

regression were restricted to spot returns of smaller magnitude, as suggested visually by the scatterplots of Figure 4.

Although empirical relationships between changes in ATM volatility and changes in spot are clearly relevant to volatility trading and risk management, the magnitude of β_T itself has direct implications for volatility modeling as well. In both local and SV models, $\beta_T \rightarrow 2$ in the short-expiration limit. Under SV, β_T is typically a decreasing function of T , whereas under the local volatility assumption where the local volatility surface is fixed with respect to a given level of the underlying, β_T is typically an increasing function of T .

Market participants often adopt a phenomenological approach and characterize surface dynamics as following one of these rules: “sticky strike,” “sticky delta,” or “local volatility”; each rule has an associated value of β_T . Under the sticky-strike assumption, $\beta_T = 1$ and the volatility surface is fixed by strike; under the sticky-delta assumption, $\beta_T = 0$ and the volatility surface is a fixed function of K/S ; and under the local volatility assumption, as mentioned earlier, $\beta_T = 2$ for short expirations.

Neither “sticky-strike” nor “sticky-delta” rules imply reasonable dynamics [2]: In a sticky-delta model, the log of the spot has independent increments and the only arbitrage-free sticky-strike model is Black–Scholes (where there is no smile).

Although the estimates of β_T in Table 2 are all around 1.5, consistent with SV, this does not exclude the possibility that there may be periods where the β_T may substantially depart from these average values. Derman [7] identified seven distinct regimes for S&P500 daily volatility changes between September 1997 and November 1998, finding evidence for all three of the alternatives listed above. A subsequent study [6] looked at S&P500 monthly data between June 1998 and April 2002 (47 points) and found that for that period, the data were much more consistent with the sticky-delta rule than with the sticky-strike rule.

References

- [1] Aït-Sahalia, Y. & Kimmel, R. (2007). Maximum likelihood estimation of stochastic volatility models, *Journal of Financial Economics* **83**, 413–452.
- [2] Balland, P. (2002). Deterministic implied volatility models, *Quantitative Finance* **2**, 31–44.
- [3] Bergomi, L. (2008). Smile dynamics III, *Risk* **21**, 90–96.
- [4] Bouchaud, J.-P. & Potters, M. (2003). *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*, Cambridge University Press, Cambridge.
- [5] Cont, R. & Fonseca, J. (2002). Dynamics of implied volatility surfaces, *Quantitative Finance* **2**, 45–60.
- [6] Daglish, T., Hull, J. & Suo, W. (2007). Volatility surfaces: theory, rules of thumb, and empirical evidence, *Quantitative Finance* **7**, 507–524.
- [7] Derman, E. (1999). Regimes of Volatility, *Risk* **12**, 55–59.
- [8] Gatheral, J. (2006). *The Volatility Surface*, John Wiley & Sons, Hoboken.
- [9] Kamal, M. & Derman, E. (1997). The patterns of change in implied index volatilities, in *Goldman Sachs Quantitative Research Notes*, Goldman Sachs, New York.

Related Articles

Black–Scholes Formula; Implied Volatility in Stochastic Volatility Models; Implied Volatility: Large Strike Asymptotics; Implied Volatility: Long Maturity Behavior; Implied Volatility: Market Models; SABR Model.

MICHAEL KAMAL & JIM GATHERAL

Moment Explosions

Let $(S_t, V_t)_{t \geq 0}$ be a Markov process, representing a (not necessarily purely continuous) stochastic volatility model. $(S_t)_{t \geq 0}$ is the (discounted) price of a traded asset, such as a stock, and $(V_t)_{t \geq 0}$ represents a latent factor, such as stochastic volatility, stochastic variance, or the stochastic arrival rate of jumps. A **moment explosion** takes place, if the moment $\mathbb{E}[S_t^u]$ of some given order $u \in \mathbb{R}$ becomes infinite (“explodes”) after some finite time $T_*(u)$. This time is called the *time of moment explosion* and formally defined by

$$T_*(u) = \sup \{t \geq 0 : \mathbb{E}[S_t^u] < \infty\} \quad (1)$$

We say that no moment explosion takes place for some given order u , if $T_*(u) = \infty$.

Moment explosions can be considered both under the physical and the pricing measure, with most applications belonging to the latter. If $(S_t)_{t \geq 0}$ is a martingale, then Jensen’s inequality implies that moment explosions can only occur for moments of order $u \in \mathbb{R} \setminus [0, 1]$.

Conceptually, the notion of a moment explosion has to be distinguished from an explosion of the process itself, which refers to the situation that the process $(S_t)_{t \geq 0}$, not one of its moments, becomes infinite with some positive probability.

Applications

In *equity* and *foreign exchange* models, where $(S_t)_{t \geq 0}$ represents a stock price or an exchange rate, moment explosions are closely related to the shape of the implied volatility surface, and can be used to obtain approximations for the implied volatility of deep in-the-money and out-of-the-money options (see **Implied Volatility: Large Strike Asymptotics**, and the references therein). According to [5, 14], the asymptotic shape of the implied volatility surface for some fixed maturity T is determined by the smallest and largest moment of S_T that is still finite. These critical moments $u_-(T)$ and $u_+(T)$ are the piecewise inverse functions^a of the moment explosion time. Often the explosion time is easier to calculate, so a feasible approach is to first calculate explosion times, and then to invert to obtain the critical moments. Let

us note that finite critical moments of the underlying S_T correspond, in essence, to exponential tails of $\log(S_T)$. There is evidence that refined knowledge of how moment explosion occurs (or the asymptotic behavior of $u \mapsto \mathbb{E}[S_T^u]$ in the case of nonexplosion) can lead to refined results about implied volatility, see [6, 11] for some examples of stochastic alpha beta rho (SABR) type.

In *fixed-income* markets $(S_t)_{t \geq 0}$ might represent a forward LIBOR rate or swap rate. Andersen and Piterbarg [2] give examples of derivatives with super-linear payoff, whose pricing involves calculation of the second moment of S_T . It is clear that an explosion of the second moment will lead to infinite prices of such derivatives.

For *numerical procedures*, such as discretization schemes for stochastic differential equations (SDEs), error estimates that depend on higher order moments of the approximated process may break down if moment explosions occur [1]. Moment explosions may also lead to *infinite expected utility* in utility maximization problems [12].

Moment Explosions in the Black–Scholes and Exponential Lévy Models

In the Black–Scholes model, moment explosions never occur, since moments of all orders exist for all times. In an exponential Lévy model (see **Exponential Lévy Models**), S_t is given by $S_t = S_0 \exp(X_t)$, where X_t is a Lévy process. It holds that $\mathbb{E}[S_t^u] = e^{t\kappa(u)}$, where $\kappa(u)$ is the cumulant-generating function (cgf) of X_1 . Thus in an exponential Lévy model, the time of moment explosion is given by

$$T_*(u) = \begin{cases} +\infty & \kappa(u) < \infty \\ 0 & \kappa(u) = \infty \end{cases} \quad (2)$$

Let us remark that, from Theorem 25.3 in [16], $\kappa(u) < \infty$ iff $\int e^{ux} 1_{|x|>1} \nu(dx) < \infty$ where $\nu(dx)$ denotes the Lévy measure of X .

Moment Explosions in the Heston Model

The situation becomes more interesting in a stochastic volatility model, like the Heston model (see **Heston**

Model):

$$\begin{aligned} dS_t &= S_t \sqrt{V_t} dW_t^1, \quad S_0 = s \\ dV_t &= -\lambda(V_t - \theta) dt + \eta \sqrt{V_t} dW_t^2, \\ V_0 &= v, \quad \langle dW_t^1, dW_t^2 \rangle = \rho dt \end{aligned} \quad (3)$$

We now discuss how to compute the moments of S_t (equivalently, the moment-generating function of $X_t = \log S_t/S_0$). The joint process $(X_t, V_t)_{t \geq 0}$ is a (time-homogenous) diffusion, started at $(0, v)$, with generator

$$\begin{aligned} \mathcal{L} &= \frac{v}{2} \left(\frac{\partial^2}{\partial x^2} - \frac{\partial}{\partial x} \right) + \frac{v}{2} \eta^2 \frac{\partial^2}{\partial v^2} \\ &\quad + \lambda(\theta - v) \frac{\partial}{\partial v} + \rho \eta v \frac{\partial^2}{\partial x \partial v} \end{aligned} \quad (4)$$

Note that $(X_t, V_t)_{t \geq 0}$ has affine structure in the sense that the coefficients of $\mathcal{L}$ are affine linear in the state variables.^b

Now

$$\begin{aligned} \mathbb{E} [e^{uX_t} | X_t = x, V_t = v] \\ = e^{ux} \mathbb{E} [e^{uX_t} | X_t = 0, V_t = v] \end{aligned} \quad (5)$$

$$T_*^{\text{Heston}}(u) = \begin{cases} +\infty & \text{if } \Delta(u) \geq 0, \chi(u) < 0 \\ \frac{1}{\sqrt{\Delta(u)}} \log \left(\frac{\chi(u) + \sqrt{\Delta(u)}}{\chi(u) - \sqrt{\Delta(u)}} \right) & \text{if } \Delta(u) \geq 0, \chi(u) > 0 \\ \frac{2}{\sqrt{-\Delta(u)}} \left(\arctan \frac{\sqrt{-\Delta(u)}}{\chi(u)} + \pi \mathbf{1}_{\{\chi(u) < 0\}} \right) & \text{if } \Delta(u) < 0 \end{cases} \quad (10)$$

satisfies, as a function of (t, x, v) , the backward equation $\partial_t + \mathcal{L} = 0$ with terminal data e^{ux} and after replacing $T - t$ with t we can rewrite this as an initial value problem. Indeed, setting $f = f(t, v; u) := \mathbb{E} [e^{uX_t} | X_0 = 0, V_0 = v]$, and noting that

$$\begin{aligned} \left(\frac{\partial^2}{\partial x^2} - \frac{\partial}{\partial x} \right) (e^{ux} f) &= e^{ux} (u^2 - u) f \text{ and} \\ \frac{\partial^2}{\partial x \partial v} (e^{ux} f) &= e^{ux} u \frac{\partial}{\partial v} f \end{aligned} \quad (6)$$

we see that f satisfies a parabolic partial differential equation (PDE).

$$\begin{aligned} \partial_t f &= \mathcal{A}f := \left(\frac{v}{2} \eta^2 \frac{\partial^2}{\partial v^2} + [\lambda(\theta - v) + \rho \eta u v] \frac{\partial}{\partial v} \right. \\ &\quad \left. + \frac{v}{2} (u^2 - u) \right) f \end{aligned} \quad (7)$$

with initial condition $f(0, \cdot; u) = 1$, in which (again) all coefficients depend in an affine-linear way on v . The exponentially affine ansatz $f(t, v; u) = \exp(\phi(t, u) + v\psi(t, u))$ then immediately reduces this PDE to a system of ordinary differential equations (ODEs) for $\phi(t, u)$ and $\psi(t, u)$:

$$\frac{\partial}{\partial t} \phi(t, u) = F(u, \psi(t, u)), \quad \phi(0, u) = 0 \quad (8)$$

$$\frac{\partial}{\partial t} \psi(t, u) = R(u, \psi(t, u)), \quad \psi(0, u) = 0 \quad (9)$$

where $F(u, w) = \lambda \theta w$ and $R(u, w) = \frac{w^2}{2} \eta^2 + (\rho \eta u - \lambda)w + \frac{1}{2}(u^2 - u)$. Equation (9) is a Riccati differential equation, whose solution blows up at finite time, corresponding to the moment explosion of S_t . Explicit calculations ([2], for instance) yield^c

where $\chi(u) = \rho \eta u - \lambda$ and $\Delta(u) = \chi(u)^2 - \eta^2(u^2 - u)$. A simple analysis of this condition (cf. [2]) then allows to express the no-explosion condition in terms of the correlation parameter ρ . With focus on positive moments of the underlying, $u \geq 1$, we have

$$T_*^{\text{Heston}}(u) = +\infty \iff \rho \leq -\sqrt{\frac{u-1}{u}} + \frac{\lambda}{\eta u} \quad (11)$$

Similar results for a class of nonaffine stochastic volatility models is discussed below.

Moment Explosions in Time-changed Exponential Lévy Models

Stochastic volatility can also be introduced in the sense of running time at a stochastic “business” clock. For instance, when $\rho = 0$ the (log-price) in the Heston model is a Brownian motion with drift, $W_t - t/2$, run at a Cox–Ingersoll–Ross^d (CIR) clock $\tau(t, \omega) = \tau_t$ where

$$dV_t = -\lambda(V_t - \theta) dt + \eta\sqrt{V_t} dW_t, \quad V_0 = v \quad (12)$$

$$d\tau_t = V dt, \quad \tau_0 = 0 \quad (13)$$

Since (V, τ) has an affine structure, there is a tractable moment-generating/characteristic function in the form

$$\begin{aligned} \mathbb{E}(\exp(u\tau_T)) &= \mathbb{E}\left(\exp\left(u \int_0^T V(t, \omega) dt\right)\right) \\ &= \exp(A(u, T) + vB(u, T)) \end{aligned} \quad (14)$$

where^e

$$\begin{aligned} A(u, t) &= \lambda^2 \theta t / \eta^2 - \frac{2\lambda\theta}{\eta^2} \log \\ &\quad \times \left[\sinh(\gamma t/2) \cdot \left(\coth(\gamma t/2) + \frac{\lambda}{\gamma} \right) \right] \end{aligned} \quad (15)$$

$$\begin{aligned} B(u, t) &= 2u/(\lambda + \gamma \coth(\gamma t/2)), \quad \gamma = \sqrt{\lambda^2 - 2\eta^2 u} \end{aligned} \quad (16)$$

We can replace $W_t - t/2$ above by a general Lévy process $L = L_t$ and run it again at some independent clock $\tau = \tau(t, \omega)$, assuming only knowledge of the cgf $\kappa_T(u) = \log \mathbb{E}(\exp(u\tau_T))$. If we also set $\kappa_L(u) = \log \mathbb{E}[\exp(uL_1)]$, a simple conditioning argument shows that the moment-generating function of L_τ is given by

$$\begin{aligned} M(u) &= \mathbb{E}[\mathbb{E}(e^{uL_\tau} | \tau)] = \mathbb{E}[e^{\kappa_L(u)\tau}] \\ &= \exp[\kappa_\tau(\kappa_L(u))] \end{aligned} \quad (17)$$

From here on, moment explosions of L_τ can be investigated analytically, provided κ_τ, κ_L are known in sufficiently explicit form. For some computations in this context, also with regard to the asymptotic behavior of the implied volatility smile, see [5].

Moment Explosions in Non-affine Diffusion Models

Both [2] and [15] study existence of u th moments, $u \geq 1$, for (not necessarily affine) diffusion models of the type

$$dS_t = V_t^\delta S_t^\beta dW_t^1, \quad S_0 = s \quad (18)$$

$$\begin{aligned} dV_t &= \eta V_t^\gamma dW_t^2 + b(V_t) dt, \\ V_0 &= v, \quad \langle dW_t^1, dW_t^2 \rangle = \rho dt \end{aligned} \quad (19)$$

where $\delta, \gamma > 0, \beta \in [0, 1]$ and the function $b(v)$ are subject to suitable conditions that ensure a unique solution. For instance, the SABR model falls into this class. Lions and Musiela [15] first show that if $\beta < 1$, no moment explosions occur. For $\beta = 1$, the same reasoning as in the Heston model shows that $f(t, v; u) = \mathbb{E}[(S_t/s)^u]$ satisfies the PDE^f

$$\begin{aligned} \frac{\partial}{\partial t} f &= \mathcal{A}f := \frac{v^{2\gamma}}{2} \eta^2 \frac{\partial^2 f}{\partial v^2} + [b(v) + \eta \rho u v^{\delta+\gamma}] \frac{\partial f}{\partial v} \\ &\quad + \frac{v^{2\delta}}{2} (u^2 - u) f \end{aligned} \quad (20)$$

with initial condition $f(0, \cdot; u) \equiv 1$. Note that the Heston model is recovered as the special case $\beta = 1, \delta = \gamma = 1/2, b(v) = -\lambda(v - \theta)$. Using the (exponentially-affine in v^q) ansatz $f(t, v; u) = \exp(\phi(t, u) + v^q \psi(t, u))$, with suitably chosen q, ϕ , and ψ , Lions and Musiela [15] construct supersolutions of equation (20), leading to lower bounds for $T_*(u)$, and then subsolutions, leading to matching upper bounds.^g We report the following results from [15]:

1. $\beta < 1$: no moment explosion occurs, that is, $\mathbb{E}(S_t^u) < \infty$ for all $u \geq 1, t \geq 0$;
2. $\beta = 1, \gamma + \delta < 1$: as in 1. no moment explosion occurs;
3. $\beta = 1, \gamma + \delta = 1$: If $\gamma = \delta = \frac{1}{2}$, then this choice of parameters yields a Heston-type model, where

the mean-reversion term $-\lambda(V_t - \theta)dt$ has been replaced by the more general $b(V_t)dt$. With λ replaced by $\lim_{v \rightarrow \infty} -b(v)/v$ the formula (10) remains valid. If $\gamma \neq \delta$, then the model can be transformed into a Heston-like model by the change of variables $\tilde{V}_t := V_t^{2\delta}$. The time of moment explosion $T_*(u)$ can be related to the expression in equation (10), by

$$T_*(u) = \frac{1}{2\delta} T_*^{\text{Heston}}(u) \quad (21)$$

4. $\beta = 1, \gamma + \delta > 1$: Let $b_\infty = \lim_{v \rightarrow \infty} b(v)/v^{\delta+\gamma}$, and $\rho^*(u) = -\sqrt{(u-1)/u} - b_\infty/(\eta u)$, then

$$T_*(u) = \begin{cases} +\infty & \rho < \rho^*(u) \\ 0 & \rho > \rho^*(u) \end{cases} \quad (22)$$

The borderline case $\rho = \rho^*(u)$ is delicate and we refer to [15, page 13]. Observe that, the condition on $\rho < \rho^*(u)$ is consistent with the Heston model (11), upon setting $\gamma = \delta = 1/2, b_\infty = -\lambda$, whereas the behavior of $\rho > \rho^*(u)$ is different in the sense that there is no immediate moment explosion in the Heston model.

Moment Explosions in Affine Models with Jumps

Recall that in the Heston model

$$\begin{aligned} \mathbb{E}[e^{uX_t} | X_0 = x, V_0 = v] \\ = e^{ux} \exp(\phi(t, u) + v\psi(t, u)) \end{aligned} \quad (23)$$

and it was this form of exponentially affine dependence on x, v that allowed an analytical treatment *via* Riccati equations. Assuming validity only of equation (23), for all $u \in \mathbb{C}$ for which the expectation exists, and that $(X_t, V_t)_{t \geq 0}$ is a (stochastically continuous, time-homogenous) Markov process on $\mathbb{R} \times \mathbb{R}_{\geq 0}$ puts us in the framework of affine processes [8], which, in fact, includes the bulk of analytically tractable stochastic volatility models with and without jumps.

The infinitesimal generator $\mathcal{L}$ of the process $(X_t, V_t)_{t \geq 0}$ now includes integral terms corresponding to the jump effects and thus is a partial integro-differential operator. Nevertheless, the exponentially

affine ansatz $f(t, v; u) = \exp(\phi(t, u) + v\psi(t, u))$ still reduces the Kolmogorov equation to ordinary differential equations of the type equation (8). The functions $F(u, w)$ and $R(u, w)$ are no longer quadratic polynomials, but of Lévy–Khintchine form (*see Infinite Divisibility*). The time of moment explosion can be determined by calculating the blow-up time for the solutions of these *generalized* Riccati equations. This approach can be applied to a Heston model with an additional jump term:

$$\begin{aligned} dX_t &= \left(c(V_t) - \frac{V_t}{2} \right) dt + \sqrt{V_t} dW_t^1 + dJ_t(V_t), \\ X_0 &= 0 \end{aligned} \quad (24)$$

$$\begin{aligned} dV_t &= -\lambda(V_t - \theta) dt + \eta\sqrt{V_t} dW_t^2, \\ V_0 &= v, \quad \langle dW_t^1, dW_t^2 \rangle = \rho dt \end{aligned} \quad (25)$$

The process $J_t(V_t)$ is a pure-jump process based on a fixed Lévy measure $\nu(dx)$. More precisely, writing μ for the uncompensated and $\tilde{\mu}$ for the compensated Poisson random measure, independent of (W_t^1, W_t^2) , with intensity $\nu(dx) \otimes dt$, we assume that

$$\begin{aligned} dJ_t(V_t) &= \begin{cases} \int_{|x| < 1} x \tilde{\mu}(dx, dt) & \dots \text{ case (a)} \\ + \int_{|x| \geq 1} x \mu(dx, dt) & \\ \int_{|x| < 1} V_t x \tilde{\mu}(dx, dt) & \dots \text{ case (b)} \\ + \int_{|x| \geq 1} V_t x \mu(dx, dt) & \end{cases} \end{aligned} \quad (26)$$

In case (a), the process J_t is a genuine (pure-jump) Lévy process; in case (b) jumps are amplified linearly with the variance level, as proposed by Bates [4]. We focus first on case (a). Assuming $\mathbb{E}[e^{J_t}] < \infty$, or equivalently $\int e^x 1_{|x| \geq 1} \nu(dx) < \infty$, so that $e^{\tilde{J}_t} := e^{J_t + ct}$ is a martingale for suitable drift, $c = -\log \mathbb{E}[e^{J_1}]$, we have^h

$$\begin{aligned} \mathbb{E}[e^{uX_t}] &= \mathbb{E}[e^{uX_t^{\text{Heston}}}] \mathbb{E}[e^{u\tilde{J}_t}] \\ &= \mathbb{E}[e^{uX_t^{\text{Heston}}}] e^{t\tilde{\kappa}(u)} \end{aligned} \quad (27)$$

Here, $\tilde{\kappa}(u) = \int (e^{ux} - 1 - u(e^x - 1)) \nu(dx)$ is well defined with values in $(-\infty, \infty]$ and finiteness of $\tilde{\kappa}(u) < \infty$ is tantamount to $\int e^{ux} 1_{|x| \geq 1} \nu(dx) < \infty$. Hence in case (a), we can link the time of moment explosion $T_*(u)$ to $T_*^{\text{Heston}}(u)$, given by equation (10), and have

$$T_*(u) = \begin{cases} T_*^{\text{Heston}}(u) & \tilde{\kappa}(u) < \infty \\ 0 & \tilde{\kappa}(u) = \infty \end{cases} \quad (28)$$

In the case (b), the jump process $J_t(V_t)$ depends on V_t and the above argument cannot be used. A direct analysis of the (generalized) Riccati equations [13] shows that in the case $\tilde{\kappa}(u) < \infty$ the time of moment explosion is given by formula (10), only now $\Delta(u) = \chi(u)^2 - \eta^2(2\tilde{\kappa}(u) + u^2 - u)$, and immediate moment explosion happens in the case $\tilde{\kappa}(u) = \infty$.

Also the model introduced by Barndorff-Nielsen and Shephard [3] (see **Barndorff-Nielsen and Shephard (BNS) Models**), which features simultaneous jumps in price and variance, falls into the class of affine models. It is given by

$$dX_t = \left(c - \frac{V_t}{2}\right) dt + \sqrt{V_t} dW_t + \rho dJ_{\lambda t}, \quad X_0 = 0 \quad (29)$$

$$dV_t = -\lambda V_t dt + dJ_{\lambda t}, \quad V_0 = v, \quad (30)$$

where $\lambda > 0$, $\rho < 0$ and $(J_t)_{t \geq 0}$ is a pure-jump Lévy process with positive jumps only, and with Lévy measure $\nu(dx)$. The drift parameter c is determined by the martingale condition for $(S_t)_{t \geq 0}$. The time of moment explosion can be calculated [13] and is given by

$$T_*(u) = -\frac{1}{\lambda} \log \max \left(0, 1 - \frac{2\lambda \max(0, \kappa_+ - \rho u)}{u(u-1)} \right) \quad (31)$$

where $\kappa_+ := \sup \{u > 0 : \kappa(u) < \infty\}$ and $\kappa(u) = \int_0^\infty (e^{ux} - 1) \nu(dx) \in [0, \infty]$.

Moment Explosions in Affine Diffusion Models of Dai–Singleton Type

For affine diffusion models with an arbitrary number of stochastic factors, the analysis of moment explosions through the Riccati equations has been studied by Glasserman and Kim [10]. Without structural restrictions, this approach will lead to multiple coupled Riccati differential equations, whose blow-up behavior is tedious to analyze in full generality. However, for concrete specifications, this approach can still lead to explicit results. Glasserman and Kim [10] consider affine models (see **Affine Models**), of Dai–Singleton type [7], which are given by a diffusion process

$$dY_t = -A^\top (\Theta - Y_t) dt + \sqrt{\text{diag}(b + B^\top Y_t)} dW_t \quad (32)$$

evolving on the state space $\mathbb{R}_{\geq 0}^m \times \mathbb{R}^{n-m}$. The state vector Y is partitioned correspondingly, into components (Y^v, Y^d) , called *volatility factors* and *dependent factors*. The vector $b \in \mathbb{R}^n$ and matrices $A, B \in \mathbb{R}^{n \times n}$ are subject to the following structural constraints:

- (C1) $A = \begin{pmatrix} A^v & A^c \\ 0 & A^d \end{pmatrix}$, with real and strictly negative eigenvalues.
- (C2) The off-diagonal entries of A^v are nonnegative.
- (C3) The vector $\Theta = (\Theta^v, \Theta^d)$ has $\Theta^d = 0$, $\Theta^v \geq 0$, and $(-A^\top \Theta)^v \gg 0$.ⁱ
- (C4) $B = \begin{pmatrix} I & B^c \\ 0 & 0 \end{pmatrix}$, and $b = (b^v, b^d)$ with $b^v = 0$ and $b^d = (1, \dots, 1)$.

Note that condition C1 assumes strict mean reversion in all components, which is a typical assumption for interest rate models. Most equity pricing models, however, will not satisfy this condition in the strict sense: The Heston model, for example, is of the form (32), but has an eigenvalue of 0 in the matrix A , and thus does not satisfy C1. Nevertheless, relaxing this condition is in general not a problem, see for example [9]. Glasserman and Kim [10] show that the moments

of Y_t are represented by the *transform formula*

$$\begin{aligned} \mathbb{E}[\exp(2u \cdot Y_t)] &= \exp \left(-2 \int_0^t \Theta^\top A x(s) ds \right. \\ &\quad \left. + 2 \int_0^t |x^d(s)|^2 ds + 2x(t) \cdot Y_0 \right) \end{aligned} \quad (33)$$

where $x(t)$ is a solution to the coupled system of Riccati equations, given by

$$\begin{aligned} \begin{pmatrix} \dot{x}_1(t) \\ \vdots \\ \dot{x}_n(t) \end{pmatrix} &= \begin{pmatrix} A^v & A^c \\ 0 & A^d \end{pmatrix} \cdot \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix} \\ &+ \begin{pmatrix} I & B^c \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1^2(t) \\ \vdots \\ x_n^2(t) \end{pmatrix} \end{aligned} \quad (34)$$

with initial condition $x(0) = u$. Equation (33) holds in the sense that if either side is well defined and finite, the other one is also finite, and equality holds. Thus, moment explosions can again be linked to the blow-up time of the ODE (34). [10] considers two concrete specifications of the above model, with one volatility factor and one dependent factor in each case. Owing to conditions C1–C4, the model parameters are of the form $A = \begin{pmatrix} p & q \\ 0 & r \end{pmatrix}$, $B = \begin{pmatrix} 1 & s \\ 0 & 0 \end{pmatrix}$, and $\Theta = (\theta_1, 0)$, with $p < 0$, $q \geq 0$, $r < 0$, $s \geq 0$, and $\theta_1 \geq 0$.

- $q = s = 0$: This specification decouples the system (34) fully, which can then easily be solved explicitly. In this case, the moment explosion time is given by

$$T_*(u_1, u_2) = \begin{cases} +\infty, & u_1 \leq -p \\ \frac{1}{p} \log \left(1 + \frac{p}{u_1} \right), & u_1 > -p \end{cases} \quad (35)$$

Note that the moment explosion time does not depend on u_2 .

- $s > 0$, $q = 0$, $r = p < 0$: In this case, the system (34) decouples only partially; The equation for the second component becomes $\dot{x}_2 =$

px_2 , with the solution $x_2(t) = u_2 e^{pt}$. Substituting into the equation for the first component yields $\dot{x}_1(t) = px_1 + x_1^2 + su_2^2 e^{2pt}$, a nonautonomous Riccati equation. After the transformation $\xi(t) = e^{-pt}x(t)$ it can be solved explicitly, and the moment explosion time is determined as

$$\begin{aligned} T_*(u_1, u_2) &= \frac{1}{p} \log \max \left\{ 0, \right. \\ &\quad \left. \frac{p}{|u_2|\sqrt{s}} \operatorname{arccot} \left(\frac{u_1}{|u_2|\sqrt{s}} + 1 \right) \right\} \end{aligned} \quad (36)$$

End Notes

- ^aOn the intervals $(-\infty, 0)$ and $(1, \infty)$, respectively.
^bIn fact, it does not even depend on x , which implies the homogeneity properties in equation (5).
^cOnly $u \notin [0, 1]$ needs to be discussed; in this case, $\chi(u) = 0 \implies \Delta(u) < 0$.
^dWhen $u=1$, equation (14) is precisely the Cox–Ingersoll–Ross bond pricing formula.
^eFor $u < u^*$ since equation (14) explodes as $u \uparrow u^*$, where $u^* > 0$ is determined by $I(u^*) \equiv \lambda + \gamma(u) \coth(\gamma(u)t/2) = 0$.
^fCare is necessary since f can be $+\infty$; see [15] for a proper discussion *via* localization.
^gA supersolution $\bar{f}$ of equation (20) satisfies $\mathcal{A}\bar{f} - \frac{\partial \bar{f}}{\partial t} \leq 0$, a subsolution $\underline{f}$ satisfies $\mathcal{A}\underline{f} - \frac{\partial \underline{f}}{\partial t} \geq 0$.
^h X_t^{Heston} denotes the usual log-price process in the classical Heston model, that is, with $J \equiv 0$.
ⁱFollowing the notation of [7], “ $\gg$ ” denotes strict inequality, simultaneously in all components of the vectors.

References

- [1] Alfonsi, A. (2008). *High Order Discretization Schemes for the CIR Process: Application to Affine Term Structure and Heston Models*. Preprint.
- [2] Andersen, L.B.G. & Piterbarg, V.V. (2007). Moment explosions in stochastic volatility models, *Finance and Stochastics* **11**, 29–50.
- [3] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics, *Journal of the Royal Statistical Society B* **63**, 167–241.
- [4] Bates, D.S. (2000). Post-’87 crash fears in the S&P 500 futures option market, *Journal of Econometrics* **94**, 181–238.
- [5] Benaim, S. & Friz, P. (2008). Smile asymptotics ii: models with known moment generating functions, *Journal of Applied Probability* **45**(1), 16–32.

-
- [6] Benaim, S., Friz, P. & Lee, R. (2008). The Black Scholes implied volatility at extreme strikes, in *Frontiers in Quantitative Finance: Volatility and Credit Risk Modeling*, R. Cont, ed, Wiley, Chapter 2.
 - [7] Dai, Q. & Singleton, K.J. (2000). Specification analysis of affine term structure models, *The Journal of Finance* **55**, 1943–1977.
 - [8] Duffie, D., Filipovic, D. & Schachermayer, W. (2003). Affine processes and applications in finance, *The Annals of Applied Probability* **13**(3), 984–1053.
 - [9] Filipović, D. & Mayerhofer, E. (2009). *Affine Diffusion Processes: Theory and Applications*, Preprint, arXiv:0901.4003.
 - [10] Glasserman, P. & Kim, K.-K. (2009). Moment explosions and stationary distributions in affine diffusion models, *Mathematical Finance*, Forthcoming, available at SSRN: <http://ssrn.com/abstract=1280428>.
 - [11] Gulisashvili, A. & Stein, E. (2009). Implied volatility in the Hull-White model, *Mathematical Finance*, to appear.
 - [12] Kallsen, J. & Muhle-Karbe, J. (2008). *Utility Maximization in Affine Stochastic Volatility Models*, Preprint.
 - [13] Keller-Ressel, M. (2008). Moment explosions and long-term behavior of affine stochastic volatility models, arXiv:0802.1823, forthcoming. in *Mathematical Finance*.
 - [14] Lee, R. (2004). The moment formula for implied volatility at extreme strikes, *Mathematical Finance* **14**(3), 469–480.
 - [15] Lions, P.-L. & Musiela, M. (2007). Correlations and bounds for stochastic volatility models, *Annales de l'Institut Henri Poincaré* **24**, 1–16.
 - [16] Sato, K.-I. (1999). *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press.

PETER K. FRIZ & MARTIN KELLER-RESSEL

Implied Volatility in Stochastic Volatility Models

Given the geometric Brownian motion (hence constant volatility) dynamics of an underlying share price, the Black–Scholes formula finds the no-arbitrage prices of call or put options. Given, on the other hand, the price of a call or put, the Black–Scholes implied volatility is by definition the unique volatility parameter such that the Black–Scholes formula recovers the given option price.

If the share price truly follows geometric Brownian motion, then the Black–Scholes implied volatility matches the constant realized volatility of the shares. Empirically, however, stock prices do not exhibit constant volatility, which explains the description in [23] of implied volatility as “the wrong number to put in the wrong formula to obtain the right price”. Nonetheless, the Black–Scholes implied volatility remains, at the very least, a language/scale/metric by which option prices may be quoted and compared across strikes, expiries, underliers, and observation times, as noted in [17].

Moreover, even under stochastic volatility dynamics, Black–Scholes implied volatility is not only a language but indeed carries meaningful information about realized volatility. This article surveys those relationships between implied and realized stochastic volatility, in particular, the following:

- Expected realized variance equals the weighted average of implied variance across strikes, with “implied normal” weights.
- Implied volatility of an option is the *break-even realized volatility* for “business-time delta hedging” of that option.
- Implied volatility at-the-money *approximates expected realized volatility*, under an independence condition.

Aside from Black–Scholes implied volatility, alternative notions of options-implied volatility have robust relationships to realized volatility. We define and discuss two notions of model-free implied volatility (MFIV):

- “VIX-style” MFIV equals the square root of expected variance.
- “Synthetic volatility swap (SVS) style” MFIV equals expected volatility under an independence condition, and approximates expected volatility under perturbations of that condition.

Unless otherwise noted, the only assumptions on the underlying price process are positivity and continuity.

Specifically, on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$, let S be a positive continuous martingale. Regard S as the share price of an underlying tradable asset, and $\mathbb{P}$ as risk-neutral measure, with respect to a bond having price 1 at all times. Extensions to arbitrary deterministic interest rates are straightforward. Let $\mathbb{E}_t$ denote $\mathcal{F}_t$ -conditional expectation, with respect to $\mathbb{P}$. Let

$$X_t := \log(S_t/S_0) \quad (1)$$

denote the log returns process, and let $\langle X \rangle_t$ denote its *quadratic variation* process, which may be regarded as the unannualized running total of the squared realized returns of S continuously monitored on $[0, t]$.

Fixing a time horizon $T > 0$, define *realized variance* to be

$$\langle X \rangle_T$$

and define *realized volatility* to be the square root $\sqrt{\langle X \rangle_T}$ of realized variance. For example, if S has dynamics

$$dS_t = \sigma_t S_t dW_t \quad (2)$$

with respect to Brownian motion W , then $\langle X \rangle_T = \int_0^T \sigma_t^2 dt$.

Black–Scholes Implied Volatility

Fix a time horizon $T > 0$. Define the Black–Scholes [3] function, for S, K , and σ positive, by

$$C^{\text{bs}}(\sigma, S, K) := SN\left(\frac{\log(S/K)}{\sigma} + \frac{\sigma}{2}\right) - KN\left(\frac{\log(S/K)}{\sigma} - \frac{\sigma}{2}\right) \quad (3)$$

where N is the standard normal cdf. Define $C^{\text{bs}}(0, S, K) := (S - K)^+$.

For each $K > 0$, define the time-0 dimensionless Black–Scholes *implied volatility* $IV_0(K)$ to be the unique solution of

$$C^{\text{bs}}(IV_0(K), S_0, K) = \mathbb{E}_0(S_T - K)^+ =: C(K) \quad (4)$$

Dividing $IV_0(K)$ by $\sqrt{T}$ produces the usual annualized implied volatility.

Often, it is more convenient to regard the Black–Scholes formula as a function of dimensionless variance instead of dimensionless volatility, so define

$$C^{\text{BS}}(V, S, K) := C^{\text{bs}}(\sqrt{V}, S, K) \quad (5)$$

Moreover, it may be convenient to regard the Black–Scholes implied volatility as a function of log strike instead of strike, so define

$$IV_0(k) := IV_0(S_0 e^k) \quad (6)$$

We survey how the realized volatility $\sqrt{\langle X \rangle_T}$ (or realized variance $\langle X \rangle_T$) relates to the time-0 implied volatility (or its square, the *implied variance*).

Expected Realized Variance Equals Weighted Average Implied Variance Across Strikes

Black–Scholes implied variance at one strike does not determine the risk-neutral expectation of realized variance, but the weighted average of implied variance at all strikes does so. This result facilitates, for instance, analysis [14] of how the implied volatility skew’s slope and convexity relate to expected variance.

The “implied normal” weights are given by the standard normal distribution, applied to the log strike standardized by “implied standard deviations”. Specifically, assuming $IV_0(k) > 0$, define the standardized log strike by

$$z(k) := -d_2(k) := \frac{k}{IV_0(k)} + \frac{IV_0(k)}{2} \quad (7)$$

The result

$$\mathbb{E}_0 \langle X \rangle_T = \int_{-\infty}^{\infty} IV_0^2(k) dN(z(k)) \quad (8)$$

follows from relating each side to the value of a log contract.

Expected Realized Variance Equals Log Contract Value. Realized variance admits replication

by holding a log contract and dynamically trading shares, *via* the strategy developed in [7, 10, 21], and [9]. Specifically,

$$\begin{aligned} dX_t &= d \log S_t = \frac{1}{S_t} dS_t - \frac{1}{2S_t^2} d\langle S \rangle_t \\ &= \frac{1}{S_t} dS_t - \frac{1}{2} d\langle X \rangle_t \end{aligned} \quad (9)$$

hence,

$$\langle X \rangle_T = -2 \log(S_T/S_0) + \int_0^T \frac{2}{S_t} dS_t \quad (10)$$

Therefore, the log contract payoff $-2 \log(S_T/S_0)$, plus the profit/loss from a dynamic position long $2/S_t$ shares, replicates $\langle X \rangle_T$. A corollary is that

$$\mathbb{E}_0 \langle X \rangle_T = \mathbb{E}_0(-2X_T) \quad (11)$$

if they are finite.

Log Contract Value Equals Weighted Average Implied Variance. In turn, the expectation of $-2X_T$ equals weighted average implied variance. Proofs appear in [19] and [22]. The following is due to [22]. Let $P(K) := \mathbb{E}_0(K - S_T)^+$. Assuming differentiability of IV_0 ,

$$\mathbb{E}_0(-2X_T) = \int_0^{S_0} \frac{2}{K^2} P(K) dK + \int_{S_0}^{\infty} \frac{2}{K^2} C(K) dK \quad (12)$$

$$= \int_0^{S_0} \frac{2}{K} P'(K) dK + \int_{S_0}^{\infty} \frac{2}{K} C'(K) dK \quad (13)$$

$$\begin{aligned} &= 2 \int_{-\infty}^0 \left(N'(d_2) + N(-d_2) IV_0' \right) dk \\ &\quad + 2 \int_0^{\infty} \left(N'(d_2) IV_0' - N(d_2) \right) dk \end{aligned} \quad (14)$$

$$= 2 \int_{-\infty}^{\infty} k N'(d_2) d_2' dk + 2 \int_{-\infty}^{\infty} N'(d_2) IV_0' dk \quad (15)$$

$$= 2 \int_{-\infty}^{\infty} k N'(d_2) d_2' + N'(d_2) d_2 d_2' IV_0 dk \quad (16)$$

$$= \int_{-\infty}^{\infty} IV_0^2(k) N'(z(k)) z'(k) dk \quad (17)$$

where $'$ denotes derivative (unambiguously, as C , P , d_2 , IV_0 , N are defined as single-variable functions).

For brevity, we suppress the argument (k) of d_2 and IV_0 and their derivatives.

To justify the integration by parts in equations (15, 16), it suffices to assume the existence of $\varepsilon > 0$ such that $\mathbb{E}S_T^{1+\varepsilon} < \infty$ and $\mathbb{E}S_T^{-\varepsilon} < \infty$. Then the *moment formula* [18] implies that for some $\beta < 2$ and all $|k|$ sufficiently large, we have $IV_0^2(k) < \beta|k|$; hence

$$\begin{aligned} kN(d_2)|_{-\infty}^0 - kN(-d_2)|_0^\infty &= 0 \\ \text{and } N'(d_2)IV_0|_{-\infty}^\infty &= 0 \end{aligned} \quad (18)$$

Combining equations (11) and (17) gives the conclusion in equation (8).

Implied Volatility Equals Break-even Realized Volatility

Suppose that we buy at time 0 a T -expiry K -strike call or put; to be definite, let us say a call. We pay a premium of $C_0 := C^{\text{BS}}(IV_0^2(K), S_0, K)$.

Dynamically, delta hedging this option using shares, we have, in principle, a position that is delta neutral and “long vega”. Indeed, the implied volatility is the option’s *break-even realized volatility* in the following sense: There exists a model-independent share trading strategy N_t , such that

$$\begin{aligned} P\&L := -C_0 + \int_0^T N_t dS_t + (S_T - K)^+ \\ &< 0 \text{ in the event } \sqrt{\langle X \rangle_T} < IV_0(K) \end{aligned} \quad (19)$$

and $P\&L \geq 0$ in the event $\sqrt{\langle X \rangle_T} \geq IV_0(K)$.

In other words, total profit/loss (from the time-0 option purchase, the trading in shares, and the time- T option payout) is negative if and only if volatility realizes to less than the initial implied volatility.

Implied Volatility is Break-even Realized Volatility for Business-time Delta Hedging. Define the *business-time* delta hedging strategy by letting

$$\tau := \inf\{t : \langle X \rangle_t = IV_0^2(K)\} \quad (20)$$

and holding N_t shares at each time $t \in [0, T]$, where

$$\begin{aligned} N_t &:= -\frac{\partial C^{\text{BS}}}{\partial S}(IV_0^2(K) - \langle X \rangle_t, S_t, K) \\ &t \in [0, \tau \wedge T] \end{aligned} \quad (21)$$

$$N_t := -\mathbb{I}(S_t > K), \quad t \in (\tau \wedge T, T] \quad (22)$$

The break-even property follows from applying Ito’s rule to the process

$$C_t := C^{\text{BS}}(IV_0^2(K) - \langle X \rangle_t, S_t, K), \quad t \in [0, \tau \wedge T] \quad (23)$$

to obtain

$$\begin{aligned} dC_t &= -\frac{\partial C^{\text{BS}}}{\partial V} d\langle X \rangle_t + \frac{\partial C^{\text{BS}}}{\partial S} dS_t \\ &\quad + \frac{1}{2} \frac{\partial^2 C^{\text{BS}}}{\partial S^2} d\langle S \rangle_t \\ &= \left(\frac{1}{2} S_t^2 \frac{\partial^2 C^{\text{BS}}}{\partial S^2} - \frac{\partial C^{\text{BS}}}{\partial V} \right) d\langle X \rangle_t \\ &\quad + \frac{\partial C^{\text{BS}}}{\partial S} dS_t = \frac{\partial C^{\text{BS}}}{\partial S} dS_t \end{aligned} \quad (24)$$

where the partials of C^{BS} are evaluated at $(IV_0^2(K) - \langle X \rangle_t, S_t, K)$. Therefore,

$$-C_{\tau \wedge T} = -C_0 - \int_0^{\tau \wedge T} \frac{\partial C^{\text{BS}}}{\partial S} dS_t \quad (25)$$

as shown in [2, 11, 20].

In the event $\langle X \rangle_T < IV_0^2(K)$, hence $T < \tau$, we have

$$\begin{aligned} P\&L &= (S_T - K)^+ - C_T \\ &= (S_T - K)^+ - C^{\text{BS}}(IV_0^2(K) - \langle X \rangle_T, S_T, K) \\ &< 0 \end{aligned} \quad (26)$$

and in the event $\langle X \rangle_T \geq IV_0^2(K)$, hence $\tau \leq T$, we have

$$\begin{aligned} P\&L &= (S_T - K)^+ - C_\tau - \int_\tau^T \mathbb{I}(S_t > K) dS_t \\ &= (S_T - K)^+ - (S_\tau - K)^+ \\ &\quad - \mathbb{I}(S_\tau > K)(S_T - S_\tau) \geq 0 \end{aligned} \quad (27)$$

as claimed. This break-even result is a special case of a proposition in [6].

Implied Volatility is Not Break-even Realized Volatility for Standard Delta Hedging. The break-even property of the previous section does not extend

to standard “calendar time” delta hedging, defined by share holdings

$$-\frac{\partial C^{\text{BS}}}{\partial S}((T-t)\bar{IV}_0^2, S_t, K), \quad t \in [0, T] \quad (28)$$

where $\bar{IV}_0^2 := IV_0^2(K)/T$ denotes the time-0 annualized implied variance.

This strategy guarantees neither a profit in the event that $\langle X \rangle_T > IV_0^2(K)$ nor a loss in the opposite event. To see this, under the dynamics (2), let

$$Y_t = C^{\text{BS}}((T-t)\bar{IV}_0^2, S_t, K) \quad (29)$$

and apply Ito’s rule to obtain

$$\begin{aligned} dY_t &= -\frac{\partial C^{\text{BS}}}{\partial V}\bar{IV}_0^2 dt + \frac{\partial C^{\text{BS}}}{\partial S} dS_t \\ &\quad + \frac{1}{2} \frac{\partial^2 C^{\text{BS}}}{\partial S^2} d\langle S \rangle_t \\ &= -\frac{1}{2} \bar{IV}_0^2 S_t^2 \frac{\partial^2 C^{\text{BS}}}{\partial S^2} dt + \frac{\partial C^{\text{BS}}}{\partial S} dS_t \\ &\quad + \frac{1}{2} \sigma_t^2 S_t^2 \frac{\partial^2 C^{\text{BS}}}{\partial S^2} dt \end{aligned} \quad (30)$$

where the partial derivatives of C^{BS} are evaluated at $((T-t)\bar{IV}_0^2, S_t, K)$. Hence,

$$\begin{aligned} P\&L = Y_T - Y_0 - \int_0^T \frac{\partial C^{\text{BS}}}{\partial S} dS_t \\ &= \int_0^T \frac{1}{2} (\sigma_t^2 - \bar{IV}_0^2) S_t^2 \frac{\partial^2 C^{\text{BS}}}{\partial S^2} dt \end{aligned} \quad (31)$$

which is half the time-integrated *cash-gamma-weighted* difference of instantaneous variance σ_t^2 and implied variance $\bar{IV}_0^2$, as shown in [7] and [12]. So if, along some trajectory, $\sigma_t > \bar{IV}_0$ at points where gamma is low, but $\sigma_t < \bar{IV}_0$ at points where gamma is high, then it can occur that realized variance $\int_0^T \sigma_t^2 dt$ exceeds implied variance IV_0^2 , yet this long-vega strategy incurs a *loss*.

In conclusion, implied volatility is the option’s break-even realized volatility for business-time delta hedging, but not for calendar-time delta hedging.

Implied Volatility ATM Approximates Expected Realized Volatility, Under an Independence Condition

In this section, we specialize to dynamics (2) such that σ and W are *independent*.

Let $K_{\text{atm}} := S_0$ be the at-the-money (ATM) strike. Then

$$\mathbb{E}(S_T - K_{\text{atm}})^+ = \mathbb{E}C^{\text{bs}}(\sqrt{\langle X \rangle_T}, S_0, K_{\text{atm}}) \quad (32)$$

$$\leq C^{\text{bs}}(\mathbb{E}\sqrt{\langle X \rangle_T}, S_0, K_{\text{atm}}) \quad (33)$$

by the conditioning argument of [15], independence, and the concavity of

$$v \mapsto C^{\text{bs}}(v, S_0, K_{\text{atm}}) \quad (34)$$

It follows that

$$IV_0(K_{\text{atm}}) \leq \mathbb{E}\sqrt{\langle X \rangle_T} \quad (35)$$

The function (34), while concave, is nearly linear for small v ; indeed, its second derivative vanishes at $v = 0$, as observed in [4]. Therefore, the inequalities (33) and (35) are nearly equalities, as shown in [13]. In that sense,

$$IV_0(K_{\text{atm}}) \approx \mathbb{E}\sqrt{\langle X \rangle_T} \quad (36)$$

assuming the independence of σ and W .

Model-free Implied Volatility (MFIV)

Inverting Black–Scholes is not the only way to extract an implied volatility from option prices. While the ATM Black–Scholes implied volatility approximates expected volatility under the independence assumption, alternative definitions of MFIV use call/put data at all strikes, in order to reflect the expected variance or volatility under more general conditions.

VIX-style MFIV Equals the Square Root of Expected Realized Variance

Motivated by equation (11), define the VIX-style model-free implied volatility by

$$\begin{aligned} VIXIV_0 &:= \sqrt{\mathbb{E}_0[-2X_T]} \\ &:= \sqrt{\mathbb{E}_0[-2\log(S_T/S_0) + 2(S_T/S_0) - 2]} \end{aligned} \quad (37)$$

$VIXIV_0$ is an observable function of option prices, specifically the square root of the time-0 value of the portfolio

$$\begin{aligned} 2/K^2 dK & \text{ calls at strikes } K > S_0 \\ 2/K^2 dK & \text{ puts at strikes } K < S_0 \end{aligned} \quad (38)$$

Indeed in 2003, the Chicago board options exchange (CBOE) [8] adopted an implementation of equation (38) to define the VIX volatility index (but due to the availability of only finitely many strikes in practice, the CBOE VIX is not precisely identical to $VIXIV_0$; see [16]).

By equation (11), the square of VIX-style MFIV equals expected realized variance:

$$VIXIV_0 = \sqrt{\mathbb{E}_0 \langle X \rangle_T} \quad (39)$$

However, by Jensen's inequality, for random $\langle X \rangle_T$,

$$VIXIV_0 > \mathbb{E}_0 \sqrt{\langle X \rangle_T} \quad (40)$$

thus, VIX-style MFIV differs from expected realized volatility, due to convexity.

SVS-style MFIV Equals Expected Realized Volatility

For nonconstant $\langle X \rangle_T$, the VIX-style MFIV never equals $\mathbb{E}_0 \sqrt{\langle X \rangle_T}$; in contrast, the SVS-style MFIV will equal $\mathbb{E}_0 \sqrt{\langle X \rangle_T}$ exactly under an independence condition, and approximately under perturbations of that condition. Define SVS-style model-free implied volatility (where SVS stands for “synthetic volatility swap”) by

$$\begin{aligned} SVSIV_0 &:= \mathbb{E}_0 \sqrt{\frac{\pi}{2}} e^{X_T/2} \\ &\times \left| X_T I_0(X_T/2) - X_T I_1(X_T/2) \right| \end{aligned} \quad (41)$$

where $X_T := \log(S_T/S_0)$ and I_ν is the modified Bessel function of order ν .

$SVSIV_0$ is observable from option prices, as the time-0 value of the portfolio

$$\begin{aligned} & \sqrt{\pi/2}/S_0 \\ & \text{straddles at strike } K = S_0, \\ & \sqrt{\frac{\pi}{8K^3 S_0}} \left[I_1(\log \sqrt{K/S_0}) - I_0(\log \sqrt{K/S_0}) \right] dK \\ & \text{calls at strikes } K > S_0, \\ & \sqrt{\frac{\pi}{8K^3 S_0}} \left[I_0(\log \sqrt{K/S_0}) - I_1(\log \sqrt{K/S_0}) \right] dK \\ & \text{puts at strikes } K < S_0 \end{aligned} \quad (42)$$

Under the dynamics (2) with σ and W independent, the exact equality

$$SVSIV_0 = \mathbb{E}_0 \sqrt{\langle X \rangle_T} \quad (43)$$

is proved in [5]. Moreover, it still holds approximately, under perturbations of the independence assumption. To be precise, consider a family of processes $S^{[\rho]}$, indexed by parameters $\rho \in [-1, 1]$, and defined by

$$\begin{aligned} dS_t^{[\rho]} &= \sqrt{1 - \rho^2} \sigma_t S_t^{[\rho]} dW_{1t} + \rho \sigma_t S_t^{[\rho]} dW_{2t} \\ S_0^{[\rho]} &= S_0 \end{aligned} \quad (44)$$

where W_1 and W_2 are $\mathcal{F}_t$ -Brownian motions, and σ and W_2 are adapted to some filtration $\mathcal{H}_t \subseteq \mathcal{F}_t$, where $\mathcal{H}_T$ and $\mathcal{F}_T^{W_1}$ are independent. This includes all the standard stochastic volatility models of the form $d\sigma_t = \alpha(\sigma_t) dt + \beta(\sigma_t) dW_{2t}$.

Changing the ρ parameter does not affect the σ dynamics, and hence cannot affect $\mathbb{E}_0 \sqrt{\langle X \rangle_T}$. However, changing ρ does change the S dynamics, and hence may change option prices, $IV_0(K_{\text{atm}})$, and $SVSIV_0$. Thus, the relationships (36, 43) below

$$IV_0(K_{\text{atm}}) \approx \mathbb{E}_0 \sqrt{\langle X \rangle_T}, \quad SVSIV_0 = \mathbb{E}_0 \sqrt{\langle X \rangle_T} \quad (45)$$

that are valid for the uncorrelated case $S = S^{[0]}$, may not hold for $S = S^{[\rho]}$ where $\rho \neq 0$. Unlike $IV_0(K_{\text{atm}})$, the $SVSIV_0$ has the robustness property of being immunized against perturbations of ρ around $\rho = 0$, meaning that

$$\left. \frac{\partial}{\partial \rho} \right|_{\rho=0} SVSIV_0 = 0 \quad (46)$$

can be verified. This suggests that SVS-style implied volatility $SVSIV_0$ should outperform Black–Scholes implied volatility IV_0 , as an approximation to the expected realized volatility, at least for ρ not too large.

This is confirmed in [5] for Heston dynamics with parameters from [1], and $T = 0.5$. Across essentially all correlation assumptions, the SVS notion of implied volatility exhibited the smallest bias, relative to the true expected annualized volatility. For example, in the case $\rho = -0.64$, the VIX-style implied volatility had bias +98 bp, the Black–Scholes implied volatility had bias −30 bp, and the SVS-style implied volatility had the smallest bias, −6 bp.

Acknowledgments

This article benefited from the comments of Peter Carr.

References

- [1] Bakshi, G., Cao, C. & Chen, Z. (1997). Empirical performance of alternative option pricing models, *Journal of Finance* **52**, 2003–2049.
- [2] Bick, A. (1995). Quadratic-variation-based dynamic strategies, *Management Science* **41**, 722–732.
- [3] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- [4] Brenner, M. & Subrahmanyam, M. (1988). A simple formula to compute the implied standard deviation, *Financial Analysts Journal* **44**, 80–83.
- [5] Carr, P. & Lee, R. (2008). *Robust Replication of Volatility Derivatives*, Bloomberg LP, University of Chicago.
- [6] Carr, P. & Lee, R. (2008). *Hedging Variance Options on Continuous Semimartingales*, Forthcoming in *Finance and Stochastics*.
- [7] Carr, P. & Madan, D. (1998). Towards a theory of volatility trading, in *Volatility*, R. Jarrow, ed, Risk Publications, pp. 417–427.
- [8] CBOE. (2003). The VIX White Paper, Chicago Board Options Exchange.
- [9] Derman, E., Demeterfi, K., Kamal, M. & Zou, J. (1999). A guide to volatility and variance swaps, *Journal of Derivatives* **6**, 9–32.
- [10] Dupire, B. (1992). *Arbitrage pricing with stochastic volatility*, Société Générale.
- [11] Dupire, B. (2005). *Volatility Derivatives Modeling*, Bloomberg LP.
- [12] El Karoui, N., Jeanblanc-Picqué, M. & Shreve, S. (1998). Robustness of the Black and Scholes formula, *Mathematical Finance* **8**, 93–126.
- [13] Feinstein, S.P. (1989). *The Black–Scholes Formula is Nearly Linear in Sigma for At-the-Money Options: Therefore Implied Volatilities from At-the-Money Options are Virtually Unbiased*, Federal Reserve Bank of Atlanta.
- [14] Gatheral, J. (2006). *The Volatility Surface: A Practitioner's Guide*, John Wiley & Sons.
- [15] Hull, J. & White, A. (1987). The pricing of options on assets with stochastic volatilities, *Journal of Finance* **42**, 281–300.
- [16] Jiang, G.J. & Tian, Y.S. (2005). The model-free implied volatility and its information content, *Review of Financial Studies* **18**, 1305–1342.
- [17] Lee, R. (2004). Implied volatility: statics, dynamics, and probabilistic interpretation, *Recent Advances in Applied Probability*, Springer, pp. 241–268.
- [18] Lee, R. (2004). The moment formula for implied volatility at extreme strikes, *Mathematical Finance* **14**, 469–480.
- [19] Matytsin, A. (2000). *Perturbative Analysis of Volatility Smiles*, Merrill Lynch.
- [20] Mykland, P. (2000). Conservative delta hedging, *Annals of Applied Probability* **10**, 664–683.
- [21] Neuberger, A. (1994). The log contract, *Journal of Portfolio Management* **20**, 74–80.
- [22] Polishchuk, A. (2007). *Variance swap valuation*, Bloomberg LP.
- [23] Rebonato, R. (1999). *Volatility and Correlation in the Pricing of Equity, FX and Interest Rate Options*, John Wiley & Sons.

PETER CARR & ROGER LEE

Local Volatility Model

The most important input for pricing equity derivatives comes from vanilla call and put options on an equity index or a single stock. The market convention for these options follows the classic Black–Scholes–Merton (BSM) model [3, 15]: the price of each option can be represented by a single number called the *implied volatility*, which is the unknown volatility parameter required in the BSM model to reproduce the price. The implied volatilities for different maturities and strikes are often significantly different, and they collectively form the *implied volatility surface* of the underlying. A fundamental modeling problem is to explain the implied volatility surface accurately using logical assumptions. Many interesting applications follow directly from the solution to this problem. The impact from different parts of the implied volatility surface on a product can be assessed, leading to a deeper understanding of the product, its risks, and the associated hedging strategy. Moreover, different derivatives products, including those not available from the market, may be priced and analyzed under assumptions consistent with the vanilla options.

This article discusses the local volatility surface approach for analyzing equity implied volatility surfaces and examines a common framework in which different modeling assumptions can be incorporated. Local volatility models were first developed by Dupire [11], Derman and Kani [9], and Rubinstein [19] in the last decade and have since become one of the most popular approaches in equity derivatives quantitative research [1, 2, 7, 8, 10, 13, 16, 18]. We present the model from a practitioner’s perspective, discussing calibration techniques with extension to dividends and interest rate modeling, with emphasis on the ease of application to real-world problems.

Basic Model

The basic local volatility model is an extension of BSM to the case where the diffusion volatility becomes a deterministic function of time and the spot price. In the absence of dividends, the stock dynamics can be represented by the following stochastic differential equation:

$$\frac{dS_t}{S_t} = g_t dt + \sigma(t, S_t) dW_t \quad (1)$$

where S_t is the stock price at time t , $g_t = r_t - b_t$ is the known growth rate of the stock (r_t is the interest rate and b_t is the effective stock borrowing cost) at t , $\sigma(t, S)$ is the local volatility function for given time t and stock price S , and W_t is a Brownian motion representing the uncertainty in the stock price. Dynamics (1) can also be viewed as the effective representation of a general stochastic volatility model where $\sigma^2(t, S)$ is the expectation of the instantaneous diffusion coefficient conditioning on $S_t = S$ [13, 17]. If we use $C(t, S; T, K) = \mathbf{E}[\max(S_T - K, 0) | S_t = S]$ to represent the undiscounted price for a European call option with maturity T and strike K when the stock price at time $t \leq T$ is S , then equation (1) leads to the well-known Dupire equation for C :

$$\frac{\partial C}{\partial T} = \frac{\sigma^2(T, K)K^2}{2} \frac{\partial^2 C}{\partial K^2} + g_T \left(C - K \frac{\partial C}{\partial K} \right) \quad (2)$$

Equation (2) gives the relationship between the call option price C and the local volatility function $\sigma(t, S)$. In theory, if arbitrary-free prices of $C(T, K)$ were known for arbitrary T and K , $\sigma(t, S)$ could be recovered by inverting equation (2) with differentials of C . In practice, the market option prices are only directly available on a few maturities and strikes. Schemes for interpolating and extrapolating implied volatilities are often adopted in practice to arrive at a smooth function $C(T, K)$. Such schemes, however, typically lack explicit controls on the various derivatives terms in equation (2), and the local volatility directly inverted from equation (2) can exhibit strange shapes and sometimes attain nonphysical values for reasonable implied volatility input.

Instead of assuming the implied volatilities perfectly known for all maturities and strikes and inverting equation (2), one can model the local volatility function $\sigma(t, S)$ directly as a parametric function. Solving the forward partial differential equation (2) numerically with the initial conditions

$$C(t, S, T = t, K) = \max(S - K, 0) \quad (3)$$

yields call option prices for all maturity T s and strike K s, from which the implied volatility surface can be derived. The parameters of the local volatility function can then be determined by matching the implied volatility surface generated from the model

to that from the market. With a careful design of the local volatility function, this so-called *calibration* process can be implemented very efficiently for practical use. This methodology has the advantage that the knowledge of a perfect implied volatility surface is not required and the model is arbitrage free by construction. In addition, a great amount of analytical flexibility is available, which allows tailor-made designs of different models for specific purposes.

Volatility Surface Design and Calibration

The key to the success of a volatility model lies in an understanding of how the implied volatility surface is used in practice. Empirically, option traders often refer to the implied volatility surface and its shape deformation with intuitive descriptions such as *level*, *slope*, and *curvature*, effectively approximating the shape as simple quadratic functions. In addition, for strikes away from the at-the-money (ATM) region, sometimes the ability to modify the out-of-the-money (OTM) surface independent from the central shape is desired, which traders intuitively speak of as changing the *put wing* or the *call wing*. Thus there exist several degrees of freedom on the volatility surface that a good model should be able to accommodate, and we can design the local volatility function so that each mode is captured by a distinct parameter.

To facilitate comparison across different modeling techniques, we standardize the model specification in terms of the BSM implied volatilities on a small number of strikes per maturity, typically three or five. For example, volatilities on three strikes in the ATM region can be used to provide a precise definition of the traders' level, slope, and curvature parameters. Similarly, fixing volatilities at one downside strike and one upside strike in the OTM region allows the model to agree on a five-parameter specification of level, slope, curvature, put wing, and call wing. These calibration strikes on each maturity are chosen to cover the range of practical interest, usually one to two standard deviations of diffusion at the stock's typical volatility. In the absence of fine structures such as sharp jumps in the underlying, we expect that one standard deviation in the strike range provides a natural length scale over which the stock price distribution varies smoothly. Thus the

implied volatility should have a very smooth shape over the range thus defined, and matching the implied volatilities at the calibration strikes should produce a very good match for all the implied volatilities between them.

The preceding discussions give a straightforward strategy for building the local volatility model—we specify a small number of strikes, and tune the local volatility function with the same number of parameters as the number of strikes for each maturity in a bootstrapping process. The local volatility parameters are then solved through a root-finding routine so that the implied volatilities at the specified strikes on each maturity are reproduced. As each local volatility parameter is designed to capture a distinct aspect of the surface shape, the root-finding system is well behaved and converges quickly to the solution in practice. More importantly, such a process allows a much smaller numerical noise compared to a typical optimization process, giving rise to much more stable calibration results. This is essential in ensuring robust Greeks and scenario outputs from the model.

Discrete Dividend Models

Dividend modeling is an important problem in equity derivatives. It can be shown [15] that with nonzero dividends, the original BSM model only works when the payment amount is proportional to the stock price immediately before the ex-dividend date (ex-date), through incorporating the dividend yields in g_t of equation (1). However, many market participants tend to view future dividends as absolute cash amounts, and this is especially true after trading in index dividend swaps becomes liquid. Existing literature [4, 5, 12, 14] suggests even in the case of a constant volatility, cash dividend equity models (also known as *discrete dividend models*) are much less tractable than proportional dividend ones. Recently, Overhaus *et al.* [16] proposed a theory to ship future cash dividends from the stock price to arrive at a *pure* stock process, on which one can apply the Dupire equation. This theory calls for the changes in future dividends to have a global impact, especially for maturities *before* their ex-dates, a feature that certain traders find somewhat counterintuitive.

Nontrivial dividend specifications can be naturally introduced in the framework here. We note that between ex-dates, equations (1) and (2) continue to

hold without modification. Across an ex-date T_i , a simple model for the stock price is

$$S_{T_i^+} = (1 - Y_i) S_{T_i^-} - D_i \quad (4)$$

where Y_i and D_i are respectively the dividend yield and cash dividend amount for the ex-date T_i . This is the mixed-dividend model, which includes proportional and cash dividend models as special cases. Y_i and D_i can be determined from a *nominal* dividend schedule specifying the ex-dates and the dividend payment amount, as well as a mixing schedule specifying the portion of dividends that should remain as cash, the rest being converted into proportional yield. Typically, cash dividends can be specified for the first few years to reflect the certainty on expected dividends, gradually switching to all proportional in the long term. Theoretically, equation (4) has the disadvantage of allowing negative ex-dividend stock prices. In practice, if the mixing ratio is switched to all proportional after a few years, this does not pose a serious problem. According to dividend model (4), the forward equation across the ex-date becomes

$$C(T_i^+, K) = (1 - Y_i) C\left(T_i^-, \frac{K + D_i}{1 - Y_i}\right) \quad (5)$$

and can be implemented in the same way as standard jump conditions. With equation (5) incorporated, the calibration strategy in the section Volatility Surface Design and Calibration can be applied in exactly the same way. We note that it is straightforward to extend the local volatility model here to handle more interesting dividend models, in which the dividend amount can be made a function of the spot immediately before the ex-date. As long as such a function becomes small enough when the stock price goes to zero, the issue of negative ex-dividend stock prices can be theoretically eliminated.

Stochastic Interest Rate Models

Local volatility models can be extended to cases where stochastic interest rate needs to be considered [16]. The interest rate can be modeled through classic short-rate models, and the equity process is then specified as a diffusion with stochastic growth rate. Following Brigo and Mercurio [6], we have

$$\frac{dS_t}{S_t} = (r_t - b_t) dt + \sigma(t, S_t) dW_t \quad (6a)$$

$$r_t = u_t + y_t \quad (6b)$$

$$dy_t = \alpha(\phi - y_t) dt + \sigma_t y_t^\beta dB_t \quad (6c)$$

$$dW_t dB_t = \rho dt \quad (6d)$$

where u_t is a function of time describing the deterministic part of the interest rate, y_t is a diffusion process modeling the stochastic part of the interest rate, and B_t is a Brownian motion describing the interest rate uncertainty, correlated with W_t with coefficient ρ . In equation (6c), α , β , ϕ , and σ_t are parameters describing the short-rate process. For example, when $\phi = 0$ and $\beta = 0$, equation (6c) is equivalent to the Hull–White model. With nonzero ϕ and $\beta = \frac{1}{2}$, the shifted Cox–Ingersoll–Ross (CIR++) model is obtained. Both models admit closed form pricing formula for zero-coupon bonds, interest rate caps, and swaptions, which can be used for calibration to interest rate derivatives market observables. For a given short-rate model and its parameters, the local volatility function $\sigma(t, S)$ needs to be recovered from equity derivatives market information. This can be achieved by considering the transition density for the joint evolution of the stock price S_t and short rate r_t under stochastic discount factor, that is,

$$\begin{aligned} p(t, S, y; T, K, Y) \\ = \mathbf{E} \left[\exp \left(- \int_t^T r_\tau d\tau \right) \right. \\ \left. \times \delta(S_T - K) \delta(y_T - Y) \middle| S_t = S, y_t = y \right] \quad (7) \end{aligned}$$

The Fokker–Planck equation for such a quantity can be written down as

$$\begin{aligned} \frac{\partial p}{\partial T} = & \frac{\partial^2 (K^2 \sigma^2(T, K) p)}{2 \partial K^2} + \frac{\sigma_T^2}{2} \frac{\partial^2 (Y^{2\beta} p)}{2 \partial Y^2} \\ & + \rho \sigma_T \frac{\partial^2 (K \sigma(T, K) Y^\beta p)}{\partial K \partial Y} \\ & - (u_T + Y - b_T) \frac{\partial (K p)}{\partial K} \\ & - \alpha \phi \frac{\partial p}{\partial Y} + \alpha \frac{\partial (Y p)}{\partial Y} - (u_T + Y) p \quad (8) \end{aligned}$$

By solving equation (8) subject to vanishing boundary conditions on K and Y as well as delta-function initial condition at $T = t$, one can recover the European option prices as

$$C(t, S_t; T, K) = \int_0^\infty dS \max(S - K, 0) \times \int_{-\infty}^\infty p(t, S_t, y_0; T, S, Y) dY \quad (9)$$

and hence derive the implied volatility surface from the hybrid model (6). The strategy discussed in the

section Volatility Surface Design and Calibration can once again be invoked. In practice, since the two-factor model takes significantly more time in calculation than the basic model, it is very effective to use the basic model solution as a starting point for the hybrid calibration.

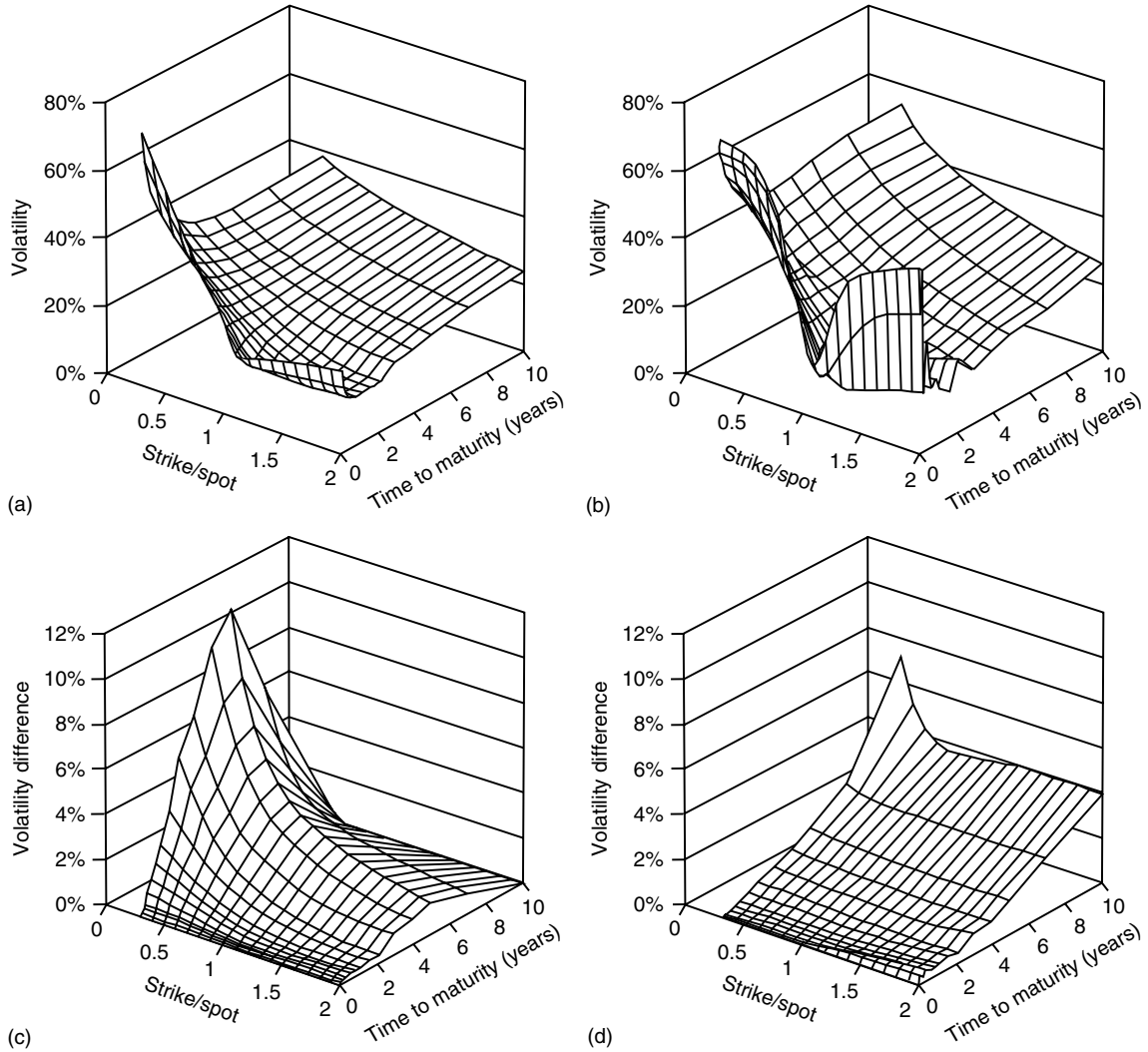


Figure 1 The implied and local volatility surface on the S&P 500 Index in November 2007. (a) The implied volatility as a function of time to maturity and strike price (expressed as a percentage of spot price). (b) The local volatility surface calibrated under the basic model. (c) Changes in the local volatility surface when cash dividends are assumed for the first five years, gradually transitioning to proportional dividends in 10 years. (d) Changes in the local volatility surface where the interest rate is assumed to follow the Hull-White model calibrated to ATM caps with correlation $\rho = 30\%$. In both (c) and (d) the new local volatility is smaller than in (b)

Examples

We use data from the S&P 500 index market as examples to illustrate the preceding discussions. Figure 1(a) and (b) shows a typical implied volatility surface and the calibrated local volatility surface under the basic model, that is, with proportional dividend and deterministic interest rate assumptions. The implied volatility surface is given by option traders. Normally it is retrieved from data in both the listed and OTC options market, interpolated, and extrapolated with trader-specified functions. The local volatility surface is built by simply calibrating to five strikes on each marked maturity, with the Libor-Swap curve and the full index dividend schedule. Excellent

calibration quality can be obtained: the option price differences computed using the input implied volatility surface and the calibrated local volatility surface are less than one basis point of the spot price for most liquid strikes and below 10 basis points across all strikes and maturities. This accuracy is sufficient for most practical purposes.

Figure 1(c) and (d) displays the changes in the local volatility surface when we include effects of cash dividends or stochastic interest rate into the model. We have assumed the Hull–White model for the interest rate in these calculations. The dividend and interest rate specifications are seen to have a significant impact on the local volatility surface and hence can be important in derivatives pricing. Cash

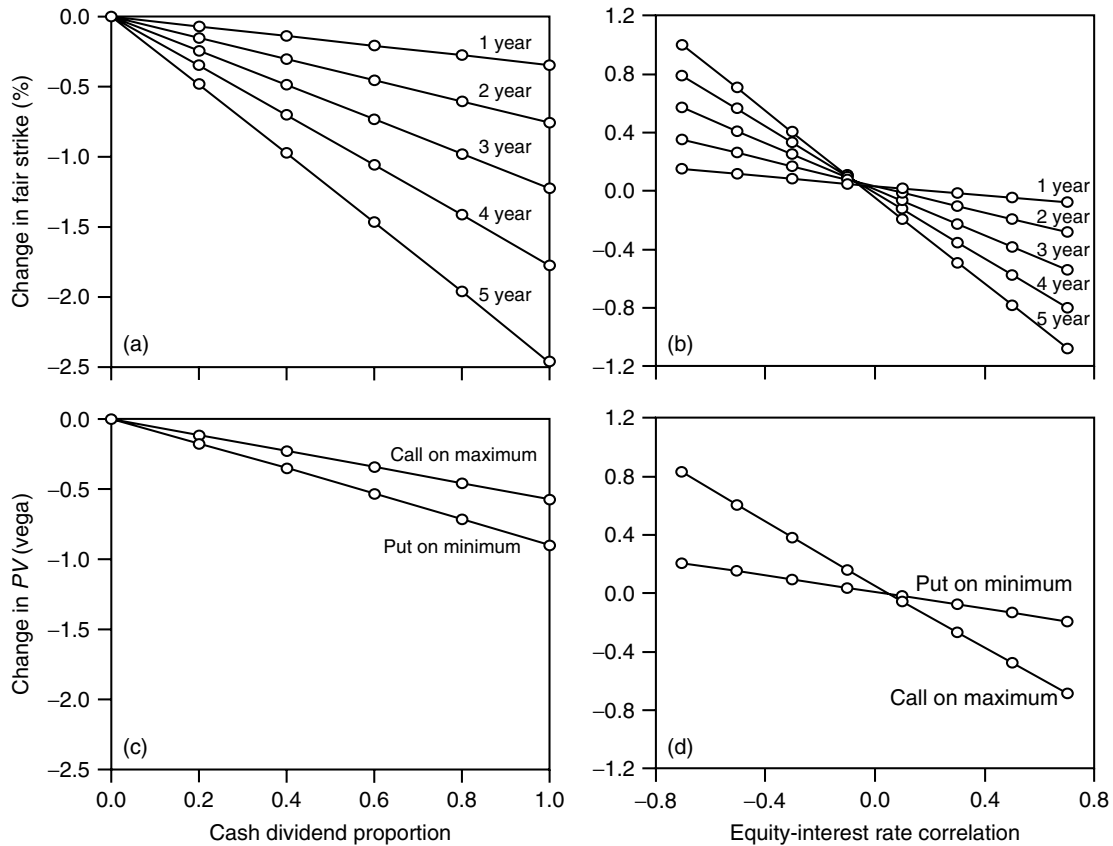


Figure 2 Impact of discrete dividends and stochastic interest rate on derivative pricing. (a) Changes to the fair strike of the variance swaps with different dividend assumptions. (b) Changes to the fair strike of the variance swaps under stochastic interest rate with different correlation. The labels indicate the maturity of the variance swaps. (c) Changes to the *PV* of the lookback options with different dividend assumption. (d) Changes to the *PV* of the lookback options under stochastic interest rate with different correlation. The numbers in (c) and (d) are in units of vega of Table 2

dividends introduce additional deterministic, nonproportional jump structures in the equity dynamics, and to maintain the same implied volatility surface the local volatility needs to become smaller. This effect depends on the dividend size relative to future spot prices, and thus become more pronounced for smaller strikes and longer maturities, producing a skewed shape in the difference. On the other hand, stochastic interest rate introduces volatility in discount bond prices and with positive correlation also reduces the equity local volatility. This effect does not depend on spot levels explicitly and is instead related to the volatility ratio between the interest rate and the equity and their correlation. Since the interest rate usually has a small volatility compared to the equity, to the leading order the effect of stochastic rates can sometimes be approximated by a parallel shift on the local volatility surface.

We can apply these local volatility models to price exotic derivatives not directly available from the vanilla market. One example is variance swaps, which are popular OTC products offered to capitalize on the discrepancy between implied and realized volatility. Another example is lookback options, which provide payoffs on the maximum/minimum index prices over a set of observation dates and can be appealing hedges to insurance companies who have sold policies with similar exposure. Tables 1 and 2 display the pricing results for these structures using the basic model.

Figure 2 shows the pricing impact on these structures when the effects of cash dividends and stochastic interest rates are considered. As the payout for the variance swap is directly linked to the equity's average local volatility, the pricing is strongly affected by the assumption of cash dividends and stochastic

Table 1 Pricing of variance swaps with the basic model

Maturity (years)	Fair strike (%)
1	27.59
2	28.18
3	28.42
4	29.14
5	30.00

The payoff for strike K at maturity is $\frac{252}{N} \sum_{i=0}^{N-1} (\ln \frac{S_{i+1}}{S_i})^2 - K^2$, where S_i is the index closing price on the i th business day from the current date ($i = N$ corresponds to the maturity). The fair strike is the value K such that the contract costs nothing to enter

Table 2 Pricing of five-year lookback options with the basic model

Option type	Payout formula	PV (%)	Vega (%)
Call on maximum	$\max_{i=0}^5 \left(\frac{S_i}{S_0} \right) - \frac{S_5}{S_0}$	25.29	1.17
Put on minimum	$1 - \min_{i=0}^5 \left(\frac{S_i}{S_0} \right)$	22.35	0.85

S_i ($i = 0, 1, \dots, 5$) is the index price at annual observation dates on year i from the current date. The PV is the calculated present value according to the payout formula at maturity. The Vega is the change in PV when a parallel shift of 1% is applied to the implied volatility surface

interest rate. For lookback options, one needs to look at the joint distribution among equity prices across different observation dates. Cash dividends generally reduce the local volatility and hence decrease the correlation between the equity prices at different dates, leading to lower lookback prices. With stochastic interest rate, the effect of modified equity diffusion volatility can either reinforce (e.g., call on maximum) or partly cancel (e.g., put on minimum) the effect of stochastic discounting.

The numerical impact of different modeling assumptions can be comparable to a full percentage difference in volatility. Hence, it may be important to take these into account when accurate and competitive pricing of exotic equity derivatives is required. An extensive and detailed discussion of the impact of stochastic interest rate on popular hybrid products can be found in [16].

References

- [1] Andersen, L. & Brotherton-Ratcliffe, R. (1997). The equity option volatility smile: an implicit finite-difference approach, *Journal of Computational Finance* **1**, 5–38.
- [2] Berestycki, H., Busca, J. & Florent, I. (2002). Asymptotics and calibrations of local volatility models, *Quantitative Finance* **2**, 61–69.
- [3] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 631–659.
- [4] Bos, M. & Vandermark, S. (2002). Finessing fixed dividends, *Risk Magazine* **15**(9), 157–158.
- [5] Bos, R., Gairat, A. & Shepeleva, S. (2003). Dealing with discrete dividends, *Risk Magazine* **16**(1), 109–112.
- [6] Brigo, D. & Mercurio, F. (2006). *Interest Rate Models—Theory and Practice with Smile, Inflation and Credit*, 2nd Edition, Springer Finance.

-
- [7] Brown, G. & Randall, C. (1999). If the skew fits, *Risk Magazine* **12**(4), 62–65.
 - [8] Coleman, T.F., Li, Y. & Verma, A. (1999). Reconstructing the unknown volatility function, *Journal of Computational Finance* **2**, 77–102.
 - [9] Derman, E. & Kani, I. (1994). Riding on a smile, *Risk Magazine* **7**(2), 32–39.
 - [10] Dumas, B., Fleming, J. & Whaley, R.E. (1998). Implied volatility functions: empirical tests, *Journal of Finance* **53**, 2059–2106.
 - [11] Dupire, B. (1994). Pricing with a smile, *Risk Magazine* **7**(1), 18–20.
 - [12] Frishling, F. (2002). A discrete question, *Risk Magazine* **15**(1), 115–116.
 - [13] Gatheral, J. (2006). *The Volatility Surface: A Practitioner's Guide*, Wiley, Hoboken, New Jersey.
 - [14] Haug, E., Haug, J. & Lewis, A. (2003). Back to basics: a new approach to the discrete dividend problem, *Wilmott Magazine* **5**, 37–47.
 - [15] Merton, R.C. (1973). Theory of rational option pricing, *The Bell Journal of Economics and Management Science* **4**, 141–183.
 - [16] Overhaus, M., Bermúdez, A., Buehler, H., Ferraris, A., Jorindson, C. & Lamnouar, A. (2007). *Equity Hybrid Derivatives*, Wiley, Hoboken, New Jersey.
 - [17] Piterbarg, V. (2007). Markovian projection method for volatility calibration, *Risk Magazine* **20**(4), 84–89.
 - [18] Rebonato, R. (2004). *Volatility and Correlation*, 2nd Edition, Wiley, Chichester, West Sussex.
 - [19] Rubinstein, M. (1994). Implied binomial trees, *Journal of Finance* **69**, 771–818.

Related Articles

Corridor Variance Swap; Dividend Modeling; Dupire Equation; Lookback Options; Model Calibration; Optimization Methods; Stochastic Volatility Interest Rate Models; Tikhonov Regularization; Variance Swap; Yield Curve Construction.

CHIYAN LUO & XINMING LIU

Dividend Modeling

A dividend is a portion of a company's earnings paid to its shareholders. In the process of dividend payment, the following stages are distinguished: (i) declaration date, when the dividend size and the ex-dividend date are announced; (ii) ex-dividend date, when the share starts trading net of dividend; (iii) record date, when holders eligible to dividend payment are identified; and (iv) payment date, when delivery is made. At the ex-dividend date, the stock price drops by an amount proportional to the size of the dividend; the proportionality factor depends on the tax regulations. There are a lot of issues, research streams, and approaches in dividend modeling; here the issue is considered mainly in the context of option pricing theory.

The usual way to price derivatives on dividend-paying stocks is to take a model for non-dividend-paying stocks and extend it to take the dividends into account. The dividends then are commonly modeled as (i) continuously paid dividend yield, (ii) proportional dividends (known fractions of the stock price) paid at known discrete times, or (iii) fixed dividends (known amounts), paid at known discrete times. It is also possible to model the dividend amounts and the dividend dates stochastically (though there is evidence that this has a negligible impact on vanilla options [10]). In fact, there is an alternative approach where the stochastic dividends are the primary quantities and the stock followed by option price are derived from these, which was pioneered in [9]. As usual, one has to choose the complexity of the model depending on dividend exposure of the derivative to be priced.

In practice, one comes across the notion of implied dividends: the value of the dividends (independent of how they are modeled) can be inverted from the synthetic forward or future contract; the fact that one can get quite different (from analyst predictions) numbers reflects various uncertainties. Among them are the sundry tax regulations in different countries for various market players, timing, and value of the dividends, just to name a few.

The impact of dividends can be illustrated, starting simply by adding a continuous dividend yield to the drift. For the sake of the simplicity of notations, it is

written in lognormal terms:

$$dS_t = (r - q)S_t dt + \sigma S_t dW_t \quad (1)$$

This approach is especially popular when modeling options on indexes, where dividend payments are numerous and spread through time. Another choice, of proportional amounts $d_i = f_i S_{t_i}$ paid at ex-dividend dates $t_1 < t_2 < \dots$ for single shares, can be justified by the fact that dividends tend to increase when a company is doing well, which is correlated with a high share price:

$$dS_t = (r - Q_t)S_t dt + \sigma S_t dW_t \quad \text{with} \\ Q_t = \sum_i \delta(t - t_i) f_i \quad (2)$$

In both these cases, the stock price at each time still has a lognormal distribution, so the prices of European options are given by straightforward modifications of the Black-Scholes (BS) pricing formula. This is no longer true, however, for discrete cash dividends:

$$dS_t = (rS_t - D_t) dt + \sigma S_t dW_t \quad \text{with} \\ D_t = \sum_i \delta(t - t_i) d_i \quad (3)$$

The stock price S_t jumps down with the amount of dividend d_i paid at time t_i and between the dividends it follows a geometric Brownian motion. In this setting, the stock price can become negative, but this is usually so unlikely that, in practice, it is not a problem. Still, one might want to use a more robust dividend policy in the model, such as capping the dividend at the stock price. Obviously, different dividend policies result in different option prices [7].

Impact on Option Pricing

To compute an option price under equation (3) the standard collection of numerical methods can be employed: finite difference (FD) method with jump conditions across ex-dividend date [11], Monte Carlo simulations, or nonrecombining trees [8]. There is no real closed-form solution with multiple dividends for European option under equation (3); however, several approximations are available. All of them are

based on bootstrapping, that is, repeatedly computing the convolution of the option value at one dividend date with the density kernel from that date to the previous dividend date and applying the jump condition at the dividend dates, starting from the payoff at maturity. One can use a piecewise linear or a more sophisticated approximation of the option value at each convolution step and enjoy having a finite sum of closed-form solutions. On the basis of the fact that diffusion preserves monotonicity and convexity, it can be shown that the result converges to the true value (unpublished work of Amaro de Matos *et al.*). Another choice of parameterization was made in [7]: at each step of the integration the option value is approximated by BS-like function where strike and volatility are adjusted to obtain the best fit. Such methods can be used for any underlying process where one can compute the density kernel (Green function, propagator) for the convolution, though it will probably be not much faster or more accurate than employing the standard finite difference method, especially in the case of multiple dividends. For the handling of American options, one can find an extensive list of references in [4] and the relation between early exercise and dividends is explained in **American Options; Finite Difference Methods for Early Exercise Options** or [8].

A Common Approach

As already pointed out in the discrete cash dividend model (3), the price of a European option has no closed-form solution and trees do not recombine. In order to remedy this, traders often split the stock price into a risky net-dividend part and a deterministic dividend part:

$$\begin{aligned} d\tilde{S}_t &= r\tilde{S}_t dt + \tilde{\sigma}\tilde{S}_t dt \\ S_t &= \tilde{S}_t + \tilde{D}_t \end{aligned} \quad (4)$$

$$\begin{aligned} \tilde{D}_t &= \tilde{D}_t^{\text{fut}} - \tilde{D}_t^{\text{past}} \\ \tilde{D}_t^{\text{fut}} &= \sum_{t < t_i \leq T} \alpha_i d_i e^{-r(t_i-t)} \\ \tilde{D}_t^{\text{past}} &= \sum_{0 < t_i \leq t} (1 - \alpha_i) d_i e^{-r(t_i-t)} \end{aligned} \quad (5)$$

Note that the dependence on the option maturity T in the notation $\tilde{D}_t^{\text{fut}}$ is suppressed. The most common

choices used by practitioners [3, 5] are

$$\alpha_i = 1 \quad \text{so} \quad \tilde{D}_t = \sum_{t < t_i \leq T} d_i e^{-r(t_i-t)} \quad (6)$$

$$\alpha_i = 0 \quad \text{so} \quad \tilde{D}_t = \sum_{0 < t_i \leq t} d_i e^{-r(t_i-t)} \quad (7)$$

$$\alpha_i = 1 - \frac{t_i}{T} \quad (8)$$

In this approach, the tree for $\tilde{S}_t$ is recombining (*see Binomial Tree* or [8]), and the price of a European option is again given by a BS type formula, where the spot and the strike are adjusted as $S_0 \rightarrow S_0 - \tilde{D}_0^{\text{fut}}$ and $K \rightarrow K + \tilde{D}_T^{\text{past}}$. Needless to say, however, the volatility for each of these processes will be different. Namely, choice (6) underestimates and choice (7) overestimates the volatility compared to the “true” model (3); the weighted choice (8) aims to minimize this effect.

Arbitrage Opportunities

In reference [1], it was shown that arbitrage opportunities exist in the most standard approach (6) if the volatility surface is continuously interpolated around ex-dividend dates. They apply a rough volatility adjustment to prevent the arbitrage opportunities. The following example demonstrates that the continuous interpolation of volatility around ex-dividend dates can lead to significant mispricing. Figure 1 shows

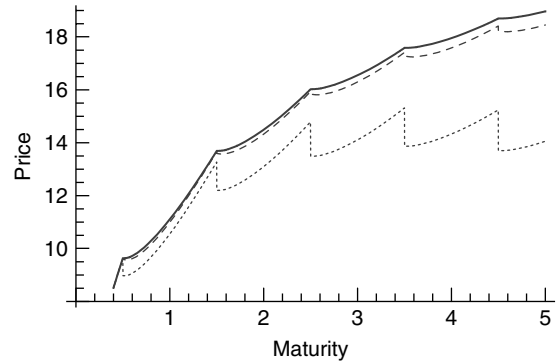


Figure 1 Price of an American call as a function of the time to maturity T for the following models: (3) (solid line), (6) (dotted line), and (6) with the volatility adjustment (10) (dashed line). The parameters are $S_0 = 100$, $K = 100$, $r = 0.05$, $\sigma = 0.3$, $d_i = 8$, and $t_i = i - \frac{1}{2}$

the price of an American call as a function of the time to maturity T for different models. Given a flat volatility, the prices under equation (6) jump down after an ex-dividend date, which is not realistic, because a risk-free profit can be locked in by selling an option maturing just before the ex-dividend date and by buying a similar option maturing just after the ex-dividend date. Although the mispricing under equation (6) is most evident for American options, it is equally present in the valuation of European options. Note that equation (4) produces continuous prices around dividend payments and the price differences between the two models increase dramatically with maturity (see [1, 5] for similar results). Another primitive pitfall with equation (6) is to use constant extrapolation of the implied volatility to longer maturities. For example, if the present value of dividends of a long-term contract is half of the current stock price, then with simple flat term structure extrapolation one will underestimate the volatility by a factor of at least one and a half [2].

Volatility Adjustments

To understand the difference in terms of volatility for the specified models, consider the local volatility model. Substitution of equation (4) into equation (3) yields the result that $\tilde{S}$ follows the process with local volatility

$$\tilde{\sigma}(\tilde{S}, t) = \left(1 + \frac{\tilde{D}_t}{\tilde{S}}\right) \sigma \quad (9)$$

It is handy to translate this result into implied terms [6]. Following the line of reasoning presented in [2], a slightly generalized result for the implied volatility can be derived:

$$\tilde{\sigma}^2 \approx \sigma^2 + \sigma \sqrt{\frac{\pi}{2T}} A \quad (10)$$

where

$$\begin{aligned} A = & 4e^{\frac{b_1^2}{2}-s} \sum_i \tilde{d}_i \left(\omega_i [\mathcal{N}(b_1) - \mathcal{N}(b_1 - \xi_i)] \right. \\ & \left. + \tilde{\omega}_i [\mathcal{N}(b_2) - \mathcal{N}(b_1 - \xi_i)] \right) \\ & + e^{\frac{c_1^2}{2}-2s} \sum_{i,j} \tilde{d}_i \tilde{d}_j \left(\omega_i \omega_j [\mathcal{N}(c_1) - \mathcal{N}(c_1 - \psi_{ij})] \right. \\ & \left. - \tilde{\omega}_i \tilde{\omega}_j [\mathcal{N}(c_2) - \mathcal{N}(c_1 - \psi_{ij})] \right. \\ & \left. + \Omega_{ij} [\mathcal{N}(c_1 - \Psi_{ij}) - \mathcal{N}(c_1 - \psi_{ij})] \right) \quad (11) \end{aligned}$$

with $s = \ln(S_0 - \tilde{D}_0^{\text{fut}})$, $k = \ln(K + \tilde{D}_T^{\text{past}}) - rT$, $x_t = (s + (k - s)\frac{t}{T}) + rt$, $y_t = \frac{t(T-t)}{T}$, $\tilde{\omega}_i = 1 - \omega_i$, $\tilde{d}_i = d_i e^{-r t_i}$, $a = \frac{s-k}{\sigma\sqrt{T}}$, $b_{1/2} = a \pm \frac{\sigma\sqrt{T}}{2}$, $c_{1/2} = a \pm \sigma\sqrt{T}$, $\xi_i = \sigma \frac{t_i}{\sqrt{T}}$, $\psi_{ij} = 2 \min(\xi_i, \xi_j)$, $\Psi_{ij} = 2 \max(\xi_i, \xi_j)$, $\Omega_{ij} = \omega_i \tilde{\omega}_j 1_{t_i > t_j} + \omega_j \tilde{\omega}_i 1_{t_j > t_i}$.

This gives a very simple and quick way of switching between models for trading practice as well as for understanding the essence of the feature. Moreover, it can be used independently of the option type. To give an idea of the quantitative value of presented methods, some numerical results with and without volatility adjustment are summarized in Table 1. All results are compared to the numerical solution calculated with the FD method. Bootstrapping with the piecewise linear interpolation converges to FD with sufficiently many points; HHL approximation of [7] differs a bit, probably due to the fact that the BS-like formula cannot exactly fit the shape of the integrand. Clearly, even in approximate form, equation (10) gives a fair correction in all cases, performing especially well for the weighted model (8).

When pricing equity derivatives, one should be aware that these instruments can be sensitive to

Table 1 European call prices with parameter set of Figure 1 for different strikes. HHL refers to the approximation of [7]; FD to finite difference and BS to closed-form solution

Strike	FD for (3)	HHL for (3)	BS for (6)	BS for (6) with (10)	BS for (8)	BS for (8) with (10)
50	33.509	33.641	29.908	33.312	33.547	33.497
80	22.482	22.559	17.846	22.414	22.304	22.473
100	17.393	17.428	12.772	17.404	17.102	17.388
120	13.573	13.575	9.250	13.644	13.209	13.573
150	9.511	9.479	5.836	9.635	9.099	9.515

dividends. Examples are exotic options on stocks and derivatives involving realized volatility, such as variance swaps (see **Variance Swap**), volatility swaps (see **Volatility Swaps**), correlation swaps (see **Correlation Swap**), and gamma swaps (see **Gamma Swap**). This sensitivity determines the required sophistication in dividend modeling. Adding dividends to a stock price process may seem trivial at first glance, but one has to be careful in setting the model parameters. The resulting model can then be solved by the usual methods. For the plain vanilla option with dividends, a number of numerical approximations have been developed.

References

- [1] Bener, R. & Vorst, T. (2001). Options on dividends paying stocks, in *Recent Developments in Mathematical Finance*, World Scientific Printers, Shanghai, pp. 204–217.
- [2] Bos, R., Gairat, A. & Shepeleva, A. (2003). Dealing with discrete dividends, *Risk* **16**, 109–112.
- [3] Bos, M. & Vandermark, S. (2002). Finessing fixed dividends, *Risk* **15**, 157–158.
- [4] Cassimon, D., Engelen, P.J., Thomassen, L. & Van Wouwe, M. (2007). Closed-form valuation of American call options on stocks paying multiple dividends, *Finance Research Letters* **4**, 34–48.
- [5] Frishling, V. (2002). A discrete question, *Risk* **15**, 115–116.
- [6] Gatheral, J. (2006). *The Volatility Surface*, John Wiley & Sons, Hoboken, pp. 13–14.
- [7] Haug, E., Haug, J. & Lewis, A. (2003). Back to basics: a new approach to the discrete dividend problem, *Wilmott Magazine* (September), 37–47.
- [8] Hull, J.C. (2006). *Options, Futures and Other Derivatives*, 6th Edition, Prentice-Hall, Upper Saddle River.
- [9] Korn, R. & Rogers, L.C.G. (2005). Stocks paying discrete dividends: modeling and option pricing, *Journal of Derivatives* **13**(2), 44–48.
- [10] Kruchen, S. (2005). *Dividend Risk*, thesis, Uni/ETH, Zürich.
- [11] Tavella, D. & Randall, C. (2000). *Pricing Financial Instruments. The Finite Difference Method*, John Wiley & Sons, New York.

Related Articles

American Options; Finite Difference Methods for Early Exercise Options; Local Volatility Model; Monte Carlo Simulation.

ANNA SHEPELEVA & ALAIN VERBERKMOES

Implied Volatility: Volvol Expansion

In order to calibrate stochastic volatility models, it is convenient to have an accurate analytical formula or approximation for call options. However, deriving such a formula is not always an easy task. In the Heston model, the most popular technique involves numerical integration, which is necessarily time consuming. The main idea is to apply a perturbation method to the volvol parameter, calculating the first and second order of the difference between a stochastic volatility model and a Black–Scholes model. In general case, we can reduce the integration of the exact formula to some simpler integration.

Consider the following two-factor stochastic volatility model:

$$dS = (r - d)Sdt + \sigma SdB_t \quad (1)$$

$$dV = b(V)dt + \epsilon v(V)dW \quad (2)$$

$$dBdW = \rho(V)dt \quad (3)$$

where r is the short rate, d is the dividend yield, ϵ is a constant, and $b(V)$ and $v(V)$ are independent of ϵ . We assume parameters r and d to be constant for the sake of simplicity. The series expansion consists in writing the option price formula as a series in ϵ .

Fourier methods (*see* **Fourier Methods in Options Pricing**) tell us that the call option price is given by

$$C(S, V, T) = Se^{-dT} - \frac{Ke^{-rT}}{2\pi} \int_{i/2-\infty}^{i/2+\infty} \exp(-iuX) \frac{\Phi(u, V, T)}{u(u-i)} du \quad (4)$$

where $X = \ln(S/K) + (r - d)T$.

$$\frac{\partial f}{\partial \tau} = \frac{1}{2}\epsilon^2 v^2(V) \frac{\partial^2 f}{\partial V^2} + \left[b(V) - iu\epsilon\rho(V)v(V)\sqrt{V} \right]$$

$$\times \frac{\partial f}{\partial V} - \frac{u(u-i)}{2} Vf \quad (5)$$

We look for solutions which can be written as a power series of ϵ .

$$f(u, V, T) = f^{(0)}(u, V, T) + \epsilon f^{(1)}(u, V, T) + \epsilon^2 f^{(2)}(u, V, T) \quad (6)$$

Thus, we can obtain the power series of the call price using either equation (4) directly

$$C(u, V, T) = C^{(0)}(u, V, T) + \epsilon C^{(1)}(u, V, T) + \epsilon^2 C^{(2)}(u, V, T) \quad (7)$$

or expanding first the implied volatility

$$V_{\text{imp}}(u, V, T) = V_{\text{imp}}^{(0)}(u, V, T) + \epsilon V_{\text{imp}}^{(1)}(u, V, T) + \epsilon^2 V_{\text{imp}}^{(2)}(u, V, T) \quad (8)$$

and then plugging it into the option price with Black–Scholes formula.

As a matter of fact, these two methods differ significantly. The former—denoted by series A in the remainder of this article—gives the call price first and implied volatility while the latter—denoted by series B—gives the implied volatility and call price is obtained afterward. Nevertheless, in most cases, there is a slight numerical difference between the two series. However, regarding far out-of-the money options, the two series give different results as shown in Figure 1. Empirical evidence shows that series B is very often better than series A. Even though this is not a general rule, series B should be usually preferred over series A.

We now explicitly compute the two series for the following model. In particular, this model encompasses the Heston model for the special case $\phi = 0.5$.

$$dS = (r - d)Sdt + \sigma SdB_t \quad (9)$$

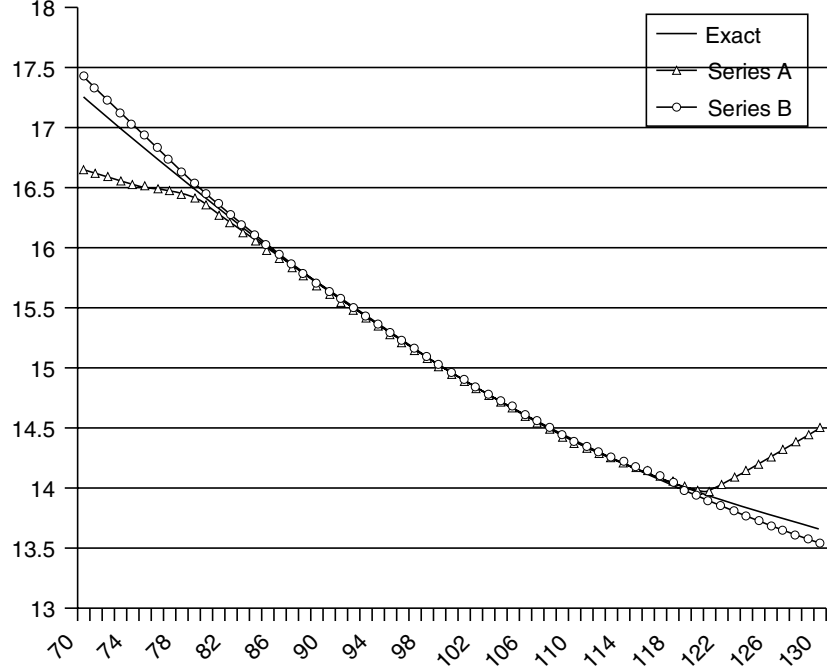


Figure 1 Series A (expansion on price), series B (expansion on implied volatility), and exact volatility

$$dV = (\omega - \theta V)dt + \epsilon V^\phi dW \quad (10)$$

$$dBdW = \rho(V)dt \quad (11)$$

First, the expansion is on the fundamental transform of the closed formula, which is presented by $\Phi(u, V, T)$ in equation (4). The idea is that we can expand this function into a simpler form, so that the integration in the equation (4) can be reduced to analytic form. Here, we do not discuss the detailed derivation but only give the result. Interested readers can refer to [2]. For series A, the expansion on price is

$$\begin{aligned} C(S, V, \tau) &= c(S, v, \tau) + \epsilon \tau^{-1} J^{(1)} \tilde{R}^{(1,1)} c_V(S, v, \tau) \\ &+ \epsilon^2 \left\{ \tau^{-1} J^2 + \tau^{-2} J^{(3)} \tilde{R}^{(2,0)} \right. \\ &\left. + \tau^{-1} J^{(4)} \tilde{R}^{(1,2)} + \frac{\tau^{-2}}{2} (J^{(1)})^2 \tilde{R}^{(2,2)} \right\} \\ &\times c_V(S, v, \tau) + O(\epsilon^3) \end{aligned} \quad (12)$$

For series B, the expansion on price is

$$\begin{aligned} V^{\text{imp}} &= v(S, v, \tau) + \epsilon \tau^{-1} J^{(1)} \tilde{R}^{(1,1)} \\ &+ \epsilon^2 \left\{ \tau^{-1} J^2 + \tau^{-2} J^{(3)} \tilde{R}^{(2,0)} \right. \\ &\left. + \tau^{-1} J^{(4)} \tilde{R}^{(1,2)} + \frac{\tau^{-2}}{2} (J^{(1)})^2 \right. \\ &\left. \times \left[\tilde{R}^{(2,2)} - \left(\tilde{R}^{(1,1)} \right)^2 \tilde{R}^{(2,0)} \right] \right\} \\ &+ O(\epsilon^3) \end{aligned} \quad (13)$$

In the above formulae, term $c(S, v, \tau)$ presents the corresponding Black-Scholes price. When volvol $\epsilon = 0$, the stochastic model reduces to a Black-Scholes model. v here is the equivalent variance for Black-Scholes, which is basically the integration of the variance from 0 to τ .

The functions $\tilde{R}^{(p,q)}$ and $J^{(s)}$ are the derivative ratios and integration, respectively. Here, for academic and practitioners' interest their expressions are listed.

$$dv_t = \kappa(\theta_t - v_t)dt + \xi_t \sqrt{v_t} dB_t, \quad v_0 \quad (20)$$

$$d\langle W, B \rangle_t = \rho_t dt \quad (21)$$

$$\begin{aligned} \tilde{R}^{(2,0)} &= \tau \left[\frac{1}{2} \frac{X^2}{Y^2} - \frac{1}{2Z} - \frac{1}{8} \right], \quad \tilde{R}^{(1,1)} = \left[-\frac{X}{Z} + \frac{1}{2} \right] \\ \tilde{R}^{(1,2)} &= \left[\frac{X^2}{Z^2} - \frac{X}{Z} - \frac{1}{4Z}(4-Z) \right], \quad \tilde{R}^{(2,2)} = \tau \left[\frac{1}{2} \frac{X^4}{Z^4} - \frac{1}{2} \frac{X^3}{Z^3} - 3 \frac{X^2}{Z^3} + \frac{1}{8} \frac{X}{Z^2}(12+Z) + \frac{1}{32} \frac{1}{Z^2}(48-Z^2) \right] \end{aligned} \quad (14)$$

with $Z = V\tau$ and

Here, we have adjusted to risk neutral probability; as a consequence, it introduces the drift term in spot process by the change of probability.

$$J^{(1)}(V, \tau) = \frac{\rho}{\theta} \int_0^\tau (1 - e^{-\theta(\tau-s)}) \left[\frac{\omega}{\theta} + e^{-\theta s} \left(V - \frac{\omega}{\theta} \right) \right]^{\phi+\frac{1}{2}} ds, \quad J^{(2)}(V, \tau) = 0 \quad (15)$$

$$J^{(3)}(V, \tau) = \frac{1}{2\theta^2} \int_0^\tau (1 - e^{-\theta(\tau-s)})^2 \left[\frac{\omega}{\theta} + e^{-\theta s} \left(V - \frac{\omega}{\theta} \right) \right]^{2\phi} ds \quad (16)$$

$$J^{(4)}(V, \tau) = \left(\phi + \frac{1}{2} \right) \int_0^\tau \left[\frac{\omega}{\theta} + e^{-\theta(\tau-s)} \left(V - \frac{\omega}{\theta} \right) \right]^{\phi+\frac{1}{2}} J^{(6)}(V, \tau) ds \quad (17)$$

$$\text{with } J^{(6)}(V, \tau) ds = \int_0^\tau (e^{-\theta(\tau-s)} - e^{-\theta s}) \left[\frac{\omega}{\theta} + e^{-\theta(\tau-u)} \left(V - \frac{\omega}{\theta} \right) \right]^{\phi-\frac{1}{2}} du \quad (18)$$

Example of Volvol Expansion: Heston Model

We will now show the expansion of volvol for the Heston model (*see Heston Model*). By the asymptotic expansion, we can finally obtain an approximate analytic formula for the European call option. This work comes from the result of Benhamou *et al.* [1]. Consider a Heston model

$$dX_t = \sqrt{v_t} dW_t - \frac{v_t}{2} dt, \quad X_0 = x_0 \quad (19)$$

To expand the model, we add ϵ in the model.

$$dX_t^\epsilon = \sqrt{v_t^\epsilon} dW_t - \frac{v_t^\epsilon}{2} dt, \quad X_0^\epsilon = x_0 \quad (22)$$

$$dv_t^\epsilon = \kappa(\theta_t - v_t^\epsilon)dt + \epsilon \xi_t \sqrt{v_t^\epsilon} dB_t, \quad v_0^\epsilon = v_0 \quad (23)$$

Now, we will expand the European call option price formula with respect to ϵ . Note that, when $\epsilon = 0$, we have a Black-Scholes model; while $\epsilon = 1$, we have a Heston model. We already have the closed

formula of Black–Scholes for $\epsilon = 0$. We expand at $\epsilon = 0$, and let $\epsilon = 1$ to obtain the approximate formula. In mathematical language, this can be written as follows:

$$P_{\text{Heston}} = P_{\text{BS}} + \mathbb{E} \left[\frac{\partial P_{\text{BS}}}{\partial \epsilon} \right] + \frac{1}{2} \mathbb{E} \left[\frac{\partial^2 P_{\text{BS}}}{\partial \epsilon^2} \right] + \mathcal{E} \quad (24)$$

Here, we will take another approximation to simulate the partial derivatives in the above equation by the linear combination of the Greek letters of Black–Scholes. Here, the idea is that use of the chain rule in the derivative can result in $\frac{\partial P_{\text{BS}}}{\partial \epsilon} = \frac{\partial P_{\text{BS}}}{\partial S} \frac{\partial S}{\partial \epsilon} + \frac{\partial P_{\text{BS}}}{\partial \sigma} \frac{\partial \sigma}{\partial \epsilon}$. The same idea holds for the second derivative.

$$P_{\text{Heston}} = P_{\text{BS}}(x_0, \text{var}_T) + \sum_{i=1}^2 a_{i,T} \frac{\partial^{i+1} P_{\text{BS}}}{\partial x^i y} (x_0, \text{var}_T) + \sum_{i=0}^1 b_{2i,T} \frac{\partial^{2i+2} P_{\text{BS}}}{\partial x^{2i} y^2} (x_0, \text{var}_T) + \mathcal{E} \quad (25)$$

We refer to [1] for proofs and intermediate derivation. The parameters in the formula are as follows:

$$\begin{aligned} \text{var}_T &= m_0 v_0 + m_1 \theta, \quad a_{1,T} = \rho \xi (p_0 v_0 + p_1 \theta) \\ a_{2,T} &= (\rho \xi)^2 (q_0 v_0 + q_1 \theta), \quad b_{0,T} = \xi^2 (r_0 v_0 + r_1 \theta) \end{aligned} \quad (26)$$

$$\text{var}_T = \int_0^T v_{0,t} dt \quad (27)$$

$$m_0 = \frac{e^{-\kappa T} (-1 + e^{\kappa T})}{\kappa},$$

$$m_1 = T - \frac{e^{-\kappa T} (-1 + e^{\kappa T})}{\kappa},$$

$$p_0 = \frac{e^{-\kappa T} (-\kappa T + e^{\kappa T} - 1)}{\kappa^2},$$

$$p_1 = \frac{e^{-\kappa T} (\kappa T + e^{\kappa T} (\kappa T - 2) + 2)}{\kappa^2},$$

$$q_0 = \frac{e^{-\kappa T} (-\kappa T (\kappa T + 2) + 2e^{\kappa T} - 2)}{2\kappa^3},$$

$$q_1 = \frac{e^{-\kappa T} (2e^{\kappa T} (\kappa T - 3) + \kappa T (\kappa T + 4) + 6)}{2\kappa^3},$$

$$r_0 = \frac{e^{-2\kappa T} (-4e^{\kappa T} \kappa T + 2e^{2\kappa T} - 2)}{4\kappa^3},$$

$$r_1 = \frac{e^{-2\kappa T} (4e^{\kappa T} (\kappa T + 1) + e^{2\kappa T} (2\kappa T - 5) + 1)}{4\kappa^3} \quad (28)$$

The advantage is that there is no integration in the approximate formula. So the calculations are done much faster than in the exact formula. We will discuss this point in the section Numerical Results.

The error in the approximation is estimated as $\mathcal{E} = O([\xi_{\text{Sup}} \sqrt{T}]^3 \sqrt{T})$.

Numerical Results

We test the approximate formula with the following strikes. We take strikes from 70% to 130% for short maturity and 10% to 730% for long maturity. Implied Black–Scholes volatilities of the closed formula, of the approximation formula, and related errors (in bp), are expressed as a function of maturities in fractions of years and relative strikes. The values of the parameters are as follows: $\theta = 6\%$, $\kappa = 3$, $\xi = 30\%$, and $\rho = 0\%$. Except for short maturity plus very small strikes, where we observe the largest difference (18.01 bp); the difference is less than 5 bp (1 bp = 0.01%) in almost all other cases. With regard to the speed of calculation, the approximate formula is about 100 times quicker than the exact formula (with the optimization in integral).

References

- [1] Benhamou, E., Gobet, E. & Miri, M. (2009). Time dependent Heston model, *SIAM Journal on Financial Mathematics*.
- [2] Lewis, A.L. (2000). *Option Valuation Under Stochastic Volatility: With Mathematica Code*. February 2000.

Related Articles

Heavy Tails; Heston Model; Hull–White Stochastic Volatility Model; Implied Volatility in Stochastic Volatility Models; Implied Volatility Surface;

Model Calibration; Partial Differential Equations; SABR Model; Stochastic Volatility Models; Stylized Properties of Asset Returns.

ZAIZHI WANG & MOHAMMED MIRI

Implied Volatility: Long Maturity Behavior

We discuss some properties of implied volatility surfaces (*see Implied Volatility Surface*) for large times to maturity. The key result is that the implied volatility smile flattens at long maturities, independently of the model of the underlying asset price, so long as there exists an equivalent martingale measure. An asymptotic formula for the long implied volatility is given and illustrated by examples from stochastic volatility models. The dynamics of the long implied volatility are shown to be almost surely non-decreasing as a function of calendar time, in complete analogy with the Dybvig *et al.* [2] theorem for long interest rates. For a more in-depth treatment of these issues, consult [3, 4], and [7].

Flattening of the Smile

To set up notation and present the main result, we consider in this section a market in which the riskless interest rate is zero, and the underlying stock pays no dividends. The results described here are valid for price processes with either continuous or discontinuous sample paths, or even discrete-time models.

We let $S = (S_t)_{t \geq 0}$ be a nonnegative martingale with $S_0 > 0$ modeling the price of a given stock under a fixed risk-neutral measure. All calculations will be performed with respect to this measure, so we do not include it in the notation.

Introduce a function $C_{BS} : \mathbb{R} \times [0, \infty) \rightarrow [0, 1)$ by

$$C_{BS}(k, v) = \begin{cases} \Phi\left(-\frac{k}{\sqrt{v}} + \frac{\sqrt{v}}{2}\right) - e^k \Phi\left(-\frac{k}{\sqrt{v}} - \frac{\sqrt{v}}{2}\right) & \text{if } v > 0 \\ (1 - e^k)^+ & \text{if } v = 0 \end{cases} \quad (1)$$

where $a^+ = \max\{a, 0\}$ denotes the positive part of the real number a as usual. Since $v \mapsto C_{BS}(k, v)$ is strictly increasing for each $k \in \mathbb{R}$, we now define the Black-Scholes implied volatility $\Sigma(k, T)$ for log moneyness k and maturity T as the unique

nonnegative solution to the equation

$$\mathbb{E} \left[\left(\frac{S_T}{S_0} - e^k \right)^+ \right] = C_{BS}(k, T \Sigma(k, T)^2) \quad (2)$$

Note that we are using the convention that the log moneyness k corresponds to the strike $K = S_0 e^k$.

The main result is that the implied volatility smile flattens at long maturities:

Theorem 1 *For any $M > 0$, we have*

$$\lim_{T \rightarrow \infty} \sup_{k_1, k_2 \in [-M, M]} |\Sigma(k_2, T) - \Sigma(k_1, T)| = 0 \quad (3)$$

That the implied volatility smile flattens at long maturities seems to be a folk theorem, and has been verified for various models for which the long implied volatility can be calculated explicitly. The result is sometimes attributed erroneously to the central limit theorem, but it is true in complete generality without any notion of mean reversion of the spot volatility process. Indeed, Theorem 1 contains no assumption on the dynamics of the stock price other than that it is a nonnegative martingale. Also note that we have not even assumed that $\lim_{T \rightarrow \infty} \Sigma(k, T)$ exists for any k .

A proof of the flattening of the implied volatility smile under some mild regularity assumptions appears in [1]. A proof of Theorem 1 in the form that it appears here can be found in [9].

It turns out that the rate of flattening can be precisely bounded:

Theorem 2

1. *For any $0 \leq k_1 < k_2$, we have*

$$\frac{\Sigma(k_2, T)^2 - \Sigma(k_1, T)^2}{k_2 - k_1} \leq \frac{4}{T} \quad (4)$$

2. *For any $k_1 < k_2 \leq 0$, we have*

$$\frac{\Sigma(k_2, T)^2 - \Sigma(k_1, T)^2}{k_2 - k_1} \geq -\frac{4}{T} \quad (5)$$

3. *If $S_t \rightarrow 0$ in probability as $t \rightarrow \infty$, for any $M > 0$, we have*

$$\limsup_{T \rightarrow \infty} \sup_{\substack{k_1, k_2 \in [-M, M] \\ k_1 \neq k_2}} T \left| \frac{\Sigma(k_2, T)^2 - \Sigma(k_1, T)^2}{k_2 - k_1} \right| \leq 4 \quad (6)$$

The inequality in part 3 of Theorem 2 is sharp, as there exists a martingale $(S_t)_{t \geq 0}$ such that $S_t \rightarrow 0$ in probability and such that

$$T \frac{\partial}{\partial k} \Sigma(k, T)^2 \rightarrow -4 \quad (7)$$

as $T \rightarrow \infty$ uniformly for $k \in [-M, M]$. A proof of Theorem 2 is in [9].

Remark The condition $S_t \rightarrow 0$ in probability appearing in part 3 of Theorem 2 has a natural financial interpretation. Indeed, we have $S_t \rightarrow 0$ in probability (equivalently, almost surely) if and only if $C(K, T) \rightarrow S_0$ as $T \rightarrow \infty$ for some $K > 0$ (equivalently, for all $K > 0$) where

$$C(K, T) = \mathbb{E}[(S_T - K)^+] \quad (8)$$

is the price of a European call option. Since the long maturity call prices converge to stock price in many models of interest (including, of course, the Black–Scholes model), we see that the assumption $S_t \rightarrow 0$ is not particularly onerous.

In fact, since $(S_t)_{t \geq 0}$ is a nonnegative martingale, it must converge almost surely to some random variable S_∞ by the martingale convergence theorem. If $S_\infty > 0$ with positive probability, then $\lim_{T \rightarrow \infty} \sqrt{T} \Sigma(k, T)$ exists and is finite for each k , and hence $\lim_{T \rightarrow \infty} \Sigma(k, T) = 0$.

A Representation Formula

Now that we know that the volatility smile flattens in the limit as the maturity goes to infinity, we can study the behavior of the long implied volatility.

Theorem 3 *For any $M > 0$, we have*

$$\lim_{T \rightarrow \infty} \sup_{k \in [-M, M]} \left| \Sigma(k, T) - \left(-\frac{8}{T} \log \mathbb{E}[S_T \wedge 1] \right)^{1/2} \right| = 0 \quad (9)$$

where $a \wedge b = \min\{a, b\}$ as usual. In particular, we have the following representation formula

$$\Sigma(\infty) = \lim_{T \rightarrow \infty} \left(-\frac{8}{T} \log \mathbb{E}[S_T \wedge 1] \right)^{1/2} \quad (10)$$

whenever the limit exists.

Formula (10) of Theorem 3 can be found, for instance, in [8]. It can be used to calculate the long implied volatility for some examples.

Example (Exponential Lévy Models (see **Exponential Lévy Models**)). The simple inequality

$$\mathbb{1}_{[1, \infty)}(x) \leq x \wedge 1 \leq x^p \quad (11)$$

which holds for all $0 \leq p \leq 1$ and $x \geq 0$, gives the bound

$$\limsup_{T \rightarrow \infty} \sup_{p \in [0, 1]} \left(-\frac{8}{T} \log \mathbb{E}[S_T^p] \right)^{1/2} \leq \Sigma(\infty) \leq \liminf_{T \rightarrow \infty} \left(-\frac{8}{T} \log \mathbb{P}[S_T \geq 1] \right)^{1/2} \quad (12)$$

If $(\log(S_t))_{t \geq 0}$ has independent identically distributed increments, then the above bounds hold with equality by the large deviation principle. Indeed, let $(L_t)_{t \geq 0}$ be a Lévy process with cumulant-generating function

$$\Lambda(p) = \log \mathbb{E}(e^{pL_1}) \quad (13)$$

such that $\Lambda(1) < \infty$, and model the stock price by the martingale $S_t = e^{L_t - t\Lambda(1)}$. Then the long implied volatility satisfies

$$\Sigma(\infty)^2 = 8 \sup_{p \in [0, 1]} \{p\Lambda(1) - \Lambda(p)\} \quad (14)$$

which is eight times the Legendre transform of the cumulant generating function evaluated at $\Lambda(1)$.

Example (*Stochastic volatility model.*) Lewis [8] has proposed a saddle-point approximation method for calculating long implied volatility in stochastic volatility models. For instance, suppose that the asset price satisfies the following system of stochastic differential equations:

$$dS_t = S_t \sqrt{V_t} dW_t \quad (15)$$

$$dV_t = \kappa(\theta - V_t) dt + \eta \sqrt{V_t} dZ_t \quad (16)$$

where κ , θ , and η are real constants, and $(W_t)_{t \geq 0}$ and $(Z_t)_{t \geq 0}$ are correlated standard Wiener processes with $\langle W, Z \rangle_t = \rho t$. Lewis [8] has shown that the long implied volatility is given by the following formula:

$$\Sigma(\infty)^2 = \frac{4\kappa\theta}{(1 - \rho^2)\eta^2} \times \left[\sqrt{(2\kappa - \rho\eta)^2 + (1 - \rho^2)\eta^2} - (2\kappa - \rho\eta) \right] \quad (17)$$

See [5] for further asymptotics of stochastic volatility models based on this method, and see [4] for asymptotics based on perturbation methods.

Long Implied Volatility Cannot Fall

In many models of interest, the long implied volatility, if it exists, is constant as a function of the calendar time. However, the long implied volatility need not be a constant in general. In this section, we consider the dynamics of the long implied volatility, and in fact, we will see that the long implied volatility can never fall. In this section, we also assume that the stock price is strictly positive, rather than merely non-negative. We define the implied volatility $\Sigma_t(k, \tau)$ for log moneyness k and time to maturity τ as the unique nonnegative $\mathcal{F}_t$ -measurable random variable that satisfies

$$\mathbb{E} \left[\left(\frac{S_{t+\tau}}{S_t} - e^k \right)^+ \right] = C_{BS}(k, \tau \Sigma_t(k, \tau)^2) \quad (18)$$

The following theorem was proved in [9].

Theorem 4 For all k_1, k_2 and $0 \leq s \leq t$ we have

$$\limsup_{\tau \rightarrow \infty} \Sigma_t(k_1, \tau) - \Sigma_s(k_2, \tau) \geq 0 \quad (19)$$

almost surely.

This result is an exact analog of the Dybvig–Ingersoll–Ross theorem that long zero-coupon rates never fall. See [6] for a nice proof of this fact.

Extensions

The previous discussion has considered the case where the stock pays no dividend and the risk-free interest rate is zero. In the general case, a stock pays a dividend and there is a cost to borrow money. The situation is usually modeled as follows. Let S_t be the stock price, let D_t be the cumulative dividends, and let B_t be the price of a numéraire asset such as a bank account at time t . There is no arbitrage if there exists a probability measure such that the process

$$\frac{S_t}{B_t} + \int_0^t \frac{dD_s}{B_s} \quad (20)$$

is a martingale. In the case of proportional continuous dividends $dD_t = \delta S_t dt$ and constant interest rate

$dB_t = r B_t dt$, there is no arbitrage if $\tilde{S}_t = S_t e^{(\delta-r)t}$ defines a martingale. In this case, everything from above applies if we define the implied volatility by

$$\mathbb{E} \left[\left(\frac{\tilde{S}_T}{\tilde{S}_0} - e^k \right)^+ \right] = C_{BS}(k, T \Sigma(k, T)^2) \quad (21)$$

where the log-moneyness parameter k now corresponds to the strike $K = S_0 e^{k+(r-\delta)T}$.

However, it is unclear which of the above results can be suitably extended to the general case with arbitrary increasing adapted processes $(D_t)_{t \geq 0}$ and $(B_t)_{t \geq 0}$.

References

- [1] Carr, P. & Wu, L. (2003). The finite moment log stable process and option pricing, *Journal of Finance* **58**(2), 753–778.
- [2] Dybvig, P., Ingersoll, J. & Ross, S. (1996). Long forward and zero-coupon rates can never fall, *Journal of Business* **69**, 1–25.
- [3] Gatheral, J. (2006). *The Volatility Surface: A Practitioner's Guide*, John Wiley & Sons, Hoboken, NJ.
- [4] Fouque, J.-P., Papanicolaou, G. & Sircar, K.R. (2000). *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press.
- [5] Jacquier, A. (2007). *Asymptotic Skew Under Stochastic Volatility*, Pre-print, Birkbeck College, University of London.
- [6] Hubalek, F., Klein, I. & Teichmann, J. (2002). A general proof of the Dybvig–Ingersoll–Ross theorem: long forward rates can never fall, *Mathematical Finance* **12**(4), 447–451.
- [7] Lee, R. (2004). Implied volatility: statics, dynamics, and probabilistic interpretation, in *Recent Advances in Applied Probability*, R. Baeza-Yates, et al., eds, Springer-Verlag, Springer, New York, 241–268.
- [8] Lewis, A. (2000). *Option Valuation Under Stochastic Volatility*, Finance Press, Newport Beach.
- [9] Rogers, L.C.G. & Tehranchi, M.R. (2008). *Can the Implied Volatility Surface Move by Parallel Shifts?* Pre-print, University of Cambridge.

Related Articles

Exponential Lévy Models; Heston Model; Implied Volatility Surface; Moment Explosions.

MICHAEL R. TEHRANCHI

SABR Model

The SABR model [4] is a stochastic volatility (*see Stochastic Volatility Models*) model in which the forward asset price follows the dynamics in a forward measure $\mathbb{P}^T$:

$$df_t = a_t C(f_t) dW_t \quad (1)$$

$$da_t = \nu a_t dZ_t, \quad \alpha \equiv a_0 \quad (2)$$

$$dW_t dZ_t = \rho dt, \quad C(f) \equiv f^\beta, \quad \beta \in [0, 1) \quad (3)$$

The stochastic volatility a_t , is described by a geometric Brownian motion. The model depends on four parameters: α , ν , ρ , and β . By using singular perturbation techniques, Hagan *et al.* [4] obtained a closed-form formula for the implied volatility $\sigma_{BS}(\tau, K)$, at the first-order in the maturity τ . Here, we display a corrected version of the formula [8]:

$$\sigma_{BS}(\tau, K) = \frac{\nu \ln \frac{f_0}{K}}{\hat{x}(\zeta)} (1 + \sigma_1(f_{av})\tau) \quad (4)$$

with

$$\begin{aligned} \hat{x}(\zeta) &= \ln \left(\frac{\sqrt{1 - 2\rho\zeta + \zeta^2} - \rho + \zeta}{1 - \rho} \right) \\ \zeta &= \frac{\nu}{\alpha} \int_K^{f_0} \frac{dx}{C(x)}, \quad f_{av} = \sqrt{f_0 K} \\ \sigma_1(f) &= \frac{\alpha \nu \partial_f C(f) \rho}{4} + \frac{2 - 3\rho^2}{24} \nu^2 + \frac{(\alpha C(f))^2}{24} \\ &\quad \times \left(\frac{1}{f^2} + \frac{2\partial_{ff} C(f)}{C(f)} - \left(\frac{\partial_f C(f)}{C(f)} \right)^2 \right) \end{aligned} \quad (5)$$

Though this formula is popular, volatility does not mean-revert in the underlying SABR model, so for given α , ν , β , and ρ , the SABR formula cannot simultaneously calibrate to the implied volatility smile at more than one expiry.

By mapping Hagan *et al.*'s computations into a geometrical framework based on the heat kernel expansion, approximate implied volatility formulae may be derived for more general stochastic volatility models (SVMs), in particular for the models where

volatility mean-reverts. The use of geometrical methods in quantitative finance originates from [1, 2] and was investigated in detail in [5, 6, 7].

A More General Stochastic Process

In the following, we will assume arbitrary local volatility functions $C(\cdot)$ and a general time-homogeneous one-dimensional stochastic differential equation (SDE) for the stochastic volatility process

$$da_t = b(a_t) dt + \sigma(a_t) dZ_t \quad (6)$$

In principle, $b(\cdot)$ and $\sigma(\cdot)$ could depend on the forward f as well, but the models we are interested in here do not exhibit this additional dependence.

This strategy for computing the short-time implied volatility asymptotics induced by the SVM involves two main steps:

- Derive the short-time limit of the *effective* local volatility function. The computation involves the use of the heat kernel expansion.
- Derive an approximate expression for the implied volatility corresponding to this effective local volatility function.

Effective Local Volatility Model

The square of the Dupire effective local volatility function (*see Model Calibration*) [3] is equal to the mean of the square of the stochastic volatility when the forward is fixed to the strike

$$\begin{aligned} \sigma_{loc}(t, K)^2 &= C(K)^2 \mathbb{E}^{\mathbb{P}}[a_t^2 | f_t = K] \\ &= C(K)^2 \frac{\int_{-\infty}^{\infty} a^2 p(t, K, a | f_0, \alpha) da}{\int_{-\infty}^{\infty} p(t, K, a | f_0, \alpha) da} \end{aligned} \quad (7)$$

where $p(t, K, a | f, \alpha)$ is the conditional probability density for the forward and the volatility at time t . As we now proceed to explain, $p(t, K, a | f, \alpha)$ is the fundamental solution of a heat kernel equation depending on two important geometrical quantities: first a metric tensor in equation (9), which is the inverse of the local covariance matrix and second an

Abelian connection in equation (10) which depends on the drift $b(a)$.

Heat Kernel Expansion

A short-time expansion of the density for a multi-dimensional $It\hat{o}$ diffusion process can be obtained using the heat kernel expansion: the Kolmogorov equation is rewritten as a heat kernel equation on an n -dimensional Riemannian manifold endowed with an Abelian connection as explained in [7].

Suppose the stochastic equations are written as

$$dx^\mu = b^\mu(x) dt + \sigma^\mu(x) dW^\mu \quad (8)$$

with $dW^\mu dW^\nu = \rho_{\mu\nu} dt$. The associated metric $g_{\mu\nu}$ depends only on the diffusion terms σ_μ , while the connection $\mathcal{A}_\mu(x)$ involves drift terms b^μ as well:

$$\begin{aligned} g_{\mu\nu}(x) &= 2 \frac{\rho^{\mu\nu}}{\sigma_\mu(x)\sigma_\nu(x)} \\ \rho^{\mu\nu} &\equiv [\rho^{-1}]_{\mu\nu}, \quad \mu, \nu = 1 \cdots n \\ \mathcal{A}^\mu(x) &= \frac{1}{2} \left(b^\mu(x) - g^{-\frac{1}{2}} \partial_\nu (g^{1/2} g^{\mu\nu}(x)) \right) \end{aligned} \quad (9)$$

with

$$g(x) \equiv \det[g_{\mu\nu}(x)] \quad (10)$$

Here, we have used the Einstein convention meaning that two repeated indices are implicitly summed. We set

$$\mathcal{A}_\mu(x) = g_{\mu\nu}(x) \mathcal{A}^\nu(x) \quad (11)$$

The asymptotic solution to the Kolmogorov equation in the short-time limit is given by

$$\begin{aligned} p(t, x|x_0) &= \frac{\sqrt{g(x)}}{(4\pi t)^{\frac{n}{2}}} \sqrt{\Delta(x)} \mathcal{P}(x^0, x) e^{-\frac{d(x)^2}{4t}} \\ &\times \sum_{n=1}^{\infty} a_n(x) t^n \end{aligned} \quad (12)$$

- $d(x)$ is the geodesic distance between x and x^0 measured in the metric $g_{\mu\nu}$. $d(x)$ is defined as

the minimizer of

$$d(x)^2 = \min_{\mathcal{C}} \int_0^1 g_{\mu\nu} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda} d\lambda \quad (13)$$

where λ parameterizes the curve $\mathcal{C}(x^0, x)$ joining $x(\lambda = 0) \equiv x^0$ and $x(\lambda = 1) \equiv x$.

- $\Delta(x)$ is the so-called Van Vleck-Morette determinant:

$$\Delta(x) = g(x)^{-\frac{1}{2}} \det \left(-\frac{\partial^2 d(x)^2}{2\partial x \partial x^0} \right) g(x^0)^{-\frac{1}{2}} \quad (14)$$

- $\mathcal{P}(x^0, x)$ is the parallel transport of the Abelian connection along the geodesic $\mathcal{C}(x^0, x)$ from the point x^0 to x :

$$\mathcal{P}(x^0, x) = e^{-\int_{\mathcal{C}(x^0, x)} \mathcal{A}_\mu(x) dx^\mu} \quad (15)$$

- The $a_i(x)$ coefficients ($a_0(x) = 1$) are smooth functions and depend on geometric invariants such as the scalar curvature. More details can be found in [7].

The Short-time Limit

Plugging the general short-time limit for p at the first-order in time as given by equation (12) in equation (7) and using a saddle-point approximation for the integration over a , we obtain the short-time limit of the effective local volatility function.

Getting implied volatility from the effective local volatility function boils down to calculating the geodesic distance between any two given points in the metric defined by the SVM. While this is generally a nontrivial task, the geodesic distance is known analytically in the special case of the geometry associated with the SVM defined by equations (1) and (6). Details are given in [7].

Asymptotic Implied Volatility

Applying these techniques, we find that the general asymptotic implied volatility at the first order for any

time-homogeneous SVM, depending implicitly on the metric g_{ij} (9) and the connection $\mathcal{A}_i$ (10) is given by

$$\begin{aligned} \sigma_{BS}(\tau, K) = & \frac{\ln \frac{K}{f_0}}{\int_{f_0}^K \frac{df'}{\sqrt{2g^{ff}}}} \left(1 + \frac{g^{ff} \tau}{12} \right. \\ & \left(-\frac{3}{4} \left(\frac{\partial_f g^{ff}}{g^{ff}} \right)^2 + \frac{\partial_f^2 g^{ff}}{g^{ff}} + \frac{1}{f_{av}^2} \right) \\ & + \frac{g^{ff'} \tau}{2g^{ff} \phi''(a_{\min})} \\ & \left. \left(\ln(\Delta g \mathcal{P}^2)'(a_{\min}) - \frac{\phi'''(a_{\min})}{\phi''(a_{\min})} + \frac{g^{ff''}}{g^{ff'}} \right) \right) \end{aligned} \quad (16)$$

with $a_{\min}$ the volatility a , which minimizes the geodesic distance $d(a, f_{av} | \alpha, f_0)$. The g^{ff} are the ff -components of the inverse metric evaluated at $a_{\min}$. Δ is the Van Vleck–Morette determinant as in equation (14), g is the determinant of the metric, and $\mathcal{P}$ is the parallel gauge transport as in equation (15). The prime symbol $'$ indicates a derivative according to a . This formula in equation (16) is particularly useful as we can use it to rapidly calibrate any given SVM. In the following, we apply it to the SABR model with an arbitrary local volatility $C(\cdot)$.

Improved SABR Formula

The asymptotic implied volatility in the SABR model with arbitrary local volatility $C(\cdot)$ is then given by

$$\sigma_{BS}(\tau, K) = \frac{\ln \frac{K}{f_0}}{\sigma(K)} (1 + \sigma_1(f_{av}) \tau) \quad (17)$$

with

$$\begin{aligned} \sigma_1(f) = & \frac{\alpha \nu \rho}{4} \partial_f C(f) \frac{\sinh(d(f))}{d(f)} + \frac{(C(f) a_{\min})^2}{24} \\ & \left(\frac{1}{f^2} + \frac{2 \partial_{ff} (C(f) a_{\min})}{C(f) a_{\min}} \right. \\ & \left. - \left(\frac{\partial_f (C(f) a_{\min})}{C(f) a_{\min}} \right)^2 \right) \end{aligned} \quad (18)$$

with

$$\begin{aligned} \sigma(f) = & \frac{1}{\nu} \\ & \times \ln \left(\frac{q \nu + \alpha \rho + \sqrt{\alpha^2 + q^2 \nu^2 + 2q \alpha \nu \rho}}{\alpha(1 + \rho)} \right) \\ q = & \int_{f_0}^f \frac{dx}{C(x)} \\ a_{\min}(f) = & \sqrt{\alpha^2 + 2\alpha \nu \rho q + \nu^2 q^2} \\ d(f) = & \cosh^{-1} \left(\frac{-q \nu \rho - \alpha \rho^2 + a_{\min}(f)}{\alpha(1 - \rho^2)} \right) \end{aligned} \quad (19)$$

The original SABR formula (6) can be reproduced by approximating $a_{\min}$ for strikes near the money by

$$a_{\min} \simeq \alpha + q \rho \nu$$

and

$$\frac{\sinh(d(a_{\min}))}{d(a_{\min})} \simeq 1 \quad (20)$$

An asymptotic formula for a SABR model with a mean-reversion term, called λ -SABR, has been obtained similarly in [7].

Calibration of the Short-term Smile

Moreover, by inverting equation (17) to lowest order in τ , we see that for any values α , ρ , and ν , a given short-term smile $\sigma_{BS}(f)$ is calibrated by construction if the local volatility function is chosen as

$$\begin{aligned} C(f) = & \frac{f \sigma_{BS}(f) \left(1 - f \ln \left(\frac{f}{f_0} \right) \frac{\sigma'_{BS}(f)}{\sigma_{BS}(f)} \right)^{-1}}{\alpha \sqrt{1 - \rho^2} \cosh \left(\frac{\rho}{|\rho|} \nu \frac{\ln \left(\frac{f}{f_0} \right)}{\sigma_{BS}(f)} + \cosh^{-1} \left(\frac{1}{\sqrt{1 - \rho^2}} \right) \right)} \end{aligned} \quad (21)$$

References

- [1] Avellaneda, M., Boyer-Olson, D., Busca, J. & Friz, P. (2002). Reconstructing the smile, *Risk Magazine* October 91–95.

- [2] Berestycki, H., Busca, J. & Florent, I. (2004). Computing the implied volatility in stochastic volatility models, *Communications on Pure and Applied Mathematics* **57**(10), 1352–1373.
- [3] Dupire, B. (2004). A unified theory of volatility, in *Derivatives Pricing: The Classic Collection*, P. Carr, ed., Risk Publications.
- [4] Hagan, P., Kumar, D., Lesniewski, A.S. & Woodward, D.E. (2002). Managing smile risk, *Wilmott Magazine* September, 84–108.
- [5] Henry-Labordère, P. (2007). Combining the SABR and BGM models, *Risk Magazine* October 102–107.
- [6] Henry-Labordère, P. (2008). A geometric approach to the asymptotics of implied volatility, in R. Cont, ed., *Frontiers in Quantitative Finance: Volatility and Credit Risk Modeling*, Wiley, Chapter 4.
- [7] Henry-Labordère, P. (2008). *Analysis, Geometry and Modeling in Finance: Advanced Methods in Option Pricing*, Financial Mathematics Series, Chapman & Hall/CRC 102–104.
- [8] Obłój, J. (2008). Fine-tune your smile, *Wilmott Magazine* May.

Further Reading

- Benaim, S., Friz, P., Lee, R. (2008). On the Black-Scholes implied volatility at extreme strikes, in *Frontiers in Quantitative Finance: Volatility and Credit Risk Modeling*, R. Cont, ed., Wiley, Chapter 3.
- Lee, R. (2004). The moment formula for implied volatility at extreme strikes, *Mathematical Finance* **14**(3), July 469–480.

PIERRE HENRY-LABORDÈRE

Implied Volatility: Large Strike Asymptotics

Let S_t be the price of a risky asset at time $t \in [0, T]$ and $B_t = B(t, T)$ the time- t value of one monetary unit received at time T . Assuming suitable no-arbitrage conditions, there exists a probability measure $\mathbb{P} = \mathbb{P}_T$, called the *(T -forward) pricing measure*, under which the B_t -discounted asset price

$$F_t = F(t, T) = S(t) / B(t, T) \quad (1)$$

is a martingale and so are B_t -discounted time- t option prices, such as $C_t / B(t, T)$, where C_t denotes the time- t value of a European call option with maturity T and payoff $(S_T - K)^+$. With focus on $t = 0$ and writing C instead of C_0 , we have

$$C = B(0, T) \mathbb{E} [C_T / B(T, T)] \quad (2)$$

$$\begin{aligned} &= B(0, T) \mathbb{E} [(F_T - K)^+] \\ &= S_0 \mathbb{E} \left[\left(\frac{F_T}{F_0} - \frac{K}{F_0} \right)^+ \right] \end{aligned} \quad (3)$$

Let us remark that in the case of deterministic interest rates $r(\cdot)$, one can rewrite this as^a

$$C = \mathbb{E} \left[\exp \left(- \int_0^T r(u) du \right) (S_T - K)^+ \right] \quad (4)$$

If we now make the assumption that there exists $\sigma > 0$, the *Black–Scholes volatility*, such that F_t satisfies

$$dF_t = \sigma F_t dW \quad (5)$$

where W is a Brownian motion under $\mathbb{P}$, then we have *normal returns* $X_{BS} := \log(F_T / F_0)$. More precisely,

$$X_{BS} \sim \text{Normal}(-V^2/2, V^2) \text{ with } V \equiv \sigma\sqrt{T} \quad (6)$$

and an elementary integration of equation (3) yields the classical Black–Scholes formula,^b

$$C_{BS} = S_0 \left\{ \Phi(d_1) - \frac{K}{F_0} \Phi(d_2) \right\} \quad (7)$$

with $d_{1,2} = -\log(K/F_0)/V \pm V/2$. It follows that we can express the *normalized Black–Scholes call price*

$$c_{BS} := C_{BS} / S_0 \quad (8)$$

as a function of two variables: *log-strike* $k := \log(K/F_0)$ and (scaled) *Black–Scholes volatility* $V = \sigma\sqrt{T}$,

$$\begin{aligned} c_{BS}(k, V) &= \Phi(d_1) - e^k \Phi(d_2) \\ \text{with } d_{1,2} &= -k/V \pm V/2 \end{aligned} \quad (9)$$

Let us now return to the general setting and just assume that, for fixed T , the *returns*

$$X := \log \frac{F_T}{F_0} \quad (10)$$

have a known distribution, fully specified by the probability distribution function

$$F(x) := \mathbb{P}[X \leq x] \quad (11)$$

From equation (2), the value of a normalized call price $c = C/S_0$ is then given by

$$\begin{aligned} \mathbb{E} \left[\left(\frac{F_T}{F_0} - e^k \right)^+ \right] &= \int_{-\infty}^{\infty} (e^x - e^k)^+ dF(x) \\ &=: c(k) \end{aligned} \quad (12)$$

Definition 1 (Implied volatility). *Let $T > 0$ be a fixed maturity and assume F is the distribution function of the returns $\log(F_T/F_0)$ under the pricing measure. Then, the scaled (Black–Scholes) implied volatility is the unique value $V(k)$ such that*

$$c(k) = c_{BS}(k, V(k)) \text{ for all } k \in \mathbb{R} \quad (13)$$

We also write $\sigma(k, T) := V(k)/\sqrt{T}$ for the (annualized, Black–Scholes) implied volatility.

By the very definition, the volatility smile $V(\cdot, T)$ is flat, namely constant equal to $V = \sigma\sqrt{T}$, in the Black–Scholes model. To see existence/uniqueness of implied volatility, in general, it suffices to note that $c_{BS}(k, \cdot)$ is strictly increasing in the volatility parameter and that

$$\begin{aligned} c_{BS}(k, V = 0) &= (1 - e^k)^+ \leq \mathbb{E} \left[(F_T/F_0 - e^k)^+ \right] = c(k) \\ c(k) &\leq \mathbb{E} [F_T/F_0] = 1 = c_{BS}(k, V = +\infty) \end{aligned} \quad (14)$$

$$c(k) \leq \mathbb{E} [F_T/F_0] = 1 = c_{BS}(k, V = +\infty) \quad (15)$$

It is clear from the afore mentioned monotonicity of $c_{BS}(k, \cdot)$ that the fatness of the tail of the returns, for example, the behavior of

$$\bar{F}(k) = 1 - F(k) = \mathbb{P}[X > k] \text{ as } k \rightarrow \infty \quad (16)$$

is related to the shape of the “wing” of the implied volatility (smile) for far-out-of-the-money calls, $V(k)$ as $k \rightarrow \infty$, and similarly, for $F(k)$, $V(k)$ as $k \rightarrow -\infty$. Surprisingly, perhaps, this link can be made very explicit. Let us agree that if F admits a density, it is denoted by $f = F'$. Let us also adopt the common convention that

$$g(k) \sim h(k) \text{ means } g(k)/h(k) \rightarrow 1 \text{ as } k \rightarrow \infty \quad (17)$$

The (meta) result is the following *tail-wing formula*: as $k \rightarrow \infty$ we have

$$\begin{aligned} V(k)^2/k &\sim \psi[-1 - \log \bar{F}(k)/k] \\ &\sim \psi[-1 - \log f(k)/k] \end{aligned} \quad (18)$$

$$\begin{aligned} V(-k)^2/k &\sim \psi[-\log F(-k)/k] \\ &\sim \psi[-\log f(-k)/k] \end{aligned} \quad (19)$$

where

$$\psi(x) \equiv 2 - 4 \left(\sqrt{x^2 + x} - x \right) \quad (20)$$

An interesting special case arises when either

$$p^* = \sup \{ p \in \mathbb{R} : M(1+p) < \infty \} \quad (21)$$

$$q^* = \sup \{ q \in \mathbb{R} : M(-q) < \infty \} \quad (22)$$

is finite, where M is the *moment generating function* of F , and this is equivalent to *moment explosion* of the underlying since

$$M(u) := \int e^{ux} dF(x) = \frac{1}{F_0^u} \mathbb{E}[F_T^u] = \frac{1}{F_0^u} \mathbb{E}[S_T^u] \quad (23)$$

In this case, one expects an exponential tail so that

$$-\log \bar{F}(k) \sim (p^* + 1)k \quad (24)$$

$$-\log F(-k) \sim q^*k \quad (25)$$

and one is led to *Lee's moment formula*

$$V(k)^2/k \sim \psi(p^*)$$

$$V(-k)^2/k \sim \psi(q^*) \quad (26)$$

Recall that $g(k) \sim h(k)$ stands for the precise mathematical statement that $\lim g(k)/h(k) \rightarrow 1$ as $k \rightarrow \infty$. In the same spirit, let us agree that

$$g(k) \lesssim h(k) \text{ means } \limsup g(k)/h(k) \rightarrow 1 \text{ as } k \rightarrow \infty \quad (27)$$

Proposition 1 (Lee's Moment Formula; [3, 8]). *Assume $\mathbb{E}[e^X] = \mathbb{E}[S_T]/F_0 < \infty$. The moment formula then holds in complete generality in “limsup” form. More precisely, as $k \rightarrow \infty$,*

$$V(k)^2/k \lesssim \psi(p^*) \quad (28)$$

$$V(-k)^2/k \lesssim \psi(q^*) \quad (29)$$

The power of the moment formula comes from the fact that the critical exponents p^*, q^* can often be obtained by sheer inspection of a moment generating function known in closed form. One can also make use of the recent literature on moment explosions to obtain such critical exponents in various stochastic volatility models; *see Moment Explosions* and the references therein. Let us note that it is possible [4] to construct (pathological) examples to see that one cannot hope for a genuine limit form of the above moment formula, as was suggested in (26). Another remark is that the moment formula provides little information in the absence of moment explosion. For instance, $p^* = +\infty$ only implies $V(k)^2 = o(k)$ but gives no further information about the behavior of $V(k)$ for large k . Both the issues are dealt with by the tail-wing formula. The key assumptions is a certain well behavedness of F ; but only on a crude logarithmic scale and, therefore, rather easy to check in many examples.

Definition 2 (Regular Variation; [5]). *A positive, real-valued function f , defined at least on $[M, \infty)$ for some large M , is said to be regularly varying of index α if for all $\lambda > 0$*

$$f(\lambda k) \sim \lambda^\alpha f(k) \text{ as } k \rightarrow \infty \quad \forall \lambda > 0 \quad (30)$$

and in this case we write $f \in R_\alpha$.

Theorem 1 (Right-hand Tail-wing Formula; [2]). Assume $\exists \epsilon > 0 : \mathbb{E}[e^{(1+\epsilon)X}] = \mathbb{E}[S_T^{1+\epsilon}] / F_0^{1+\epsilon} < \infty$. Let also $\alpha > 0$ and set

$$\psi(x) \equiv 2 - 4 \left(\sqrt{x^2 + x} - x \right) \quad (31)$$

Then (i) $\implies$ (ii) $\implies$ (iii) $\implies$ (iv) where

- (i) $-\log f(k) \in R_\alpha;$
- (ii) $-\log \bar{F}(k) \in R_\alpha;$
- (iii) $-\log c(k) \in R_\alpha;$

and

$$(iv) \quad V(k)^2/k \sim \psi[-\log c(k)/k].$$

If (ii) holds, then $-\log c(k) \sim -k - \log \bar{F}$ and

$$(iv') \quad V(k)^2/k \sim \psi[-1 - \log \bar{F}(k)/k],$$

if (i) holds, then $-\log f \sim -\log \bar{F}$ and

$$(iv'') \quad V(k)^2/k \sim \psi[-1 - \log f(k)/k].$$

Of course, there is a similar left-hand result, which we state such as to involve far-out-of-the-money (normalized) European puts,

$$p(k) := \int_{-\infty}^{\infty} (e^k - e^x)^+ dF(x) \quad (32)$$

Theorem 2 (Left-hand Tail-wing Formula). Assume $\exists \epsilon > 0 : \mathbb{E}[e^{-\epsilon X}] < \infty$. Then (i) $\implies$ (ii) $\implies$ (iii) $\implies$ (iv) where

- (i) $-\log f(-k) \in R_\alpha;$
- (ii) $-\log F(-k) \in R_\alpha;$
- (iii) $-\log p(-k) \in R_\alpha;$

and

$$(iv) \quad V(-k)^2/k \sim \psi[-1 - \log p(-k)/k].$$

If (ii) holds, then $-\log p(-k) \sim k - \log F(-k)$

and

$$(iv') \quad V(-k)^2/k \sim \psi[-\log F(-k)/k],$$

if (i) holds, then $-\log f(-k) \sim -\log F(-k)$

and

$$(iv'') \quad V(-k)^2/k \sim \psi[-\log f(-k)/k].$$

With focus on the right-hand tail-wing, let us single out two cases of particular importance in applications.

1. **(Asymptotically Linear Regime)** If $-1 - \log f(k)/k$ or $-1 - \log \bar{F}(k)/k$ converges to $p^* \in (0, \infty)$ then

$$V(k)^2 \sim \psi(p^*) \times k \quad (33)$$

and the *implied variance*, defined as the square of implied volatility, is asymptotically linear with slope $\psi(p^*)$. One can, in fact, check this from the moment generating function of X . Indeed, it is shown in [3] that if

$$s \mapsto M(1 + p^* - 1/s) \quad (34)$$

is regularly varying then $-\log f(k)/k \rightarrow p^* + 1$. In other words, to ensure equation (26), that is, a genuine limit in Lee's moment formula, one needs some well behavedness of the M as its argument approaches the critical exponent $1 + p^*$. Similar conditions can be given with M replaced by M' or $\log M$ and these conditions are, indeed, easy to check in a number of familiar exponential Lévy models (including Barndorff-Nielsen's Normal Inverse Gaussian model, Carr-Madan's Variance Gamma model, or Kou's Double Exponential model) and various time changes of these models (see [3] for details).

2. **(Asymptotically Sublinear Regime)** If $-\log f(k)/k \rightarrow \infty$, we can use $\psi(x) \sim 1/(2x)$ as $x \rightarrow \infty$ to see that

$$V(k)^2 \sim \frac{1}{-2 \log f(k)/k} \times k = \frac{k^2}{-2 \log f(k)} \quad (35)$$

so that the implied variance is asymptotically sublinear. As sanity check, consider the Black-Scholes model where f is the density of the (normally distributed) returns with variance $V^2 \equiv \sigma^2 T$, as given in (6); then $-\log f(k) \sim k^2/(2V^2)$ and it follows that $V(k) \sim V$, in trivial agreement with the flat smile in the Black-Scholes model. Following [2], other examples are given by Merton's jump diffusion as a borderline example

in which the sublinear behavior comes from a subtle logarithmic correction term, and *Carr–Wu’s Finite Moment Logstable model*. The tail behavior of the latter, as noted in [2], can be derived from the growth of the (nonexplosive) moment generating function by means of *Kasahara’s Tauberian theorem* [5]. Another example where this methodology works is the SABR model

$$dF = \sigma F^\beta dW, \quad d\sigma = \eta \sigma dZ \quad (36)$$

with $\sigma, \eta > 0, \beta < 1$ and two Brownian motions W, Z assumed (here) to be independent. Using standard stochastic calculus [4], one can give good enough estimates on $\mathbb{E}[|F_T/F_0|^u]$, from above and below, to see that

$$\begin{aligned} \log \mathbb{E}[|F_T/F_0|^u] &= \log \mathbb{E}[\exp(uX)] \\ &\sim \frac{\eta^2 T}{(1-\beta)^2} \frac{u^2}{2} \text{ as } u \rightarrow \infty \end{aligned} \quad (37)$$

From this, Kasahara’s theorem allows to deduce the tail behavior of X , namely

$$-\log \mathbb{P}[X > x] \sim \frac{(1-\beta)^2}{\eta^2 T} \frac{x^2}{2} \quad (38)$$

and the (right hand) tail-wing formula reveals that the implied volatility in the SABR model is asymptotically flat, $\sigma(k, T) \sim \eta/(1-\beta)$ as $k \rightarrow \infty$.

Early contributions in the study of smile asymptotics are [1, 6]. The moment formula appears in [8], the tail-wing formula in [2] with some additional criteria in [3]. A survey on the topic, together with some new examples (including CEV and SABR) is found in [4]. Further developments in the field include the refined asymptotic results of Gulisashvili and Stein [7]; in a simple log-normal stochastic volatility model of the form $dF = \sigma F dW, \quad d\sigma = \eta \sigma dZ$, with two independent Brownian motions W, Z they find

$$\sigma(k, T) \sqrt{T} = \sqrt{2k} - \frac{\log k + \log \log k}{2\eta \sqrt{T}} + O(1) \quad (39)$$

The leading order term $\sqrt{2k}$ says that implied variance grows linearly with slope 2, as one expects in a model with immediate moment explosion.

Acknowledgments

Financial support from the Cambridge Endowment of Research in Finance is gratefully acknowledged.

End Notes

^aEquation (4) is valid in a nondeterministic interest rate setting, provided the expectation is taken with respect to the *risk-neutral measure* (which is equivalent but, in general, not identical to $\mathbb{P}_T$).

^b Φ denotes the distribution function of normal (0, 1).

References

- [1] Avellaneda, M. & Zhu, Y. (1998). A risk-neutral stochastic volatility model, *International Journal of Theoretical and Applied Finance* **1**(2), 289–310.
- [2] Benaïm, S. & Friz, P.K. (2009). Regular variation and smile asymptotics, *Mathematical Finance* **19**(1), 1–12, eprint arXiv:math/0603146.
- [3] Benaïm, S. Friz, P.K. (2008). Smile asymptotics II: models with known MGF, *Journal of Applied Probability* **45**(1), 16–32.
- [4] Benaïm, S. Friz, P.K. & Lee, R. (2008). The Black–Scholes implied volatility at extreme strikes, in *frontiers, in Quantitative Finance: Volatility and Credit Risk Modeling*, Chapter 2, Wiley.
- [5] Bingham, N.H. Goldie, C.M. & Teugels, J.L. (1987). *Regular Variation*, CUP.
- [6] Gatheral, J. (2000). *Rational shapes of the Volatility Surface*, Presentation, RISK Conference.
- [7] Gulisashvili, A. & Stein, E. Implied volatility in the Hull–White model, *Mathematical Finance*, to appear.
- [8] Lee, R. (2004). The moment formula for implied volatility at extreme strikes, *Mathematical Finance* **14**(3), 469–480.

Further Reading

Gatheral, J. (2006). *The Volatility Surface, A Practitioner’s Guide*, Wiley.

PETER K. FRIZ

Constant Elasticity of Variance (CEV) Diffusion Model

The CEV Process

The constant elasticity of variance (CEV) model is a one-dimensional diffusion process that solves a stochastic differential equation (SDE)

$$dS_t = \mu S_t dt + a S_t^{\beta+1} dB_t \quad (1)$$

with the instantaneous volatility $\sigma(S) = aS^\beta$ specified to be a power function of the underlying spot price. The model has been introduced by Cox [7] as one of the early alternative processes to the geometric Brownian motion to model asset prices. Here β is the elasticity parameter of the local volatility, $d\sigma/dS = \beta\sigma/S$, and a is the volatility scale parameter. For $\beta = 0$, the CEV model reduces to the constant volatility geometric Brownian motion process employed in the Black, Scholes, and Merton model. When $\beta = -1$, the volatility specification is that of Bachelier (the asset price has the constant diffusion coefficient, while the logarithm of the asset price has the a/S volatility). For $\beta = -1/2$ the model reduces to the square-root model of Cox and Ross [8].

Cox [7] originally studied the case $\beta < 0$ for which the volatility is a decreasing function of the asset price. This specification captures the leverage effect in the equity markets: the stock price volatility increases as the stock price declines. The result of this inverse relationship between the price and volatility is the implied volatility skew exhibited by options prices in the CEV model with negative elasticity. The elasticity parameter β controls the steepness of the skew (the larger the $|\beta|$, the steeper the skew), while the scale parameter a fixes the at-the-money volatility level. This ability to capture the skew has made the CEV model popular in equity options markets.

Emanuel and MacBeth [14] extended Cox's analysis to the positive elasticity case $\beta > 0$, where the asset price volatility is an increasing function of the asset price. The driftless process with $\mu = 0$ and with positive β is a strict local martingale. It has been applied to modeling commodity prices that exhibit increasing implied volatility skews with the volatility

increasing with the strike price, but care should be taken when working with this model (see the discussion below).

The CEV diffusion has the following boundary characterization (see, e.g., [4] for Feller's boundary classification for one-dimensional diffusions). For $-1/2 \leq \beta < 0$, the origin is an exit boundary, and the process is killed the first time it hits the origin. For $\beta < -1/2$, the origin is a regular boundary point. The SDE (1) does not uniquely specify the diffusion process, and a boundary condition is needed at the origin. In the CEV model, it is specified as a killing boundary. Thus, the CEV process with $\beta < 0$ naturally incorporates the possibility of bankruptcy—the stock price can hit zero with positive probability, at which time the bankruptcy occurs. For $\beta \geq 0$, the origin is an inaccessible natural boundary.

Reduction to Bessel Processes, Transition Density, and Probability of Default

The CEV process is analytically tractable. Its transition probability density and cumulative distribution function are known in closed form.^a It is closely related to Bessel processes and inherits their analytical tractability. The CEV process with drift ($\mu \neq 0$) is obtained from the process without drift ($\mu = 0$) via a scale and time change:

$$S_t^{(\mu)} = e^{\mu t} S_{\tau(t)}^{(0)}, \quad \tau(t) = \frac{(e^{2\mu\beta t} - 1)}{2\mu\beta} \quad (2)$$

Let $\{R_t^{(\nu)}, t \geq 0\}$ be a Bessel process of index ν . Recall that for $\nu \geq 0$, zero is an unattainable entrance boundary. For $\nu \leq -1$, zero is an exit boundary. For $\nu \in (-1, 0)$, zero is a regular boundary. In our application, we specify zero as a killing boundary to kill the process at the first hitting time of zero (see, e.g., [4, pp. 133–134], for a summary of Bessel processes). Before the first hitting time of zero, the CEV process without drift can be represented as a power of a Bessel process:

$$S_t^{(0)} = (a|\beta|R_t^{(\nu)})^{-\frac{1}{\beta}} \quad (3)$$

where $\nu = 1/(2\beta)$.

The CEV transition density is obtained from the well-known expression for the transition density of

the Bessel process (see [4, p. 115, 21, p. 446]). For the driftless process, it is given by

$$p_{(0)}(S_0, S_t; t) = \frac{S_t^{-2\beta-3/2} S_0^{1/2}}{a^2 |\beta| t} I_{|\nu|} \left(\frac{S_0^{-\beta} S_t^{-\beta}}{a^2 \beta^2 t} \right) \times \exp \left(-\frac{S_0^{-2\beta} + S_t^{-2\beta}}{2a^2 \beta^2 t} \right) \quad (4)$$

where I_ν is the modified Bessel function of the first kind of order ν . From equation (2), the transition density with drift is obtained from the density equation (4) according to

$$p_{(\mu)}(S_0, S_t; t) = e^{-\mu t} p_{(0)}(S_0, e^{-\mu t} S_t; \tau(t)) \quad (5)$$

The density (5) was originally obtained by Cox [7] for $\beta < 0$ and by Emanuel and MacBeth [14] for $\beta > 0$ on the basis of the result due to Feller [15].

For $\beta < 0$, in addition to the continuous transition density, we also have a positive probability for the process started at S_0 at time zero to hit zero by time $t \geq 0$ (probability of default or bankruptcy) that is given explicitly by

$$G \left(|\nu|, \frac{\mu S_0^{-2\beta}}{a^2 \beta (e^{2\mu\beta t} - 1)} \right) \quad (6)$$

where $G(\nu, x) = (1/\Gamma(\nu)) \int_x^\infty u^{\nu-1} e^{-u} du$ is the complementary Gamma distribution function. This expression can be obtained by integrating the continuous density (5) from zero to infinity and observing that the result is less than one, that is, the density is defective. The defect is equal to the probability mass at zero equation (6).

While killing the process at zero is desirable for stock price modeling, it may be undesirable in other contexts, where one would prefer the process that stays strictly positive (e.g., in stock index models). A regularized version of the CEV process that never hits zero has been constructed by Andersen and Andreasen [1] (see also [9]). The positive probability of hitting zero comes from the explosion of instantaneous volatility as the process falls toward zero. The regularized version of the CEV process fixes a small value $\epsilon > 0$. For $S > \epsilon$, the volatility is according to the CEV specification. For $S \leq \epsilon$, the volatility is fixed at the constant level $a\epsilon^\beta$. We thus have a sequence of regularized strictly positive processes

indexed by ϵ that converge to the CEV process in the limit $\epsilon \rightarrow 0$.

The CEV process with $\beta > 0$ can similarly be regularized to prevent the volatility explosion as the process tends to infinity by picking a large value $\mathcal{E} > 0$ and fixing the volatility above $\mathcal{E}$ to equal $a\mathcal{E}^\beta$. The regularized processes with $\mu = 0$ are true martingales, as opposed to the failure of the martingale property for the driftless CEV process with $\beta > 0$ and $\mu = 0$, which is only a strict local martingale. The failure of the martingale property for the nonregularized process with $\beta > 0$ can be explicitly illustrated by computing the expectation (using the transition density (5)):

$$\mathbb{E}[S_t] = e^{\mu t} S_0 \left(1 - G \left(\nu, \frac{\mu S_0^{-2\beta}}{a^2 \beta (e^{2\mu\beta t} - 1)} \right) \right) \quad (7)$$

CEV Options Pricing

The closed-form CEV call option pricing formula with strike K , time to expiration T , and the initial asset price S can be obtained in closed form by integrating the call payoff with the risk-neutral CEV density (5) with the risk-neutral drift $\mu = r - q$ (r is the risk-free interest rate and q is the dividend yield). The result can be expressed in terms of the complementary noncentral chi-square distribution function $Q(z; \nu, k)$ ([7] for $\beta < 0$, [14] for $\beta > 0$; see also [11, 22]):

$$C(S; K, T) = e^{-rT} \mathbb{E}[(S_T - K)^+] = \begin{cases} e^{-qT} S Q(\xi; 2\nu, y_0) - e^{-rT} K (1 - Q(y_0; 2(1+\nu), \xi)), & \beta > 0 \\ e^{-qT} S Q(y_0; 2(1+|\nu|), \xi) - e^{-rT} K (1 - Q(\xi; 2|\nu|, y_0)), & \beta < 0 \end{cases} \quad (8)$$

where

$$\xi = \frac{2\mu S^{-2\beta}}{a^2 \beta (e^{2\mu\beta T} - 1)}, \quad y_0 = \frac{2\mu K^{-2\beta}}{a^2 \beta (1 - e^{-2\mu\beta T})} \quad (9)$$

and $S = S_0$ is the initial asset price at time zero. The price of the put option is obtained from the put-call parity relationship:

$$P(S; K, T) = C(S; K, T) + K e^{-rT} - S e^{-qT} \quad (10)$$

The complementary noncentral chi-square distribution function can be expressed as the series of complementary Gamma distribution functions ([22, pp. 214]):

$$Q(z; v, k) = \sum_{n=0}^{\infty} e^{-k/2} \frac{(k/2)^n}{\Gamma(n+1)} G\left(n + \frac{v}{2}, \frac{z}{2}\right) \quad (11)$$

for $k, z > 0$. Further efficient numerical methods to compute the noncentral chi-square cumulative distribution function (CDF) can be found in [3, 12, 13, 22].

The first passage time problem for the CEV diffusion can be solved analytically and, hence, barrier and lookback options can be priced analytically under the CEV process. Davydov and Linetsky [9, 10] obtained the analytical expressions for the Laplace transforms of single- and double-barrier and lookback options pricing formulas in time to expiration. Davydov and Linetsky [10] and Linetsky [18] inverted the Laplace transforms for barrier options and lookback options in terms of eigenfunction expansions, respectively.

Other types of options under the CEV process, such as American options, require numerical treatment. The pricing partial differential equation (PDE) for European options reads as follows:

$$\frac{a^2}{2} S^{2\beta+2} \frac{\partial^2 V}{\partial S^2} + (r - q) S \frac{\partial V}{\partial S} + \frac{\partial V}{\partial t} = r V \quad (12)$$

The early exercise can be dealt with in the same way as for other diffusion models *via* dynamic programming, free boundary PDE formulations, or variational inequality formulations.

Jump-to-Default Extended CEV Model

While the CEV process can hit zero and, as a result, the CEV equity model includes the positive probability of bankruptcy, the term structure of credit spreads in the CEV model is such that the instantaneous credit spread vanishes. There is no element of surprise—the event of default is a hitting time. Moreover, the probability of default is too small for practical applications of modeling stocks of firms other than the highest rated investment grades. Carr and Linetsky [6] extend the CEV model by allowing a jump to default to occur from a positive stock price. They introduce a default

intensity that is an affine function of the instantaneous variance:

$$\lambda(S) = b + c\sigma^2(S) = b + ca^2 S^{2\beta} \quad (13)$$

where $b \geq 0$ is the constant part of the default intensity and $c \geq 0$ is the sensitivity of the default intensity to the instantaneous variance. The predefault stock price follows a diffusion process solving the SDE:

$$dS_t = [\mu + \lambda(S_t)] S_t dt + a S_t^{\beta+1} dB_t \quad (14)$$

The addition of the default intensity in the drift compensates for the jump to default and makes the process with $\mu = 0$ a martingale. The diffusion process with the modified drift (14) and killed at the rate (13) is called *jump-to-default extended constant elasticity of variance (JDCEV) process*. In the JDCEV model, the stock price evolves according to equation (14) until a jump to default arrives, at which time the stock price drops to zero and equity becomes worthless. The jump to default time has the intensity (13).

The JDCEV model can be reduced to Bessel processes similar to the standard CEV model. Consequently, it is also analytically tractable. Closed-form pricing formulas for call and put options and the probability of default can be found in [6]. The first passage time problem for the JDCEV process and the related problem of pricing equity default swaps are solved in [20]. Atlan and Leblanc [2] and Campi *et al.* [5] investigate related applications of the CEV model to hybrid credit–equity modeling.

Volatility Skews and Credit Spreads

Figure 1(a) illustrates the shapes of the term structure of zero-coupon credit spreads in the CEV and JDCEV models, assuming zero recovery. The credit spread curves start at the instantaneous credit spread equal to the default intensity $b + c\sigma_*^2$ (σ_* is the volatility at a reference level S^*).^b The instantaneous credit spreads for the CEV model vanish, while they are positive for the JDCEV model. Figure 1(b) plots the Black–Scholes implied volatility against the strike price in the CEV and JDCEV models (we calculate the implied volatility by equating the price of an option under the Black–Scholes model to the corresponding option price under the (JD)CEV model). One can observe the decreasing and convex implied

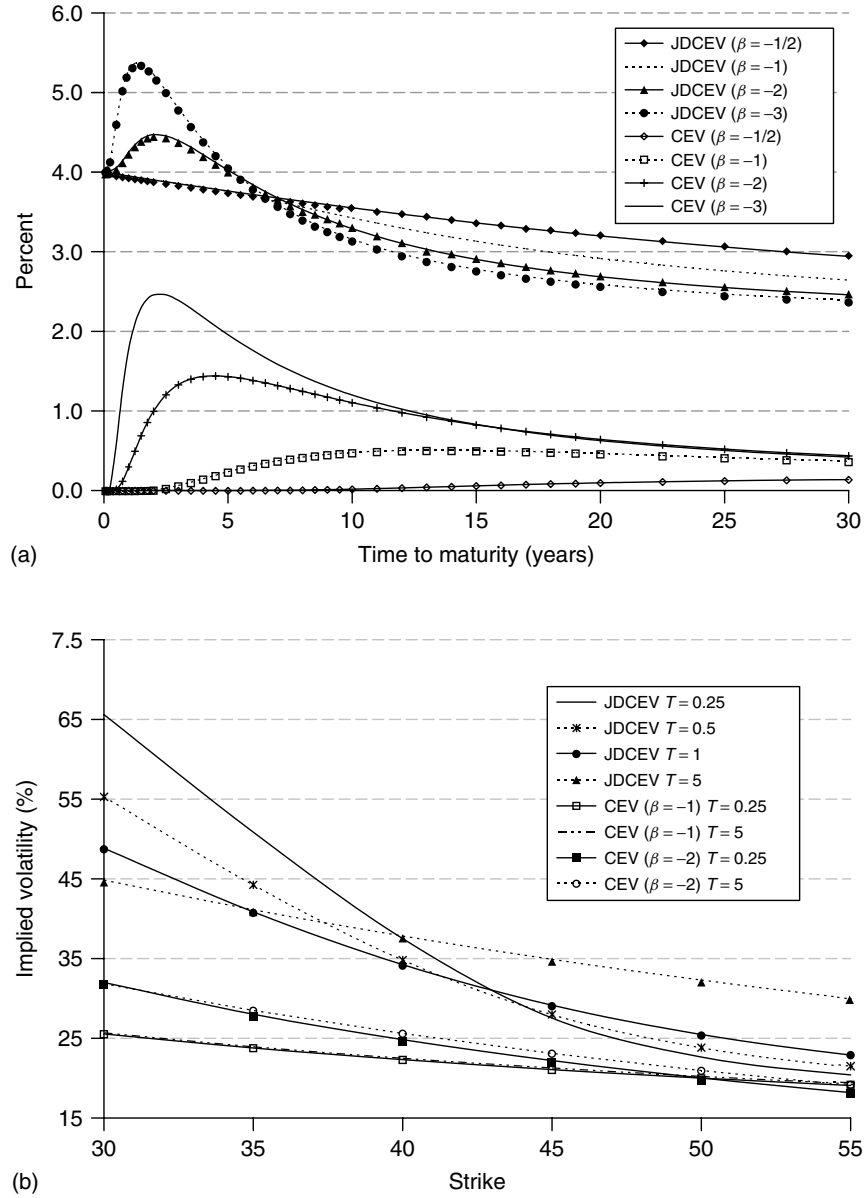


Figure 1 (a) Term structures of credit spreads. Parameter values: $S = S^* = 50$, $\sigma_* = 0.2$, $\beta = -1/2, -1, -2, -3$, $r = 0.05$, $q = 0$. JDCEV: $b = 0.02$ and $c = 1/2$. CEV: $b = 0$ and $c = 0$. (b) Implied volatility skews. Parameter values: $S = S^* = 50$, $\sigma_* = 0.2$, $r = 0.05$, $q = 0$. For JDCEV model: $b = 0.02$, $c = 1/2$ and $\beta = -1$, the times to expiration are $T = 0.25, 0.5, 1, 5$ years. For CEV model: $b = c = 0$, $\beta = -1, -2$ and times to expiration are $T = 0.25, 5$. Implied volatilities are plotted against the strike price

volatility skew with implied volatilities increasing for lower strikes, as the local volatility and the default intensity both increase as the stock price declines. The volatility elasticity β controls the slope of the

skew in the CEV model. The slope of the skew in the JDCEV model is steeper and is controlled by β , as well as the default intensity parameters b and c .

Implied Volatility and the SABR model

By using singular perturbation techniques, Hagan and Woodward [17] obtained explicit asymptotic formulas for the Black–Scholes implied volatility σ_{BS} of European calls and puts on an asset whose forward price $F(t)$ follows the CEV dynamics, that is, $dF_t = aF_t^{\beta+1} dB_t$,

$$\sigma_{BS} = af_{av}^{\beta} \left\{ 1 - \frac{\beta(\beta+3)}{24} \left(\frac{F_0 - K}{f_{av}} \right)^2 + \frac{\beta^2}{24} a^2 \tau f_{av}^{2\beta} + \dots \right\} \quad (15)$$

where τ is time to expiration, $f_{av} = (F_0 + K)/2$ and F_0 is today's forward price (Hagan and Woodward's β is equal to our $\beta + 1$). This asymptotics for the implied volatility approximates the exact CEV-implied volatilities well when the ratio F_0/K is not too far from one and when K and F_0 are far away from zero. The accuracy tends to deteriorate when the values are close to zero since this asymptotic approximation does not take into account the killing boundary condition at zero.

Hagan *et al.* [16] introduced the SABR model, which is a CEV model with stochastic volatility. More precisely, the volatility scale parameter a is made stochastic, so that the forward asset price follows the dynamics:

$$\begin{aligned} dF_t &= a_t F_t^{\beta+1} dB_t^{(1)} \quad \text{and} \\ da_t &= \eta a_t dB_t^{(2)} \end{aligned} \quad (16)$$

where $dB_t^{(1)}, dB_t^{(2)} = \rho dt$. Hagan *et al.* derive the asymptotic expression for the implied volatility in the SABR model.

Introducing Jumps and Stochastic Volatility into the CEV Model

Mendoza *et al.* [19] introduce jumps and stochastic volatility into the JDCEV model by time changing the JDCEV process. Lévy subordinator time changes introduce state-dependent jumps into the process, while absolutely continuous time changes introduce stochastic volatility. The result is a flexible family of models that exhibit the leverage effect, default

intensity linked to the stock price volatility, jumps, and stochastic volatility. These models inherit the analytical tractability of the CEV and JDCEV models as long as the Laplace transform of the time-change process is analytically tractable. The stochastic volatility version of the CEV model obtained in this approach is different from the SABR model in two respects. The advantage of the time-change approach is that it preserves the analytical tractability for more realistic choices for the stochastic volatility process, such as the Cox–Ingersoll–Rand (CIR) process with mean-reversion. Another advantage is that jumps, including the jump to default, can also be incorporated. The weakness is that it is hard to incorporate the correlation between the price and volatility.

End Notes

^aIn this article we present the results for the CEV model with constant parameters. We note that the process remains analytically tractable when μ and a are taken to be deterministic functions of time [6].

^bIt is convenient to parameterize the local volatility function as $\sigma(S) = aS^{\beta} = \sigma_*(S/S^*)^{\beta}$ so that at some reference spot price level $S = S^*$ (e.g., the at-the-money level at the time of model calibration) the volatility takes the reference value, $\sigma(S^*) = \sigma_*$. In the example presented here, the reference level is taken to equal the initial spot price level, $S^* = S_0$, and the volatility scale parameter is $a = \sigma_*/(S_0^{\beta})$.

Acknowledgments

This research was supported by the National Science Foundation under grant DMS-0802720.

References

- [1] Andersen, L. & Andreasen, J. (2000). Volatility skew and extensions of the LIBOR market model, *Applied Mathematical Finance* **7**, 1–32.
- [2] Atlan, M. & Leblanc, B. (2005). Hybrid equity-credit modelling, *Risk Magazine* **18**, 8.
- [3] Benton, D. & Krishnamoorthy, K. (2003). Computing discrete mixtures of continuous distributions: noncentral chi-square, noncentral t and the distribution of the square of the sample multiple correlation coefficient, *Computational Statistics and Data Analysis* **43**, 249–267.

- [4] Borodin, A. & Salminen, P. (2002). *Handbook of Brownian Motion: Facts and Formulae*, Probability and Its Applications, 2nd rev Edition, Birkhauser Verlag AG.
- [5] Campi, L., Sbuelz, A. & Polbennikov, S. (2008). Systematic equity-based credit risk: A CEV model with jump to default, *Journal of Economic Dynamics and Control* **33**, 93–108.
- [6] Carr, P. & Linetsky, V. (2006). A jump to default extended CEV model: an application of Bessel processes, *Finance and Stochastics* **10**, 303–330.
- [7] Cox, J.C. (1975, 1996). Notes on option pricing I: constant elasticity of variance diffusions, *Reprinted in The Journal of Portfolio Management* **23**, 15–17.
- [8] Cox, J.C. & Ross, S.A. (1976). The valuation of options for alternative stochastic processes, *Journal of Financial Economics* **3**, 145–166.
- [9] Davydov, D. & Linetsky, V. (2001). Pricing and hedging path-dependent options under the CEV process, *Management Science* **47**, 949–965.
- [10] Davydov, D. & Linetsky, V. (2003). Pricing options on scalar DIFFUSIONS: an eigenfunction expansion approach, *Operations Research* **51**, 185–209.
- [11] Delbaen, F. & Shirakawa, H. (2002). A note of option pricing for constant elasticity of variance model, *Asia-Pacific Financial Markets* **9**, 85–99.
- [12] Ding, C.G. (1992). Computing the non-central χ^2 distribution function, *Applied Statistics* **41**, 478–482.
- [13] Dyrting, S. (2004). Evaluating the noncentral chi-square distribution for the Cox-Ingersoll-Ross process, *Computational Economics* **24**, 35–50.
- [14] Emanuel, D.C. & MacBeth, J.D. (1982). Further results on the constant elasticity of variance call option pricing model, *The Journal of Financial and Quantitative Analysis* **17**, 533–554.
- [15] Feller, W. (1951). Two singular diffusion problems, *The Annals of Mathematics* **54**, 173–182.
- [16] Hagan, P.S., Kumar, D., Lesniewski, A.S. & Woodward, D.E. (2002). Managing smile risk, *Wilmott Magazine* **1**, 84–108.
- [17] Hagan, P. & Woodward, D. (1999). Equivalent black volatilities, *Applied Mathematical Finance* **6**, 147–157.
- [18] Linetsky, V. (2004). Lookback options and diffusion hitting times: a spectral expansion approach, *Finance and Stochastics* **8**, 343–371.
- [19] Mendoza, R., Carr, P. & Linetsky, V. (2007). *Time Changed Markov Processes in Credit-Equity Modeling*, *Mathematical Finance*, to appear.
- [20] Mendoza, R. & Linetsky, V. (2008). *Equity Default Swaps under the Jump-to-Default Extended CEV Model*. Working paper.
- [21] Revuz, D. & Yor, M. (1999). *Continuous Martingales and Brownian Motion*, Grundlehren Der Mathematischen Wissenschaften, Springer.
- [22] Schroder, M. (1989). Computing the constant elasticity of variance option pricing formula, *The Journal of Finance* **44**, 211–219.

VADIM LINETSKY & RAFAEL MENDOZA

Bates Model

The Bates [3] and Scott [13] option pricing models were designed to capture two features of the asset returns: the fact that conditional volatility evolves over time in a stochastic but mean-reverting fashion, and the presence of occasional substantial outliers in the asset returns. The two models combined the Heston [9] model of stochastic volatility (*see Heston Model*) with the Merton [11] model of independent normally distributed jumps in the log asset price (*see Jump-diffusion Models*). The Bates model ignores interest rate risk, while the Scott model allows interest rates to be stochastic. Both models evaluate European option prices numerically, using the Fourier inversion approach of Heston (*see also Fourier Transform and Fourier Methods in Options Pricing* for a general discussion of Fourier transform methods in finance). The Bates model also includes an approximation for pricing American options (*see American Options*). The two models were historically important in showing that the tractable class of affine option pricing models includes jump processes as well as diffusion processes.

All option pricing models rely upon a risk-neutral representation of the data generating process that includes appropriate compensation for the various risks. In the Bates and Scott models, the risk-neutral processes for the underlying asset price S_t and instantaneous variance V_t are assumed to be of the form

$$\begin{aligned} dS_t/S_t &= (b - \lambda^* \bar{k}^*) dt + \sqrt{V_t} dZ_t + k^* dq_t \\ dV_t &= (\alpha - \beta^* V_t) dt + \sigma_v \sqrt{V_t} dZ_{vt} \end{aligned} \quad (1)$$

where b is the cost of carry; Z_t and Z_{vt} are Wiener processes with correlation ρ ; q_t is an integer-valued Poisson counter with risk-neutral intensity λ^* that counts the occurrence of jumps; and k^* is the random percentage jump size, with a Gaussian distribution $\ln(1 + k^*) \sim N\left[\ln(1 + \bar{k}^*) - \frac{1}{2}\delta^2, \delta^2\right]$ conditional upon the occurrence of a jump. The Bates model assumes b is constant, while the Scott model assumes it is a linear combination of V_t and an additional state variable that follows an independent square-root process. Bates [3] examines foreign currency options, for which b is the domestic/foreign interest differential, while Scott's application [13] to nondividend paying stock options implies

the cost of carry is equal to the risk-free interest rate.

The postulated process has an associated conditional characteristic function that is exponentially affine in the state variables. For the Bates model, the characteristic function is

$$\begin{aligned} \varphi(i\Phi) &= E_0^*[e^{i\Phi S_T} | S_0, V_0, T] \\ &= \exp [i\Phi S_0 + C(T; i\Phi) + D(T; i\Phi) V_0 \\ &\quad + \lambda^* T E(i\Phi)] \end{aligned} \quad (2)$$

where $E_0^*[\cdot]$ is the risk-neutral expectational operator associated with equation (1), and

$$\gamma(z) = \sqrt{(\rho\sigma_v z - \beta^*)^2 - \sigma_v^2(z^2 - z)} \quad (3)$$

$$\begin{aligned} C(T; z) &= bTz - \frac{\alpha T}{\sigma_v^2} [\rho\sigma_v z - \beta^* - \gamma(z)] - \frac{2\alpha}{\sigma_v^2} \\ &\quad \times \ln \left\{ 1 + [\rho\sigma_v z - \beta^* - \gamma(z)] \frac{1 - e^{\gamma(z)T}}{2\gamma(z)} \right\} \end{aligned} \quad (4)$$

$$D(T; z) = \frac{z^2 - z}{\gamma(z) \frac{e^{\gamma(z)T} + 1}{e^{\gamma(z)T} - 1} + \beta^* - \rho\sigma_v z} \quad (5)$$

$$E(z) = \left(1 + \bar{k}^*\right)^z e^{\frac{1}{2}\delta^2(z^2 - z)} - 1 - \bar{k}^* z \quad (6)$$

The terms $C(\cdot)$ and $D(\cdot)$ are identical to those in the Heston [9] stochastic volatility model, while $E(\cdot)$ captures the additional distributional impact of jumps. Scott's generalization to stochastic interest rates uses an extended Fourier transform of the form

$$\begin{aligned} \varphi^*(z) &= E_0^* \left[\exp \left(- \int_0^T r_t dt + z \ln S_T \right) | S_0, r_0, V_0, T \right] \end{aligned} \quad (7)$$

which has an analytical solution for complex-valued z that is also exponentially affine in the state variables S_0 , r_0 , and V_0 .

European call option prices take the form $c = B(FP_1 - XP_2)$, where B is the price of a discount bond of maturity T , F is the forward price on the underlying asset, X is the option's exercise price, and

P_1 and P_2 are upper tail probability measures derivable from the characteristic function. The papers of Bates [3] and Scott [13] present Fourier inversion methods for evaluating P_1 and P_2 numerically. However, faster methods were subsequently developed for directly evaluating European call options, using a single numerical integration of the form

$$c = BF - BX \times \left\{ \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \operatorname{Re} \left[\frac{f(i\Phi) e^{-i\Phi \ln X}}{i\Phi(1 - i\Phi)} \right] d\Phi \right\} \quad (8)$$

where $\operatorname{Re}[z]$ is the real component of a complex variable z (see **Fourier Methods in Options Pricing**). For the Bates model, $f(i\Phi) = \varphi(i\Phi)$; for the Scott model, $f(i\Phi) = \varphi^*(i\Phi)/B$. European put options can be evaluated from European call option prices using the put–call parity relationship $p = c + B(X - F)$ (see **Put–Call Parity** for details on put–call parity).

Evaluating equation (8) typically involves integration of a dampened oscillatory function. While there exist canned programs for integration over a semi-infinite domain, most papers use various forms of integration over a truncated domain. Bates [3] uses Gauss–Kronrod quadrature (see **Quadrature Methods**). Fast Fourier transform approaches have also been proposed, but these involve substantially more functional evaluations. The integration is typically well behaved, but there do exist extreme parameter values (e.g., $|\rho|$ near 1) for which the path of integration crosses the branch cut of the log function. As all contemporaneous option prices of a given maturity use the same values of $f(i\Phi)$ regardless of the strike price X , evaluating options jointly greatly increases numerical efficiency.

Related Models

Related affine models can be categorized along four lines:

1. alternate specifications of jump processes;
2. the Bates [5] extension to stochastic-intensity jump processes;
3. models in which the underlying volatility can also jump; and
4. multifactor specifications.

Alternate jump specifications (including Lévy processes) with independent and identically distributed jumps involve modification of the functional form of $E(\cdot)$, and are discussed in other articles: **Tempered Stable Process**; **Normal Inverse Gaussian Model**; **Variance-gamma Model**; **Kou Model**; **Exponential Lévy Models**. The Bates [5] model with (risk-neutral) stochastic jump intensities of the form $\lambda^* + \lambda_1^* V_t$ involves modifying $\gamma(\cdot)$ and $D(\cdot)$:

$$\gamma(z) = \sqrt{(\rho\sigma_v z - \beta^*)^2 - \sigma_v^2[z^2 - z + 2\lambda_1^* E(z)]} \quad (9)$$

$$D(T; z) = \frac{z^2 - z + 2\lambda_1^* E(z)}{\gamma(z) \frac{e^{\gamma(z)T} + 1}{e^{\gamma(z)T} - 1} + \beta^* - \rho\sigma_v z} \quad (10)$$

See also **Time-changed Lévy Process** for other stochastic-intensity jump models.

Bates [5] also contains multifactor specifications for the instantaneous variance and jump intensity. The general class of affine jump-diffusion models is presented in [8], including the volatility-jump option pricing model. Scott's extended Fourier transform approach for stochastic interest rates was subsequently also used by Bakshi and Madan [2] and Duffie *et al.* [8]

Further Reference Material

Bates [7, pp. 943–4] presents a simple derivation of equation (8), and cites earlier papers that develop the single-integration approach. Numerical integration issues are discussed by Lee [10]. Bates [3] and Bakshi *et al.* [1] estimate and test the Bates and Scott models, respectively, while Pan [12] provides additional estimates and tests of the Bates [5] stochastic-intensity model. Bates [4, 6] surveys empirical option pricing research.

References

- [1] Bakshi, G., Cao, C. & Chen, Z. (1997). Empirical performance of alternative option pricing models, *Journal of Finance* **52**, 2003–2049.
- [2] Bakshi, G. & Madan, D.B. (2000). Spanning and derivative-security valuation, *Journal of Financial Economics* **55**, 205–238.

-
- [3] Bates, D.S. (1996). Jumps and stochastic volatility: exchange rate processes implicit in PHLX deutschemark options, *Review of Financial Studies* **9**, 69–107.
 - [4] Bates, D.S. (1996). Testing option pricing models, in *Handbook of Statistics*, G.S. Maddala & C.R. Rao, eds, (Statistical Methods in Finance), Elsevier, Amsterdam, Vol. 14, pp. 567–611.
 - [5] Bates, D.S. (2000). Post-'87 crash fears in the S&P 500 futures option market, *Journal of Econometrics* **94**, 181–238.
 - [6] Bates, D.S. (2003). Empirical option pricing: a retrospection, *Journal of Econometrics* **116**, 387–404.
 - [7] Bates, D.S. (2006). Maximum likelihood estimation of latent affine processes, *Review of Financial Studies* **19**, 909–965.
 - [8] Duffie, D., Pan, J. & Singleton, K.J. (2000). Transform analysis and asset pricing for affine jump-diffusions, *Econometrica* **68**, 1343–1376.
 - [9] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**, 327–344.
 - [10] Lee, R.W. (2004). Option pricing by transform methods: extensions, unification and error control, *Journal of Computational Finance* **7**, 51–86.
 - [11] Merton, R.C. (1976). Option pricing when underlying stock returns are discontinuous, *Journal of Financial Economics* **3**, 125–144.
 - [12] Pan, J. (2002). The jump-risk premia implicit in options: evidence from an integrated time-series study, *Journal of Financial Economics* **63**, 3–50.
 - [13] Scott, L.O. (1997). Pricing stock options in a jump-diffusion model with stochastic volatility and interest rates: applications of Fourier inversion methods, *Mathematical Finance* **7**, 413–426.

Related Articles

Barndorff-Nielsen and Shephard (BNS) Models; Heston Model; Jump-diffusion Models; Stochastic Volatility Models; Foreign Exchange; Time-changed Lévy Process.

DAVID S. BATES

Barndorff-Nielsen and Shephard (BNS) Models

Stochastic volatility models based on non-Gaussian Ornstein–Uhlenbeck (OU)-type processes were introduced in [3]. The motivation was to construct a mathematically tractable model that provides an adequate description of price fluctuations on various timescales. The main idea is to model the volatility with a non-Gaussian OU process: solution of a linear Stochastic differential equation (SDE) with Lévy increments. The non-Gaussian increments allow to build a process that is positive and linear, meaning that many computations are very simple.

Non-Gaussian Ornstein–Uhlenbeck-type Processes

The OU-type process (see [8, 11, 12] for original introduction or [2, 4, 10] for a more modern treatment) is defined as the solution of the stochastic differential equation

$$dy_t = -\lambda y_t + dZ_t \quad (1)$$

where Z is a Lévy process. It can be explicitly written as

$$y_t = y_0 e^{-\lambda t} + \int_0^t e^{-\lambda(t-s)} dZ_s \quad (2)$$

At any time t , the distribution of y_t is infinitely divisible. If the characteristic triplet of Z is (A, ν, γ) , the characteristics of y_t are given by

$$\begin{aligned} A_t^y &= \frac{A}{2\lambda} \{1 - e^{-2\lambda t}\} \\ \gamma_t^y &= \frac{\gamma}{\lambda} \{1 - e^{-\lambda t}\} + y_0 e^{-\lambda t} \\ \nu_t^y &= \int_1^{e^{\lambda t}} \nu(\xi B) \frac{d\xi}{\lambda \xi} \quad \forall B \in \mathcal{B}(\mathbb{R}) \end{aligned} \quad (3)$$

and the characteristic function of y_t is

$$E[e^{iu y_t}] = \exp\left(iu y_0 e^{-\lambda t} + \int_0^t \psi(u e^{\lambda(s-t)}) ds\right) \quad (4)$$

with $\psi(u) = \ln E[e^{iu Z_1}]$. Under the integrability condition $\int_{|x|>1} \ln |x| \nu(dx) < \infty$, the process (y_t) has a stationary distribution with characteristics

$$\begin{aligned} A^y &= \frac{A}{2\lambda}, \quad \gamma^y = \frac{\gamma}{\lambda}, \quad \nu^y(B) \\ &= \int_1^\infty \nu(\xi B) \frac{d\xi}{\lambda \xi}, \quad \forall B \in \mathcal{B}(\mathbb{R}) \end{aligned} \quad (5)$$

In the stationary case, an OU-type process has an exponential (short memory) autocorrelation structure:

$$\text{Cov}(y_t, y_{t+s}) = e^{-\lambda s} \text{Var } y_t \quad (6)$$

To obtain more interesting correlation structures, one can add up several OU-type processes [1]: if y and $\tilde{y}$ are independent stationary OU-type processes with parameters λ and $\tilde{\lambda}$ then

$$\begin{aligned} \text{Cov}(y_t + \tilde{y}_t, y_{t+s} + \tilde{y}_{t+s}) &= e^{-\lambda s} \text{Var } y_t \\ &+ e^{-\tilde{\lambda} s} \text{Var } \tilde{y}_t \end{aligned} \quad (7)$$

The price to be paid is an increased model dimension: the two-dimensional process $(y, \tilde{y})$ is Markov but the sum $y + \tilde{y}$ is not. Superpositions of OU-type processes can also be used to construct finite-dimensional approximations to non-Markov (e.g., long memory) processes.

Positive OU-type processes

Positive OU-type processes can be used as linear models for stationary financial time series such as volatility (discussed below) or commodity prices (see [6]). An OU-type process is positive if the driving Lévy process Z is a positive Lévy process, also known as a *subordinator*. In this case, the trajectory consists of a series of positive jumps with exponential decay between them like in Figure 1.

Model Specification and Examples

The “econometric” (as opposed to risk-neutral) version of the Barndorff-Nielsen and Shephard (BNS) stochastic volatility model has the form

$$\begin{aligned} S_t &= S_0 \exp(X_t) \\ dX_t &= (\mu + \beta \sigma_t^2) dt + \sigma_t dW_t + \rho dZ_t \quad \rho \leq 0 \\ d\sigma_t^2 &= -\lambda \sigma_t^2 dt + dZ_t \quad \sigma_0^2 > 0 \end{aligned} \quad (8)$$

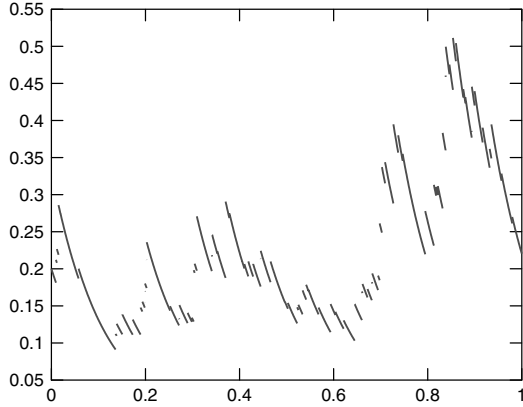


Figure 1 Sample trajectory of a positive OU-type process

The log stock price is a stochastic volatility process with downward jumps (Z has only positive jumps) and the volatility is a positive OU-type process. Introducing Z into the equation for the log price with a negative coefficient accounts for the *leverage effect*: volatility jumps up when price jumps down.

Nicolato and Venardos [9] have shown that model (8) is arbitrage-free, that is, one can always find an equivalent martingale measure. Under a martingale measure, the model takes the form

$$\begin{aligned} S_t &= S_0 \exp(X_t) \\ dX_t &= (r - l(\rho) - \frac{1}{2}\sigma_t^2) dt + \sigma_t dW_t \\ &\quad + \rho dZ_t \\ d\sigma_t^2 &= -\lambda\sigma_t^2 dt + dZ_t, \quad \sigma_0^2 > 0 \end{aligned} \quad (9)$$

where r is the interest rate and $l(u) := \ln E[e^{uZ_1}]$.

As an example of a concrete specification, suppose that the stationary distribution μ of the squared volatility process σ_t^2 is the gamma distribution with density $\mu(x) = \alpha^c / \Gamma(c) x^{c-1} e^{-\alpha x} 1_{x \geq 0}$ (this is the same as the stationary distribution of the volatility in Heston's stochastic volatility model). In this case, (Z_t) has zero drift and a Lévy measure with density $\nu(x) = \alpha \lambda c e^{-\alpha x}$, that is, it is a compound Poisson process with exponential jump size distribution. The Laplace exponent of Z is $l(u) = \lambda c u / \alpha - u$.

Option Pricing and Hedging

The BNS models are a subclass of *affine processes* (see **Affine Models**) [7]: there is an explicit

expression for the characteristic function of log stock price X . Under the risk-neutral probability,

$$\begin{aligned} \phi_t(u) &= E[e^{iuX_t}] \\ &= \exp \left\{ iu(r - l(\rho))t - i\sigma_0^2 \frac{u^2 + iu}{2} \varepsilon(\lambda, t) \right. \\ &\quad \left. + \int_0^t l \left(i\rho u - \frac{u^2 + iu}{2} \varepsilon(\lambda, t-s) \right) ds \right\} \end{aligned} \quad (10)$$

with $\varepsilon(\lambda, t) := 1 - e^{-\lambda t} / t$. This means that if the risk-neutral parameters are known, European options can be priced by Fourier inversion (see **Exponential Lévy Models**). The expected integrated variance is

$$\begin{aligned} E \left[\int_0^t \sigma_s^2 ds \right] &= \sigma_0^2 \frac{1 - e^{-\lambda t}}{\lambda} \\ &\quad + E[Z_1] \frac{e^{-\lambda t} - 1 + \lambda t}{\lambda^2} \end{aligned} \quad (11)$$

leading to a simple explicit formula for the fair rate of a variance swap.

The market being generally incomplete, there exist many risk-neutral probabilities and the prices of contingent claims are not unique. The solution is to select the risk-neutral probability implied by the market, calibrating the model parameters to a set of quoted option prices. Nicolato and Venardos [9] carry out the calibration exercise by the usual technique of nonlinear least squares:

$$\min_{\theta} \sum_{i=1}^N (C_i^M - C^\theta(T_i, K_i))^2 \quad (12)$$

where N is the total number of observations, C_i^M is the observed price of the option with strike K_i and time to maturity T_i , and $C^\theta(T_i, K_i)$ is the price of this option evaluated in a model with parameter vector θ . This method appears to work well in [9] but in other situations two problems may arise:

- **Lack of flexibility:** in BNS models, the same parameter ρ determines the size of jumps in the price process (and hence the short-maturity skew or asymmetry of the implied volatility smile) and the correlation between the price process and the volatility (the long-dated skew). For this reason, the model may be difficult to calibrate in markets

with pronounced skew changes from short to long maturities such as FX markets.

- Lack of stability: since the calibration functional (12) is not convex and the number of model parameters may be large, the calibration algorithm may be caught in a local minimum, which leads to instabilities in the calibration procedure. Usual remedies for this problem include the use of global minimization algorithms such as simulated annealing or adding a convex penalty term to the functional (12) to make the problem well posed.

The minimal variance hedging in BNS models is discussed in [5]. Let the option price at time t be given by $C(t, S_t, \sigma_t^2)$ (this can be computed by Fourier transform). The hedging strategy minimizing the variance of the residual hedging error under the risk-neutral probability is then given by

$$\begin{aligned} \phi_t = & \left[\sigma_{t-}^2 \frac{\partial C}{\partial S} + \frac{1}{S_{t-}} \int v(dz)(e^{\rho z} - 1) \right. \\ & \times [C(t, S_{t-}e^{\rho z}, \sigma_{t-}^2 + z) - C(t, S_{t-}, \sigma_{t-}^2)] \\ & \left. \times \left[\sigma_{t-}^2 + \int (e^{\rho z} - 1)^2 v(dz) \right]^{-1} \right] \end{aligned} \quad (13)$$

When there are no jumps in the stock price ($\rho = 0$), the optimal hedging strategy is just delta-hedging: $\phi_t = \partial C / \partial S$; even though there are jumps in the option price, they cannot be hedged using stock only, because the stock does not jump.

References

- [1] Barndorff-Nielsen, O.E. (2001). Superposition of Ornstein–Uhlenbeck type processes, *Theory of Probability and Its Applications* **45**, 175–194.
- [2] Barndorff-Nielsen, O.E., Jensen, J.L. & Sørensen, M. (1998). Some stationary processes in discrete and continuous time, *Advances in Applied Probability* **30**, 989–1007.
- [3] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck based models and some of their uses in financial econometrics, *Journal of the Royal Statistical Society: Series B* **63**, 167–241.
- [4] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, Chapman & Hall/CRC Press.
- [5] Cont, R., Tankov, P. & Voltchkova, E. (2007). Hedging with options in models with jumps, in *Stochastic Analysis and Applications: The Abel Symposium 2005 in Honor of Kiyosi Ito*, F.E. Benth, G. Di Nunno, T. Lindstrom, B. Øksendal & T. Zhang, eds, Springer, pp. 197–218.
- [6] Deng, S.-J. & Jiang, W. (2005). Lévy process-driven mean-reverting electricity price model: the marginal distribution analysis, *Decision Support Systems* **40**, 483–494.
- [7] Duffie, D., Filipovic, D. & Schachermayer, W. (2003). Affine processes and applications in finance, *Annals of Applied Probability* **13**, 984–1053.
- [8] Jurek, Z.J. & Vervaat, W. (1983). An integral representation for self-decomposable Banach space valued random variables, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **62**(2), 247–262.
- [9] Nicolato, E. & Venardos, E. (2003). Option pricing in stochastic volatility models of Ornstein–Uhlenbeck type, *Mathematical Finance* **13**, 445–466.
- [10] Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge.
- [11] Sato K. & Yamazato M. (1983). Stationary processes of Ornstein–Uhlenbeck type, in *Probability Theory and Mathematical Statistics, Fourth USSR–Japan Symposium*, K. Itô & V. Prokhorov, eds, Lecture Notes in Mathematics, Vol. 1021, Springer, Berlin.
- [12] Wolfe, S.J. (1982). On a continuous analogue of the stochastic difference equation $x_n = \rho x_{n-1} + b_n$, *Stochastic Processes and Applications* **12**(3), 301–312.

Related Articles

Exponential Lévy Models; Lévy Processes; Ornstein–Uhlenbeck Processes.

PETER TANKOV

Heston Model

In the class of stochastic volatility models (*see Stochastic Volatility Models*), Heston's is probably the most well-known model. The model was published in 1993 by S. Heston in his seminal article the title of which readily reveals much of its popularity, *A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options*. It is probably the only stochastic volatility model for equities which both allows very efficient computing of European option prices and fits reasonably well to market data in very different conditions.

In fact, the model was used successfully (in the sense explained below) during the boom of the end of the 1990s, in the brief recession 2001, in the very low volatility regime until 2007, and it still performed well during the very volatile period of late-2008.

However, the model also has several questionable properties: critics point out that its inherent structure as a square-root diffusion does not reflect statistical properties seen in real market data. For example, typical calibrated parameters allow the instantaneous volatility of the stock to become zero with a positive probability. From a practical point of view, the most challenging property of Heston's model is the interdependence of its parameters and the resulting inability to give these parameters a real idiosyncratic meaning. One example is the fact that moving the term structure of volatility has an impact on the shape of the implied volatility skew. This means that traders who use this model will have to have a very good understanding of the dynamics of the model and the interplay between its parameters.

Other stochastic volatility models with efficient pricing methods for European options are: SABR, Schobel–Zhou or Hull–White model (*see Hull–White Stochastic Volatility Model*) and Lewis' "3/2-model" presented in Lewis' book [13]. The n -dimensional extension of Heston's model is the class of affine models [9]. Related are Lévy'-based models that can also be computed efficiently (*see Time-changed Lévy Process*). The most natural model that is used frequently but which actually does not allow efficient pricing of Europeans is a lognormal model for instantaneous volatility.

Model Description

If we assume a prevailing instantaneous interest rate of $r = (r_t)_{t \geq 0}$ and a yield from holding a stock of $\mu = (\mu_t)_{t \geq 0}$, then Heston's model is given as the unique strong solution $Z = (S_t, v_t)_{t \geq 0}$ of the following stochastic differential equation (SDE):

$$\begin{aligned} dv_t &= \kappa(\theta - v_t) dt + \sigma \sqrt{v_t} dW_t \\ dS_t &= S_t (r_t - \mu_t) dt + S_t \sqrt{v_t} dB_t \end{aligned} \quad (1)$$

with starting values spot $S_0 > 0$ and "Short Vol" $\sqrt{v_0} > 0$. In this equation, W and B are two standard Brownian motions with a Correlation of $\rho \in (-1, +1)$. The model is usually specified directly under a risk-neutral measure.

This Correlation together with the "Vol Of Vol" $\sigma \geq 0$ can be thought of being responsible for the skew. This is illustrated in Figure 1: Vol Of Vol controls the volume of the smile and Correlation and its "tilt". A negative Correlation produces the desired downward skew of implied volatility. It is usually calibrated to a value around -70% .

The other parameters control the term structure of the model: in Figure 2, the impact of changing "Short Vol" $\sqrt{v_0} \geq 0$, "Long Vol" $\sqrt{\theta} \geq 0$, and "Reversion Speed" $\kappa > 0$ on the term structure of at-the-money (ATM) implied volatility is illustrated. It can be seen that Short Vol lives up to its name and controls the level of the short-date implied volatilities, whereas Long Vol controls the long end. Reversion Speed controls the skewness or "decay" of the curve from the Short Vol level to the Long Vol level.

This inherent mean-reversion property of Heston's stochastic volatility around a long-term mean $\sqrt{\theta}$ is one of the important properties of the model. Real market data are often mean-reverting, and it also makes economic sense to assume that volatility is not unbounded in its growth as, for example, a stock price process is. In historic data, the "natural" level of mean-reversion is often seen to be itself a mean-reverting process as Fouque *et al.* [10] have shown. Some extensions of Heston in this direction are discussed below.

Parameter Interdependence

Before we proceed, a note of caution: the above distinction of the parameters by their effect on

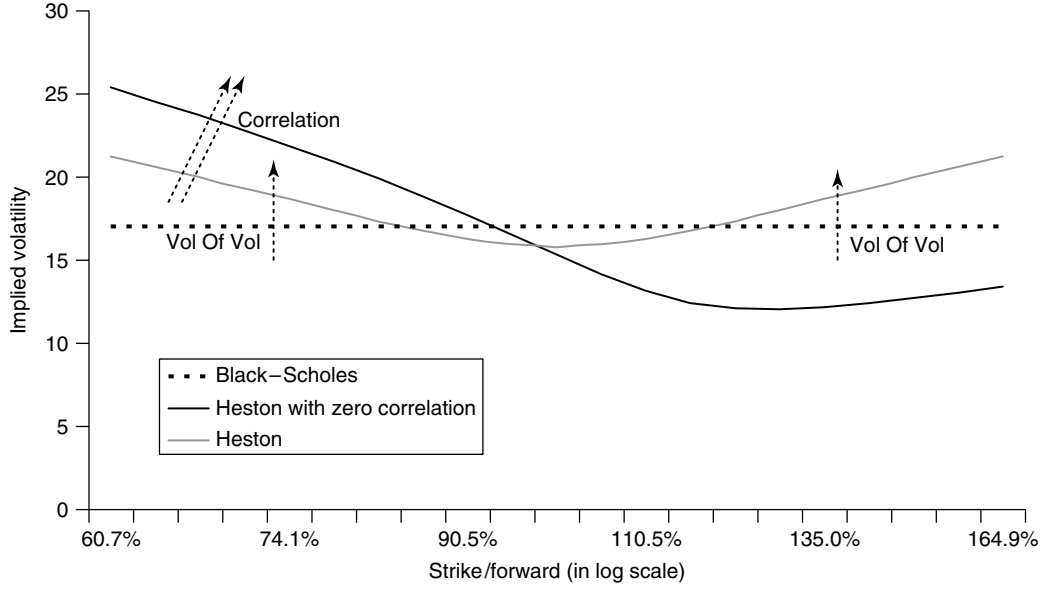


Figure 1 Stylized effects of changing Vol Of Vol and Correlation in Heston's model on the one-year implied volatility. The Heston parameters are $v_0 = 15\%^2$, $\theta = 20\%^2$, $\kappa = 1$, $\rho = -70\%$, and $\sigma = 35\%$

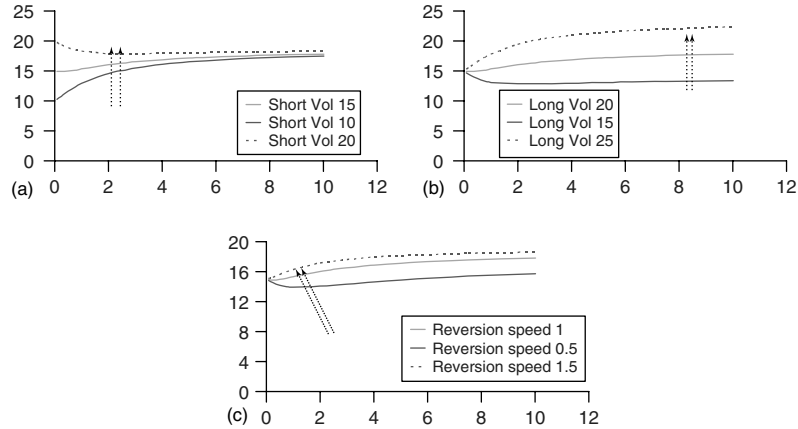


Figure 2 The effects of changing Short Vol (a), Long Vol (b), and Reversion Speed (c) on the ATM term structure of implied volatilities. Each graph shows the volatility term structure for 12 years. The reference Heston parameters are $v_0 = 15\%^2$, $\theta = 20\%^2$, $\kappa = 1$, $\rho = 70\%$, and $\sigma = 35\%$

term structure and strike structure, was made for illustration purposes only: in particular, κ and σ are strongly interdependent if the model is used in the form (1).

This is one of the most serious drawbacks of Heston's model since it means that a trader who uses it to risk-manage a position cannot independently

control the risk with the five available parameters, but has to understand very well their interdependency. For example, to hedge, say, convexity risk in strike direction of the implied volatility surface, the trader will also have to deal with the skew risk at the same time since in Heston, there is no one parameter to control either: convexity is mainly controlled

by Vol Of Vol, but the effect of Correlation on skew depends on the level of Vol Of Vol, too. Moreover, changes to the short end volatility skew will always affect the long-term skew. A similar strong codependency exists between Vol of Vol and Reversion Speed; as pointed out in [14], some of the strong interdependence between Vol Of Vol and Reversion Speed can be alleviated by using the alternative formulation

$$dv_t = (\theta - v_t)\kappa dt + \tilde{\sigma}\sqrt{v_t}\sqrt{\kappa} dW_t \quad (2)$$

In this parametrization, the new Vol Of Vol and reversion speed are much less interdependent, which stabilizes results of daily calibration to market data substantially. Mathematically, this parametrization much more naturally defines κ as the “speed” of the equation.

Such complications are a general issue with stochastic volatility models: since such models attempt to describe an unobservable, rather theoretical quantity (instantaneous variance), they do not produce very intuitive behavior when looked at through the lens of the observable measure of “implied volatility”. That said, implied volatility itself or, rather, its interpolations are also moving on a daily basis. This indicates that natural parameters such as convexity and skew of implied volatility might be a valuable tool for feeding a stochastic volatility model, but it is unreasonable to keep them as constant parameters inside the model.

Pricing European Options

Heston’s popularity is probably mainly derived from the fact that it is possible to price European options on the stock price S using semi-closed-form Fourier transformation, which in turn allows rapid calibration of the model parameters to market data. “Calibration” here means to infer values for the five unobservable parameters $\sqrt{v_0}, \sqrt{\theta}, \kappa, \sigma, \rho$ from market data by minimizing the distance between the models’ European option prices and observed market prices.

We focus on the call prices. Following Carr and Madan [7], we price them *via* Fourier inversion.^a The call price for a relative strike K at maturity T is given as

$$\mathbb{C}(T, K) := \text{DF}(T) \mathbb{E}[(S_T - K F_T)^+] \quad (3)$$

where $\text{DF}(T)$ represents the discount factor and F_T is the forward of the stock. Since the call price itself

is not an L^2 -function in K , we define a “dampened call”

$$c(T, k) := \frac{e^{\alpha k}}{\text{DF}(T) F_T} \mathbb{C}(T, e^k) \quad (4)$$

for an $\alpha > 0$,^b for which its characteristic function $\psi_t(z; k) := \int_{\mathbb{R}} e^{ikz} c(t, k) dk$ is well defined and given as

$$\psi_t(z; k) = \frac{\varphi_t(k - i(\alpha + 1))}{(ik + \alpha)(ik + \alpha + 1)} \quad (5)$$

The function $\varphi_t(z) := \mathbb{E}[\exp\{iz \log S_t/F_t\}]$ is the characteristic function of $X_t := \log S_t/F_t$. Since Heston belongs to the affine model class, its characteristic function has the form

$$\varphi_t(z) = e^{-v_0 A_t - m B_t} \quad (6)$$

with (cf. [14])

$$\begin{aligned} A_t &:= \frac{\alpha + a e^{\gamma t}}{\beta + b e^{\gamma t}} \quad \text{and} \\ B_t &:= \tilde{\kappa} \frac{\alpha b \gamma t + (a\beta - \alpha b) \log \frac{\beta + b e^{\gamma t}}{\beta + b}}{\beta b \gamma} \end{aligned} \quad (7)$$

where $\mu := (iz + z^2)/2$, $\tilde{\kappa} := \kappa - \rho i z \sigma$, $\gamma := -\sqrt{2\mu\sigma^2 + \tilde{\kappa}^2}$, $a := -2\mu$, $\alpha := 2\mu$, $b := -\tilde{\kappa} + \gamma$ and $\beta := \tilde{\kappa} + \gamma$.

We can then price a call on X using

$$\begin{aligned} \mathbb{C}(T, K) &= \text{DF}(T) F_T \frac{e^{-\alpha \ln(K)}}{\pi} \\ &\times \int_0^\infty e^{-iz \ln(K)} \psi_t(z; \ln(K)) dz \end{aligned} \quad (8)$$

The method also lends itself to Fast Fourier Transform if a range of option prices for a single maturity is required.

Similarly, various other payoffs can be computed very efficiently with the Fourier approach, for example, forward started vanilla options, options on integrated short variance, and digital options.

Time-Dependent Parameters

Moreover, for most of these products—and most importantly, plain European options—it is very straightforward to extend the model to time-dependent, piece-wise constant parameters. This is briefly

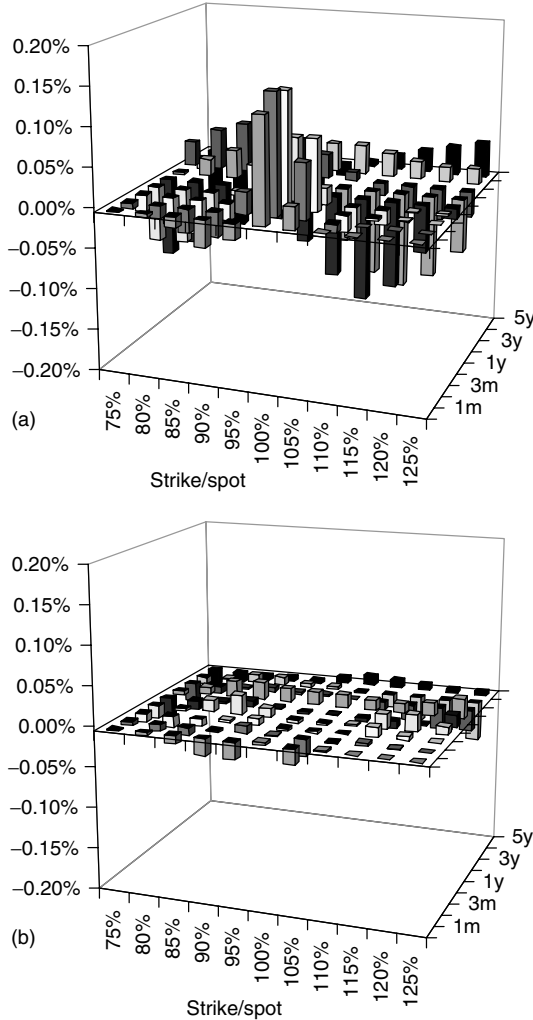


Figure 3 Heston (a) without and (b) with time-dependent parameters fitted to STOX50E for maturities from 1m to 5y. The introduction of time dependency clearly improves the fit

discussed in [14]. It improves the fit of the model to the market prices markedly, cf. Figure 3.

However, it should be noted that by introducing piecewise constant time-dependent parameters, we lose much of a model's structure. It is turned from a time-homogeneous model which "takes a view" on the actual evolution of the volatility *via* its SDE into a kind of an arbitrage-free interpolation of market data: if calibrated without additional constraints to ensure smoothness of the parameters over time, this

is reflected in large discrepancies of the parameter values for distinct periods. For example, the excellent fit of the time-dependent Heston model in Figure 3 is achieved with the following parameter values (short volatility $\sqrt{\zeta_0}$ was 15.0%):

	6m	1y	3y	∞
Long Vol $\sqrt{\theta}$	20.7%	23.6%	36.1%	46.5%
Reversion	5.0	3.2	0.4	0.3
Speed κ				
Correlation ρ	-55.2%	-70.9%	-80.1%	-69.4%
Vol Of Vol σ	78.7%	81.5%	35.3%	60.0

The increased number of parameters also makes it more difficult to hedge in such a model in practice; even though both Heston and the time-dependent Heston models create complete markets, we will always need to additionally protect our position against moves in the parameter values of our model. Just as for Vega in Black and Scholes, this is typically done by computing "parameter greeks" and neutralizing the respective sensitivities. Clearly, the more the parameters that are involved, and the less stable these are, this "parameter hedge" becomes less and less reliable.

Mathematical Drawbacks

The underlying mathematical reason for the relative tractability of Heston's model is that v is a squared Bessel process, which is well understood and reasonably tractable. In fact, a statistical estimation on S&P 500 by Aït-Sahalia and Kimmel [1] of $\alpha \in [1/2, 2]$ in the extended model

$$dv_t = \kappa(\theta - v_t) dt + \sigma v_t^\alpha dW_t^1 \quad (9)$$

has shown that, depending on the observation frequency, a value around 0.7 would probably be more adequate (see **Econometrics of Diffusion Models**). What is more, the square-root volatility terms mean that unless

$$2\kappa\theta \geq \sigma^2 \quad (10)$$

the process v can reach zero with nonzero probability. The crux is that this condition is regularly violated if the model is calibrated freely to observed market data. Although a vanishing short variance is not a problem in itself (after all, a variance of zero simply means absence of trading activity), it makes

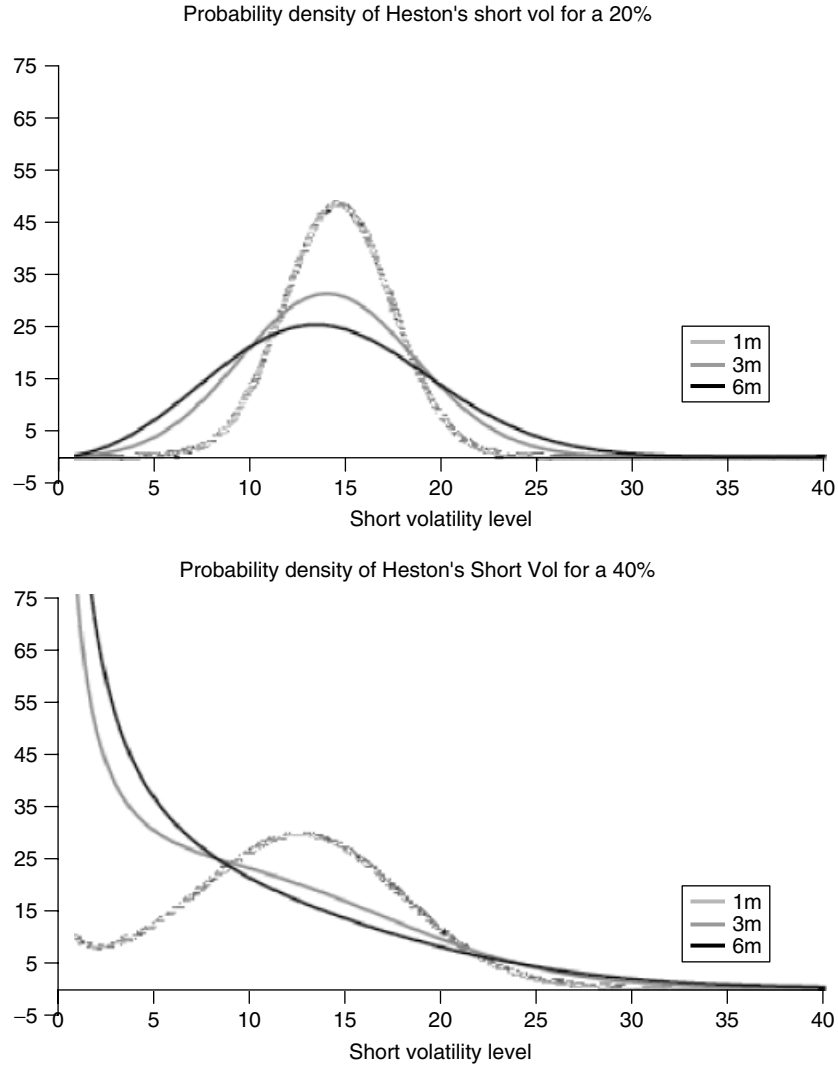


Figure 4 This graphs shows the density of v_t for one, three and six months for the case where condition (10) is satisfied (above) or not (below). Apart from Vol Of Vol, the parameters were $v_0 = 15\%^2$, $\theta = 20\%^2$ and $\kappa = 1$

numerical approximations more complicated. In a Monte Carlo simulation, for example, we have to take the event of v being negative into account. The same problem appears in a partial differential equation (PDE) solver: Heston's PDE becomes degenerate if Short Vol hits zero. A violation of Equation (10) also implies that the distribution of short variance V_t at some later time t is very wide, cf. Figure 4.

Additionally, if Equation (10) does not hold, then the stock price S may fail to have a second moment if the Correlation is not negative enough in the

sense detailed in proposition 3.1 in [2] (see **Moment Explosions** for more details). Again, this is not a problem from a purely mathematical point of view, but it makes numerical schemes less efficient. In particular, Monte Carlo simulations perform much worse: although an Euler scheme will still converge to the desired value, the speed of convergence deteriorates. Moreover, we cannot safely use control variates anymore if the payoff is not bounded.

Pricing Methods

Once we have calibrated the model using the aforementioned semiclosed form of solution for the European options, the question is how to evaluate complex products. At our disposal are PDEs and Monte Carlo schemes.

Since the conditional transition density of the entire process is not known, we have to revert to solving a discretization of the SDE (1) if we want to use a Monte Carlo scheme (*see Monte Carlo Simulation for Stochastic Differential Equations* for an overview of Monte Carlo concepts). To this end, assume that we are given fixing dates $0 = t_0 < \dots < t_N = T$ and let $\Delta t_i := t_{i+1} - t_i$ for $i = 0, \dots, N-1$. Moreover, we denote by ΔW_i for $i = 0, \dots, N-1$ a sequence of independent normal variables with variance Δt_i , and by ΔB_i a corresponding sequence where ΔB_i and ΔW_i have Correlation ρ .

When using a straightforward Euler scheme, we will face the problem that v can become negative. It works well simply to reduce the volatility term of the variance to the positive part of the variance, that is, to simulate

$$v_{t_{i+1}} = v_{t_i} + \kappa(\theta - v_{t_i})\Delta t_i + \sigma\sqrt{v_{t_i}^+}\Delta W_i \quad (11)$$

A flaw of this scheme is that it is biased. This is overcome by using the moment-matching scheme

$$v_{t_{i+1}} = \theta\Delta t_i + (\theta - v_{t_i})e^{-\kappa\Delta t_i} + \left(\sigma v_{t_i}^+ \sqrt{\frac{1 - e^{-2\kappa\Delta t_i}}{2\kappa}} \right) \Delta W_i \quad (12)$$

which works well in practice. To compute the stock price, we approximate the integrated variance over $[t_i, t_{i+1}]$ as

$$\Delta_i V := \theta\Delta t_i + (v_{t_i} - \theta) \frac{1 - e^{-\kappa\Delta t_i}}{\kappa} \quad (13)$$

and set

$$S_{t_k} := F_{t_k} \exp \left\{ \sum_{i=1}^{k-1} \left\{ \sqrt{\Delta_i V} \Delta B_i - \frac{1}{2} \Delta_i V \right\} \right\} \quad (14)$$

Note that this scheme is unbiased in the sense that $\mathbb{E}[S_T] = F_T$.

Heston's PDE

It is straightforward to derive the PDE for the previous model. Let

$$P_t(v, S) := DF_t(T) \mathbb{E}[F(S_T) | S_t = S, v_t = v] \quad (15)$$

be the price of a derivative with maturity T at time t . It satisfies

$$0 = r_t P_t + \partial_S P_t (r_t - \mu_t) S_t + \partial_v P_t \kappa (m - v_t) + \frac{1}{2} \partial_{SS}^2 P_t S_t^2 v_t + \frac{1}{2} \partial_{vv}^2 P_t \sigma^2 v_t + \partial_{vS}^2 P_t \rho v_t S_t \quad (16)$$

with boundary condition $P_T(S, v) = F(S_T)$. To solve this two-factor PDE with a potentially degenerate diffusion term in $\partial_{vv}^2 P_t$, it is recommended to use a stabilized alternating direction implicit (ADI) scheme such as the one described by Craig and Sneyd [8] (*see Alternating Direction Implicit (ADI) Method* for a discussion on ADI).

Risk Management

Provided that we consider not only the stock price itself but also a second liquid instrument V such as a listed option as hedging instrument, stochastic volatility models are complete, that is, in theory every contingent claim P can be replicated in the sense that there are hedging strategies $(\Delta_t, \mathcal{V}_t)_t$ such that

$$dP_t - r_t P_t dt = \Delta_t (dS_t - S_t(r_t - \mu_t) dt) + \mathcal{V}_t (dV_t - r_t V_t dt) \quad (17)$$

(*see Complete Markets* for a discussion on complete markets). In Heston's model, we can write the price process of both the derivative we want to hedge and the hedging instrument as a function of current spot level and short variance, that is, $P_t \equiv P_t(S_t, v_t)$ and $V_t \equiv V_t(S_t, v_t)$. Then, the correct hedging ratios are

$$\mathcal{V}_t = \frac{\partial_v P}{\partial_v V} \quad \text{and} \quad \Delta_t = \partial_S P_t - \frac{\partial_v P}{\partial_v V} \partial_S V_t \quad (18)$$

This is the equivalent of delta hedging in Black and Scholes (*see Delta Hedging*). However, as for the latter, plain theoretical hedging will not work since the other parameters in our model, Reversion Speed, Vol of Vol, Long Vol, and potentially Correlation, will not remain constant if we calibrate our model on a

daily basis. This is the effect of a change in volatility for Black and Scholes—a change of this parameter is not anticipated by the model itself and must be taken care of “outside the model”.

As a result, one way to control this risk is to engage in additional parameter hedging, that is, the desk also displays sensitivities with respect to the other model parameters including, potentially, second-order exposures. Those can then be monitored on a book level and managed explicitly. The drawback of this method is that to reduce risk with respect to those parameters, a portfolio of vanilla options has to be bought whose composition can change quickly if implemented blindly.^c

A second variant is to try to map standard risks of the desk such as implied volatility convexity, skewness, and so on into stochastic volatility risk by “recalibration”. The idea here is that, say, the convexity parameter of the implied volatility is modified, then Heston’s model is calibrated to this new implied volatility surface and the option priced off this model. The resulting change in model price is then considered the sensitivity of the option to convexity in implied volatility. This approach suffers from the fact that typical “implied vol risks” are very different from typical movements in the Heston model. For example, the standard Heston model is homogeneous so it cannot easily accommodate changes in short-term skew only.

Related Models

Owing to its numerical efficiency, Heston’s model is the base for many extensions. The first notable extension is Bates’ addition of jumps to the diffusion process in his article [3] (see **Bates Model**). Jumps are commonly seen as a necessary feature of any risk management model, even though the actual handling of the jump risk part is far from clear.

Bates’ approach can be written as follows: let X be given by

$$\begin{aligned} dv_t &= \kappa(\theta - v_t) dt + \sigma \sqrt{v_t} dW_t \\ dX_t &= X_t \sqrt{v_t} dB_t \end{aligned} \quad (19)$$

and let

$$S_t = F_t X_t e^{\sum_{j=1}^{N_t} \xi_j - \lambda m t} \quad (20)$$

where N_t is a Poisson process with intensity λ (see **Poisson Process**) and where $(\xi_j)_j$ are the normal jumps of the returns of S with mean μ and volatility v . To make sure that S_t/F_t is a martingale we stipulate that $\mu = e^{m + \frac{1}{2}v^2} - 1$.

Since the process X is independent of the jumps, the characteristic function of the log-stock process is the product of the separate characteristic functions. In other words, Bates’ model can be evaluated using the same approach as above and is equally efficient while allowing for a very pronounced short-term skew due to the jump part.^d Figure 5 shows the improvement of time-dependent Bates over time-dependent Heston.

The model has been further enhanced by Knudsen and Nguyen-Ngoc [12] who also added exponentially distributed jumps to the variance process.

Multifactor Models

Structurally, Heston’s model is a member of the class of “affine models” as introduced by Duffie *et al.* [9]. As such, it can easily be extended by mixing in further independent square-root processes. One obvious approach presented in [14] is simply to multiply several independent Heston processes. For the two-factor case, this means to set $S_t := F_t X_t^1 X_t^2$ where both X^1 and X^2 have the form (19). Jumps can be added, but to make the Fourier integration work efficiently, the processes X^1 and X^2 must remain independent.

The stochastic variance of the joint stock price is then simply the sum of the two separate variances, v^1 and v^2 , and it is intuitively assumed that one is a “short-term”, fast mean-reverting process whereas the other is mean reverting slowly. Such a structure is supported by statistical evidence, cf. [10]. However, the independence of the two processes makes it very difficult to impose enough skew into this model since the effective Correlation between instantaneous variance and stock price weakens. In practice, this model is used only rarely.

A related model “Double Heston” has been mentioned by Buehler [6], which is obtained by modeling the mean variance level θ in Heston itself as a square-root diffusion, that is,

$$\begin{aligned} dv_t &= \kappa(\theta_t - v_t) dt + \sigma \sqrt{v_t} dW_t \\ d\theta_t &= c(m - \theta_t) dt + v \sqrt{\theta_t} dW_t^\theta \\ dS_t &= S_t(r_t - \mu_t) dt + S_t \sqrt{v_t} dB_t \end{aligned} \quad (21)$$

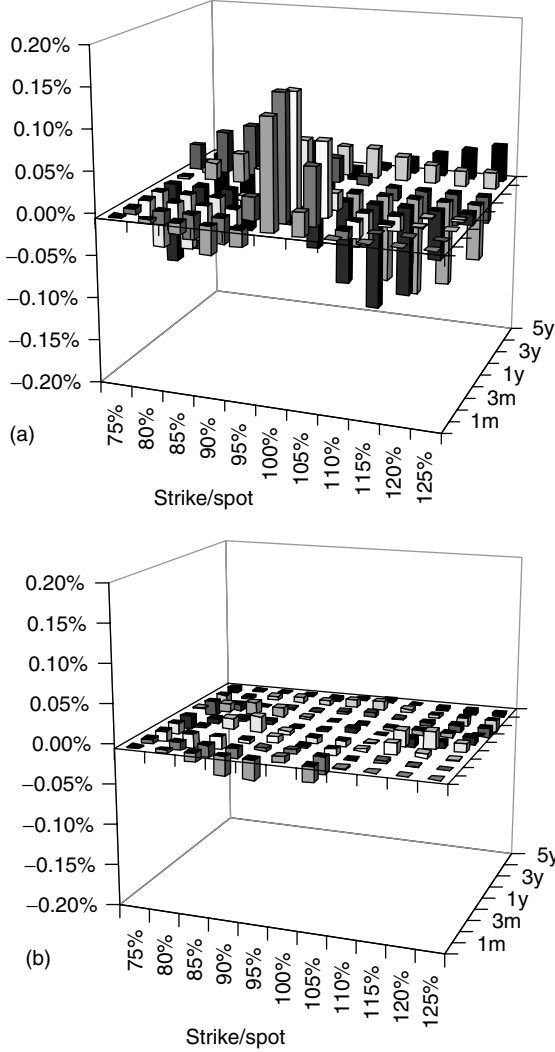


Figure 5 Heston (a) and Bates (b) with time-dependent parameters fitted to STOX50E for maturities from 1m to 5y

where W^θ is independent of W and B . While this model has a reasonably tractable characteristic function, it also suffers from the problem that long-term skew becomes too symmetric, contrary to what is observed in the market. Such a model, however, may have applications when pricing options on variance where the skew counts less and it is more important to be able to account for some dynamics of the term structure of variance. Refer to [6] for an extensive discussion on this.

Fitted Heston

A particular class of derivatives that has gained reasonable popularity in recent years are “Options on Variance”, that is, structures whose terminal payoff depends on the realized variance of the returns of the stock over a set of business days $0 = t_0 < \dots < t_n = T$,

$$\mathbb{R}\mathbb{V}(T) := \sum_{i=1}^n \left(\log \frac{S_{t_i}}{S_{t_{i-1}}} \right)^2 \quad (22)$$

The most standard of such products is a “variance swap” (see **Variance Swap**), which essentially pays the actual realized annualized variance over the period in exchange for a previously agreed fair strike. This strike is usually quoted in volatility terms, that is a variance swap with maturity T , and strike $\Sigma(T)$ pays

$$\frac{252}{n} \mathbb{R}\mathbb{V}(T) - \Sigma^2(T) \quad (23)$$

From this product, a market with options on realized variance has evolved naturally; these include capped variance swaps (mainly traded on single stocks), outright straddles on realized variance swaps, and also VIX futures and options (see **Realized Volatility Options**). Although there are several discussions around how best to approach the risk management of such products, a particularly useful Heston’s model is the “Fitted Heston” approach introduced by Buehler [4].

The main idea here is that to price an option on realized variance in a given model, it is crucial to price correctly a variance swap itself, that is, to make sure that

$$\mathbb{E}[\mathbb{R}\mathbb{V}(T)] = \frac{n}{252} \Sigma^2(T) \quad (24)$$

The idea of “fitting to the market”, say, Heston’s model (2) is now simply to force the model to satisfy this equation. First, assume that we have the term structure of the market’s expected realized variance, $M(T) = n \Sigma^2(T)/252 = \Sigma^2(T)T$, and define $m(t) := \partial_T|_{T=t} M(T)$. Take the original short variance of the model,

$$dv_t = (\theta - v_t)\kappa dt + \sigma \sqrt{v_t} \sqrt{\kappa} dW_t \quad (25)$$

and define the new “fitted” process as

$$w_t := m(t) \frac{v_t}{\mathbb{E}[v_t]} \quad (26)$$

with the stock price as

$$dS_t = S_t(r_t - \mu_t) dt + S_t\sqrt{w_t}dB_t \quad (27)$$

This now reprices all variance swaps automatically in the sense (24). Note that this method does not at all depend on using Heston's model and can be applied to any stochastic volatility model as long as the expectation of instantaneous variance is known.

As pointed out in [6], this model is naturally very attractive from a risk-management point of view if the input M is computed on the fly within the risk management system. In this case, the risk embedded in the variance swap level (called *VarSwapDelta*) is automatically reflected back in the standard implied volatility risk, and the underlying stochastic volatility model is used purely to control skew and convexity around the variance swap backbone.^c Further practical considerations and the impact of jumps are discussed by Buehler in [5].

End Notes

^aIn his original paper [11], Heston suggested a numerically more expensive approach via numerical integration that is twice as slow but still much faster than the same computation for most other models. The approach to price with Fourier inversion is due to Carr and Madan [7]; the interested reader finds more details on the subject in Lewis's book [13].

^bSee [7] for a discussion on the choice of α .

^cBermudez *et al.* discuss one approach to find such portfolios [14].

^dIn practice, calibrating all parameters (stochastic volatility plus jumps) together is relatively unstable since the two parts play similar roles for the short-term options. It is therefore customary to fix the jump parameters themselves or to calibrate them separately to very short-term options.

^eUsually, the parameters v_0 and θ are fixed to some "usual level" such as 20%. Then, they do not need to be calibrated anymore and, in addition, σ retains some comparability to the standard Heston model.

References

- [1] Aït-Sahalia, Y. & Kimmel, R. (2004). *Maximum Likelihood Estimation of Stochastic Volatility Models*, NBER Working Paper No. 10579, June 2004.
- [2] Andersen, L. & Piterbarg, V. (2007). Moment explosions in stochastic volatility models, *Finance and Stochastics* **11**(1), 20–50.
- [3] Bates, D. (1996). Jumps and stochastic volatility: exchange rate process implicit in the Deutschemark options, *Review of Financial Studies* **9**(1), 69–107.
- [4] Buehler, H. (2006). Consistent variance curve models, *Finance and Stochastics* **10**(2), 178–203.
- [5] Buehler, H. (2006). Options on variance: pricing and hedging, *Presentation, IQPC Volatility Trading Conference*, London November 28th, 2006, <http://www.quantitative-research.de/dl/IQPC2006-2.pdf>
- [6] Buehler, H. (2006). *Volatility Markets: Consistent Modeling, Hedging and Practical Implementation*, PhD thesis TU Berlin, <http://www.quantitative-research.de/dl/HansBuehlerDiss.pdf>
- [7] Carr, P. & Madan, D. (1999). Option pricing and the Fast Fourier Transform, *Journal of Computational Finance* **2**(4), 61–73. Summer.
- [8] Craig, I.J.D. & Sneyd, A.D. (1988). An alternating-direction implicit scheme for parabolic equations with mixed derivatives, *Computers and Mathematics with Applications* **16**(4), 341–350.
- [9] Duffie, D., Pan, J. & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions, *Econometrica* **68**, 1343–1376.
- [10] Fouque, J.-P., Papanicolaou, G. & Sircar, K. (2000). *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge Press.
- [11] Heston, S. (1993). A closed-form solution for option with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**(2), 327–343.
- [12] Knudsen, T. & Nguyen-Ngoc, L. (2000). The Heston model steps further Deutsche Bank Quantessence **1**(7) <https://www.dbconvertibles.com/dbquant/quantessence/Vol1Issue7External.pdf>
- [13] Lewis, A. (2000). *Option Valuation under Stochastic Volatility*, Finance Press.
- [14] Overhaus, M., Bermudez, A., Buehler, H., Ferraris, A., Jordinson, C., Lamnouar, A. (2006). *Equity Hybrid Derivatives*, Wiley.

Related Articles

Alternating Direction Implicit (ADI) Method; Bates Model; Complete Markets; Cliquet Options; Econometrics of Diffusion Models; Hedging; Hull–White Stochastic Volatility Model; Model Calibration; Monte Carlo Simulation for Stochastic Differential Equations; Moment Explosions; Realized Volatility Options; Variance Swap.

HANS BUEHLER

Hull–White Stochastic Volatility Model

Even before practitioners started using the **Black–Scholes formula** extensively, one had identified the assumption of constant volatility as unrealistic. Empirical observation of equity vanilla option market shows, indeed, that the implied volatility level depends on the strike. This feature, commonly known as the *volatility smile*, violates the constant volatility assumption. This essential remark motivated the birth of stochastic volatility models (*see Stochastic Volatility Models*).

Among the first authors to tackle this issue, Hull and White proposed in 1987 a simple extension of Black–Scholes model [1]. This article aims at presenting a sound introduction to the Hull–White stochastic volatility model and at indicating its implications in terms of volatility behavior and correlation.

Hull and White describe the variance $V = \sigma^2$ as a geometric Brownian motion. Therefore, the asset and variance satisfy the following **stochastic differential equation**:

$$dS = \phi S dt + \sigma S dw_t \quad (1)$$

$$dV = \mu V dt + \xi V dz_t \quad (2)$$

$$dw_t, dz_t = \rho dt \quad (3)$$

In its general formulation, the parameter ϕ may depend on S , σ , and t , while parameters μ and ξ may depend on σ and t . One may find that many models fall under these dynamics, including the **Heston model** for $\mu = \kappa (\theta/V - 1)$ and $\xi = (\nu/\sqrt{V})$. As in [1], we will restrict to the constant parameters case (Hull and White studied the mean-reverting variance case in [2]).

Option Pricing

Let f be the price of a security which depends on the stock price. f satisfies the **partial differential equation** (PDE)

$$\frac{\partial f}{\partial t} + \frac{1}{2} \left[\sigma^2 S^2 \frac{\partial^2 f}{\partial S^2} + 2\rho\sigma^3\xi S \frac{\partial^2 f}{\partial S \partial V} + \xi^2 V^2 \frac{\partial^2 f}{\partial V^2} \right]$$

$$-rf = -rS \frac{\partial f}{\partial S} - \mu\sigma^2 \frac{\partial f}{\partial V} \quad (4)$$

Assuming volatility and stock price are uncorrelated, we derive an analytic solution to equation (4) through risk-neutral valuation procedure

$$f(S_t, \sigma_t^2, t) = e^{-r(T-t)} \times \int_0^\infty f(S_T, \sigma_T^2, T) p(S_T | S_t, \sigma_t^2) dS_T \quad (5)$$

where T is the option maturity, S_t is the security at pricing time t , σ_t is the instantaneous volatility at time t , and $p(S_T | S_t, \sigma_t^2)$ is the conditional distribution of S_T given the security price and variance at time t .

Introducing the mean variance over the option life $\bar{V}$,

$$\bar{V} = \frac{1}{T-t} \int_t^T \sigma_\tau^2 d\tau \quad (6)$$

we express the call option value as

$$C_{\text{HW}}(S_t, \sigma_t^2, t) = \int_0^\infty C_{\text{BS}}(\bar{V}) h(\bar{V} | \sigma_t^2) d\bar{V} \quad (7)$$

where C_{BS} is the Black–Scholes call option value and $h(\cdot, \sigma_t^2)$ is the conditional density of $\bar{V}$ given σ_t^2 .

In some particular cases, such as when μ and σ are constant, this expression admits an explicit Taylor expansion which converges quickly for small values of $\xi^2(T-t)$.

Behavior of Volatility

Assuming that parameters μ and ξ are constant, the first two moments of the volatility process are given by

$$E[\sigma(t)] = \sigma(0) \times e^{\frac{1}{2} \left(\mu - \frac{1}{4} \xi^2 \right) t} \quad (8)$$

$$V[\sigma(t)] = \sigma(0)^2 \times e^{\mu t} \left(1 - e^{\frac{1}{8} \xi^2 t} \right) \quad (9)$$

For $\mu < \frac{1}{4} \xi^2$, the expectation of volatility converges to zero, whereas for $\mu > \frac{1}{4} \xi^2$, it diverges. Regarding the variance of volatility, it increases unbounded if $\mu > 0$.

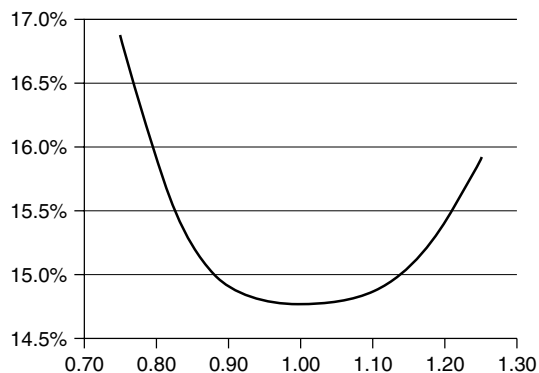


Figure 1 Implied volatility as a function of strike

When calibrating the model to the market, we are very likely to find $\mu \leq 0$, since variance of volatility is bounded. Hence, the expectation of volatility converges to either zero or its initial value.

Implied Volatility Smile

In Figure 1, we show implied volatility as a function of moneyness (strike divided by forward). The call option price has been computed using the Taylor expansion of equation (7) with $\sigma_0 = 0.15$, $\mu = 0$, $\xi = 0$, and $r = 0$. Compared to the Hull–White model, the Black–Scholes model overprices at-the-money options and underprices in- and out-of-the-money options.

Correlation between Stock Returns and Changes in Volatility

By introducing correlation between stock and variance Gaussian increments, Hull and White incorporate explicitly a cause of the volatility skew: the leverage effect. Even if they do not provide any analytic formula in the correlated case, one can still analyze the impact of correlation through numerical simulation.

As shown in Figure 2, this correlation has a huge impact since it enables to transform the smile into

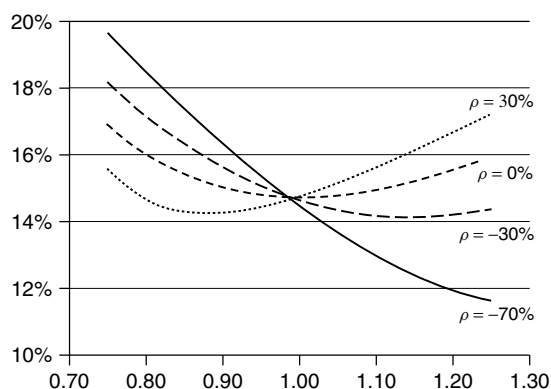


Figure 2 Volatility smile for various correlation levels

a skew. In order to fit market data, it is crucial to correctly set the correlation parameter.

One drawback of the Hull–White model is the lack of mean-reverting behavior in the volatility process (see **Stochastic Volatility Models**).

References

- [1] Hull, J. & White, A. (1987). The pricing of options on assets with stochastic volatilities, *The Journal of Finance* 17(2), 281–300.
- [2] Hull, J & White, A. (1988). An analysis of the bias in option pricing caused by a stochastic volatility, *Advances in Futures and Options Research* 3, 27–61.

Related Articles

Heavy Tails; Heston Model; Implied Volatility in Stochastic Volatility Models; Implied Volatility Surface; Partial Differential Equations; Stochastic Volatility Models; Stylized Properties of Asset Returns.

PIERRE GAUTHIER & PIERRE-YVES H.
RIVAILLE

Tempered Stable Process

A tempered stable process is a pure-jump Lévy process (see **Lévy Processes**) with infinite activity (see **Exponential Lévy Models**) whose small jumps behave like a stable process, while the large jumps are “tempered” so that the tail of the density decays exponentially. Tempered stable processes can be constructed from stable processes by exponential tilting (see **Esscher Transform**) of the Lévy measure.

Tempered stable processes were introduced in [8] and introduced in financial modeling by Cont *et al.* [4] under the name truncated stable process, where it was noted that tempered stable processes have a short-time behavior similar to stable Lévy processes while retaining finite variance and finite exponential moments. Option pricing with tempered stable processes was studied in [1], [2], and [5].

The best known example of tempered stable processes is the CGMY process introduced in [2], which is a pure-jump Lévy process with Lévy density given by

$$k_{\text{CGMY}} = C \frac{\exp(-G|x|)}{|x|^{1+Y}} 1_{\{x < 0\}} + C \frac{\exp(-M|x|)}{|x|^{1+Y}} 1_{\{x > 0\}} \quad (1)$$

The model parameters of equation (1) fulfill $C > 0$, $G, M \geq 0$, and $Y \in (-\infty, 2)$. The restriction on the parameter Y ensures that the measure is a Lévy measure.

For a given stochastic process $X(t)$, its characteristic function is given by $\Phi(u, t) = \mathbb{E}[\exp(iuX(t))]$ (see **Fourier Transform; Fourier Methods in Options Pricing**). For the CGMY model, it is derived in [2] and it is given by

$$\Phi(u, t) = \exp(tC \Gamma(-Y)((M - iu)^Y - M^Y + (G + iu)^Y - G^Y)) \quad (2)$$

On this basis, Fourier-transform methods (see **Fourier Transform; Fourier Methods in Options Pricing**) can be applied to option pricing. Carr *et al.* show that the CGMY process has completely monotone Lévy density for $Y > -1$ and is of infinite activity for $Y > 0$. The drift parameter is chosen to

make $S(t)$ into a martingale and can be determined by using equation (2), which leads to $\omega = -\ln(\Phi(-i))$. Further methods to compute an equivalent martingale measure are discussed in [7].

Tempered Stable

The CGMY process is a special case of the tempered stable process considered in Boyarenko and Levendorskii [1], Cont and Tankov [5] or Rosinski [12]. The latter process has Lévy measure with density given by

$$k_{\text{TS}} = C_- \frac{\exp(-G|x|)}{|x|^{1+Y_-}} 1_{\{x < 0\}} + C_+ \frac{\exp(-M|x|)}{|x|^{1+Y_+}} 1_{\{x > 0\}} \quad (3)$$

The parameters of equation (3) fulfill $G, M > 0$, $C_{\pm} > 0$, and $Y_{\pm} \in (-\infty, 2)$. The characteristic function is available in closed form and hence option pricing and calibration can be performed using Fourier-transform methods. Choosing $C_- = C_+$ and $Y_- = Y_+$ leads to the CGMY process and $Y_- = Y_+ = 0$ leads to a variance gamma process (see **Variance-gamma Model**).

Interpretation of the Parameters

In order to show the impact of the model parameters to asset returns, we consider the properties of the process $X(t)$. Increasing C makes the density more peaked while decreasing C flattens it. C controls the frequency of jumps. While determining the probability of jumps larger than a certain level, this parameter is incorporated. The parameter Y governs the fine structure of the process and the choice affects the overall properties of the process as explained in the previous section. It determines if the process is of finite or infinite activity.

The parameters G and M control the rate of exponential decay, that is, the tail behavior, on the right and the left of k_{CGMY} . We consider three cases: $G = M$ leads to a symmetric Lévy measure, $G < M$ makes the left tail heavier than the right one, and *vice versa* for the case $G > M$. The last two cases lead to a skewed distribution.

This behavior is illustrated in Figure 1.

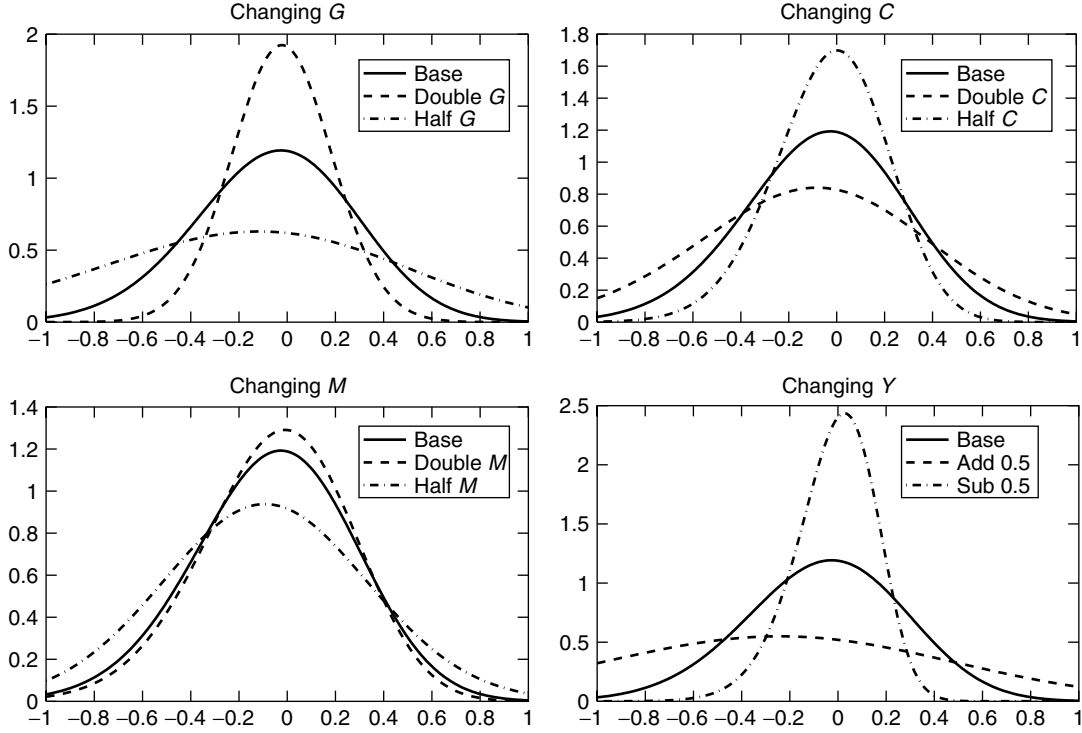


Figure 1 Illustration of the effect of changing the CGMY model parameters C , G , M , and Y on the probability density function

If variance, skewness, and kurtosis exist, they can be computed by

$$\text{Variance} = C\Gamma(2-Y) \left(\frac{1}{M^{2-Y}} + \frac{1}{G^{2-Y}} \right) \quad (4)$$

$$\text{Skewness} = \frac{C\Gamma(3-Y) \left(\frac{1}{M^{3-Y}} + \frac{1}{G^{3-Y}} \right)}{V^{3/2}} \quad (5)$$

$$\text{Kurtosis} = \frac{C\Gamma(4-Y) \left(\frac{1}{M^{4-Y}} + \frac{1}{G^{4-Y}} \right)}{V^2} \quad (6)$$

The equations for the higher moments suggest that the parameter C controls the overall size of the moments. This has already been verified by the expression for the density. In the case $\int_{\mathbb{R}} k(x) dx < +\infty$, it can be interpreted as a measure for the overall level of activity. In the case of finite activity, the process has a finite number of jumps on every compact interval.

Pricing and Calibration

We can use the characteristic function of the log-price $X(t)$ from equation (2) to apply Fourier methods described in Carr and Madan [3] or Eberlein *et al.* [6] to price European and path-dependent options (*see Fourier Methods in Options Pricing*).

Options may also be priced by Monte Carlo simulation [10, 11] using the representation of the tempered stable process as a subordinated Brownian motion [5, Proposition 4.1].

In contrast to the diffusion processes, for the pure-jump processes, the change in measure can be computed from the statistical measure, that is, using parameters computed from time-series data and risk-neutral measure, that is, using parameters obtained using quoted option prices. It holds that $k_{\tilde{\mathbb{P}}}(x) = Y(x)k_{\mathbb{P}}$; see [2] for details. Let us call the corresponding parameter sets $\mathcal{P} = \{C, G, M, Y, \mu\}$ and $\tilde{\mathcal{P}} = \{\tilde{C}, \tilde{G}, \tilde{M}, \tilde{Y}, r\}$, where r denotes the riskless rate and the corresponding measures by $\mathbb{P}$ and $\tilde{\mathbb{P}}$ respectively. If the characteristic functions are denoted by Φ and

$\tilde{\mathbb{P}}$ then using the results in [2] state that $\tilde{\mathbb{P}}$ is an equivalent martingale measure to $\mathbb{P}$ if and only if $C = \tilde{C}$, $Y = \tilde{Y}$, and $r - \Phi(-i) = \mu - \tilde{\Phi}(-i)$. The constraints on the parameters $G, M, \tilde{G}$, and $\tilde{M}$ are implicit in the last equality.

Path Properties

Path properties of the model affect the prices of exotic path-dependent options. We considered path variation when we gave the interpretation for the model parameters. Other concepts like hitting points, creeping, or regularity of the half-line are considered in [9]. We shortly introduce hitting points. The process X_t can hit a point $x \in \mathbb{R}$ if $\mathbb{P}(X_t = x \text{ for at least one } t > 0) > 0$. We denote the set of all points the process can hit by $H = \{x \in \mathbb{R} | \mathbb{P}(X_t = x \text{ for at least one } t > 0) > 0\}$. See [9] for details.

References

- [1] Boyarchenko, S.I. & Levendorskii, S.Z. (2002). *Non-Gaussian Merton-Black-Scholes theory*, Advanced Series on Statistical Science and Applied Probability, World Scientific, River Edge, NJ, Vol. 9.
- [2] Carr, P., Geman, H., Madan, D. & Yor, M. (2002). The fine structure of asset returns: an empirical investigation, *Journal of Business* **75**(2), 305–332.
- [3] Carr, P. & Madan, D. (1999). Option valuation using the fast Fourier transform, *Journal of Computational Finance* **2**(4), 61–73.
- [4] Cont, R., Potters, M. & Bouchaud, J.P. (1997). Scaling in stock market data: stable laws and beyond, in *Scale Invariance and Beyond*. B. Dubrulle, F. Graner & D. Sornette, eds, Springer.
- [5] Cont, R. & Tankov, P. (2003). *Financial Modelling with Jump Processes*, Chapman and Hall / CRC Press.
- [6] Eberlein, E., Glau, K. & Papapantoleon, A. (2008) *Analysis of Valuation Formulae and Applications to Exotic Options.. Preprint Uni Freiburg*, www.stochastic.uni-freiburg.de/~eberlein/papers/Eberlein-glau.Papapan.pdf
- [7] Kim, Y.S. & Lee, J.H. (2007). The relative entropy in CGMY processes and its applications to finance, *Mathematical Methods of Operations Research* **66**(2), 327–338.
- [8] Koponen, I. (1995). Analytic approach to the problem of convergence of truncated Lévy flights towards the Gaussian stochastic process, *Physical Review E* **52**, 1197–1199.
- [9] Kyprianou, A.E. & Loeffen, T.L. (2005). Lévy processes in finance distinguished by their coarse and fine path properties, in *Exotic Option Pricing and Advanced Lévy models*, A.E. Kyprianou, W. Schoutens & P. Wilmott, eds, Wiley, Chichester.
- [10] Madan, D. & Yor, M. (2005). *CGMY and Meixner Subordinators are Absolutely Continuous with Respect to One Sided Stable Subordinators*. *Prépublication du Laboratoire de Probabilités et Modèles Aléatoires*.
- [11] Poirrot, J. & Tankov, P. (2006). Monte Carlo option pricing for tempered stable (CGMY) processes, *Asia Pacific Financial Markets* **13**(4), 327–344.
- [12] Rosinski, J (2007). *Tempering stable processes*, *Stochastic Processes and their Applications*, **117**(6), 677–707.

Related Articles

Exponential Lévy Models; Fourier Methods in Options Pricing; Fourier Transform; Lévy Processes; Time-changed Lévy Process.

JÖRG KIENITZ

Lognormal Mixture Diffusion Model

Let us denote the time- t price of a given financial asset by $S(t)$, equivalently S_t . We say that S evolves according to a local-volatility model (see also **Local Volatility Model**) if, under the risk-neutral measure,

$$\begin{aligned} dS(t) &= \mu S(t)dt + \sigma(t, S(t))S(t)dW(t), \\ S(0) &= S_0 \end{aligned} \quad (1)$$

where S_0 is a positive constant, W is a standard Brownian motion, σ is a well-behaved deterministic function, and μ is the risk-neutral drift rate, which is assumed to be constant. For instance, in case of a stock paying a continuous dividend yield q , $\mu = r - q$, where r is the (assumed constant) continuously compounded risk-free rate.

Brigo and Mercurio [1–3] find an explicit expression for the function σ such that the resulting process has a density that, at each time, is given by a mixture of lognormal densities. Their result is briefly reviewed in the following.

Let us consider N functions σ_i 's that are deterministic and bounded from above and below by positive constants, and corresponding lognormal densities

$$\begin{aligned} p_i^j(y) &= \frac{1}{y V_i(t) \sqrt{2\pi}} \\ &\times \exp \left\{ -\frac{1}{2V_i^2(t)} \left[\ln \frac{y}{S_0} - \mu t + \frac{1}{2} V_i^2(t) \right]^2 \right\} \end{aligned} \quad (2)$$

$$V_i(t) := \sqrt{\int_0^t \sigma_i^2(u) du} \quad (3)$$

Proposition 1 *Let us assume that each σ_i is also continuous and that there exists an $\varepsilon > 0$ such that $\sigma_i(t) = \sigma_0 > 0$, for each t in $[0, \varepsilon]$ and $i = 1, \dots, N$. Then, if we set*

$$v(t, y) = \sqrt{\frac{\sum_{i=1}^N \lambda_i \sigma_i^2(t) \frac{1}{V_i(t)} \exp \left\{ -\frac{1}{2V_i^2(t)} \left[\ln \frac{y}{S_0} - \mu t + \frac{1}{2} V_i^2(t) \right]^2 \right\}}{\sum_{i=1}^N \lambda_i \frac{1}{V_i(t)} \exp \left\{ -\frac{1}{2V_i^2(t)} \left[\ln \frac{y}{S_0} - \mu t + \frac{1}{2} V_i^2(t) \right]^2 \right\}}} \quad (4)$$

for $(t, y) > (0, 0)$ and $v(t, y) = \sigma_0$ for $(t, y) = (0, S_0)$, the SDE

$$dS_t = \mu S_t dt + v(t, S_t) S_t dW_t \quad (5)$$

has a unique strong solution whose marginal density is given by the mixture of lognormals

$$\begin{aligned} p_t(y) &= \sum_{i=1}^N \lambda_i \frac{1}{y V_i(t) \sqrt{2\pi}} \\ &\times \exp \left\{ -\frac{1}{2V_i^2(t)} \left[\ln \frac{y}{S_0} - \mu t + \frac{1}{2} V_i^2(t) \right]^2 \right\} \end{aligned} \quad (6)$$

Moreover, for $(t, y) > (0, 0)$, we can write $v^2(t, y) = \sum_{i=1}^N \Lambda_i(t, y) \sigma_i^2(t)$, where, for each (t, y) and i , $\Lambda_i(t, y) \geq 0$ and $\sum_{i=1}^N \Lambda_i(t, y) = 1$. As a consequence, for each $t, y > 0$,

$$\begin{aligned} 0 < \tilde{\sigma} &:= \inf_{t \geq 0} \left\{ \min_{i=1, \dots, N} \sigma_i(t) \right\} \leq v(t, y) \leq \hat{\sigma} \\ &:= \sup_{t \geq 0} \left\{ \max_{i=1, \dots, N} \sigma_i(t) \right\} < +\infty \end{aligned} \quad (7)$$

A proof of this proposition can be found in [2], and more formally in [5].

The pricing of European options under the lognormal-mixture local-volatility model is quite straightforward (see also **Risk-neutral Pricing; Black–Scholes Formula**).

Proposition 2 *Consider a European option with maturity T , strike K , and written on the asset. The option value at the initial time $t = 0$ is given by the following convex combination of Black–Scholes prices:*

$$\pi(K, T) = \omega P(0, T) \sum_{i=1}^N \lambda_i \times \left[S_0 e^{\mu T} \Phi \left(\frac{\ln \frac{S_0}{K} + \left(\mu + \frac{1}{2} \eta_i^2 \right) T}{\eta_i \sqrt{T}} \right) - K \Phi \left(\frac{\ln \frac{S_0}{K} + \left(\mu - \frac{1}{2} \eta_i^2 \right) T}{\eta_i \sqrt{T}} \right) \right] \quad (8)$$

where $P(0, T)$ is the discount factor for maturity T , Φ is the normal cumulative distribution function, $\omega = 1$ for a call and $\omega = -1$ for a put, and

$$\eta_i := \frac{V_i(T)}{\sqrt{T}} = \sqrt{\frac{\int_0^T \sigma_i^2(t) dt}{T}} \quad (9)$$

The main advantage of the lognormal-mixture local-volatility model is its tractability (explicit marginal density and option prices). This model can be successfully used in practice to calibrate

smile-shaped implied volatility structures. Extensions allowing for nonzero slopes at the at-the-money level are introduced in [4].

References

- [1] Brigo, D. & Mercurio, F. (2000). A mixed-up smile, *Risk* September, 123–126.
- [2] Brigo, D. & Mercurio, F. (2001). Displaced and mixture diffusions for analytically-tractable smile models, in *Mathematical Finance—Bachelier Congress 2000*, H. Geman, D.B. Madan, S.R. Pliska & A.C.F. Vorst, eds, Springer Finance, Springer, Berlin, Heidelberg, New York.
- [3] Brigo, D. & Mercurio, F. (2002). Lognormal-mixture dynamics and calibration to market volatility smiles, *International Journal of Theoretical and Applied Finance* 5(4), 427–446.
- [4] Brigo, D., Mercurio, F. & Sartorelli, G. (2003). Alternative asset-price dynamics and volatility smile, *Quantitative Finance* 3(3), 173–183.
- [5] Sartorelli, G. (2004). *Density Mixture Ito Processes*. PhD thesis, Scuola Normale Superiore di Pisa.

FABIO MERCURIO

Normal Inverse Gaussian Model

The normal inverse Gaussian (NIG) process is an example of a Lévy process (see **Lévy Processes**) with no Brownian component.

We first discuss the NIG distribution and its main properties. The NIG process can be constructed either as process with NIG increments or, alternatively, defined *via* random time change of Brownian motion using the inverse Gaussian process to determine time. Further, we present the NIG market model and show how one can price European options under this model. Option pricing can be done using the NIG density function, the NIG Lévy characteristics, or the NIG characteristic function.

The Normal Inverse Gaussian Distribution

The NIG distribution with parameters $\alpha > 0$, $-\alpha < \beta < \alpha$, and $\delta > 0$ has characteristic function (see [1])

$$\phi(u; \alpha, \beta, \delta) = \exp \left\{ -\delta \left(\sqrt{\alpha^2 - (\beta + iu)^2} - \sqrt{\alpha^2 - \beta^2} \right) \right\} \quad (1)$$

We shall denote this distribution by $NIG(\alpha, \beta, \delta)$. The distribution is so named due to the fact that $NIG(\alpha, \beta, \delta)$ is a variance–mean mixture of a normal distribution with the inverse Gaussian as the mixing distribution. It follows immediately from expression (1) that this distribution is infinitely divisible. The distribution is defined on the whole real line and has the density function

$$f(x; \alpha, \beta, \delta) = \alpha \delta \pi^{-1} \exp \left\{ \delta \sqrt{\alpha^2 - \beta^2} + \beta x \right\} K_1(\alpha \sqrt{\delta^2 + x^2}) \left(\sqrt{\delta^2 + x^2} \right)^{-1}, \quad x \in \mathbb{R} \quad (2)$$

where K_1 is the modified Bessel function of third-order and index 1. If a random variable X follows an $NIG(\alpha, \beta, \delta)$ distribution and $c > 0$, then cX is $NIG(\alpha/c, \beta/c, c\delta)$ -distributed. Further, if $X \sim NIG(\alpha, \beta, \delta_1)$ is independent of $Y \sim NIG(\alpha, \beta, \delta_2)$, then $X + Y \sim NIG(\alpha, \beta, \delta_1 + \delta_2)$. If $\beta = 0$, the

distribution is symmetric. This can easily be seen from the characteristics of the NIG distribution given in Table 1.

Note that the NIG distribution is a special case of generalized hyperbolic distribution, and it can approximate most hyperbolic distributions very closely. In modeling, the NIG distribution can describe observations with considerably heavier tail behavior than the log linear rate of decrease that characterizes the hyperbolic shape (see [2]). The NIG distribution has semiheavy tails (see [3])

$$f(x; \alpha, \beta, \delta) \sim \text{const.} |x|^{-3/2} \exp\{-\alpha|x| + \beta x\}, \quad x \rightarrow \pm\infty$$

The Normal Inverse Gaussian Process

We define the NIG process

$$X^{(\text{NIG})} = \{X_t^{(\text{NIG})}, t \geq 0\} \quad (3)$$

as the Lévy process with stationary and independent NIG-distributed increments, where $X_0^{(\text{NIG})} = 0$ with probability 1. To be precise, $X_t^{(\text{NIG})}$ follows a $NIG(\alpha, \beta, \delta t)$ law.

The Lévy measure of the NIG process is given by

$$\nu_{\text{NIG}}(dx) = \delta \alpha \pi^{-1} \exp\{\beta x\} K_1(\alpha|x|) (|x|)^{-1} dx \quad (4)$$

An NIG process has no Brownian component and its Lévy triplet is given by $[\gamma, 0, \nu_{\text{NIG}}(dx)]$, where

$$\gamma = 2\delta \alpha \pi^{-1} \int_0^1 \sinh(\beta x) K_1(\alpha x) dx \quad (5)$$

The NIG Lévy process may alternatively be represented *via* random time change of Brownian motion, using the inverse Gaussian (IG) process to determine time, as

$$X_t^{\text{NIG}} = \beta \delta^2 I_t + \delta W_{I_t} \quad (6)$$

where $W = \{W_t, t \geq 0\}$ is a standard Brownian motion and $I = \{I_t, t \geq 0\}$ is an IG process with parameters $a = 1$ and $b = \delta \sqrt{\alpha^2 - \beta^2}$.

2 Normal Inverse Gaussian Model

Table 1 Mean, variance, skewness, and kurtosis of the normal inverse Gaussian distribution

	$NIG(\alpha, \beta, \delta)$
Mean	$\delta\beta(\alpha^2 - \beta^2)^{-1/2}$
Variance	$\alpha^2\delta(\alpha^2 - \beta^2)^{-3/2}$
Skewness	$3\beta\alpha^{-1}\delta^{-1}(\alpha^2 - \beta^2)^{-1/4}$
Kurtosis	$3(1 + (\alpha^2 + 4\beta^2)\delta^{-1}\alpha^{-2}(\alpha^2 - \beta^2)^{-1/2})$

The NIG Model

The NIG model belongs to the class of exponential Lévy models (*see Time-changed Lévy Process*). Consider a market with a riskless asset (the bond), with a price process given by $b_t = \exp\{rt\}$, and one risky asset (the stock or index). The model for the risky asset is

$$S_t = S_0 \exp\{X_t^{(NIG)}\} \quad (7)$$

where the log returns $\log(S_{t+s}/S_t)$ follow the $NIG(\alpha, \beta, \delta s)$ distribution (i.e., the distribution of increments of length s of the NIG process).

Equivalent Martingale Measure

Pricing financial derivatives requires that we work under an equivalent martingale measure. We present here two ways to attain equivalent martingale measures for the discounted price process $\{\exp(-(r - q)t)S_t, t \geq 0\}$, where r is the risk-free continuously compounded interest rate and q is the continuously compounded dividend yield.

One can find at least one equivalent martingale measure Q using the Esscher transform (see [6]). For the NIG model, the Esscher transform equivalent martingale measure follows an $NIG(\alpha, \theta^* + \beta, \delta)$ law (see [12]), where θ^* is the solution of the equation

$$r - q = \delta \left(\sqrt{\alpha^2 - (\beta + \theta)^2} - \sqrt{\alpha^2 - (\beta + \theta + 1)^2} \right) \quad (8)$$

Another way to obtain an equivalent martingale measure Q is by mean-correcting the exponential

of the $NIG(\alpha, \beta, \delta)$ process, that is, introducing the distribution $NIG(\alpha, \beta, \delta, m)$ with the characteristic function

$$\tilde{\phi}(u; \alpha, \beta, \delta, m) = \phi(u; \alpha, \beta, \delta) \exp\{ium\} \quad (9)$$

where

$$m = r - q + \delta \left(\sqrt{\alpha^2 - (\beta + 1)^2} - \sqrt{\alpha^2 - \beta^2} \right) \quad (10)$$

and $\phi(u; \alpha, \beta, \delta)$ is defined by expression (1).

Pricing of European Options

Given our NIG market model, we focus now on the pricing of European options whose payoffs are functions of the terminal asset value only. Denote the payoff of the option at its time of expiry T by $G(S_T)$ and let $F(X_T) = G(S_T)$

Pricing through Density Function

For a European call option with strike price K and time to expiration T , the value V_0 at time 0 is given by the expectation of the payoff under the martingale measure Q . If we take for Q the Esscher transform equivalent martingale measure, the value at time 0 is given by

$$V_0 = \exp\{-qT\}S_0 \int_c^\infty f_T^{(\theta^*+1)}(x) dx - \exp\{-rT\}K \int_c^\infty f_T^{(\theta^*)}(x) dx \quad (11)$$

where $c = \ln(K/S_0)$, $f_T^{(\theta^*)}(x)$ is the density function of the $NIG(\alpha, \theta^* + \beta, \delta)$ distribution. Similar formulas can be derived for other derivatives with a payoff function that depends only on the terminal value at time $t = T$.

Pricing through the Lévy Characteristics

Another way to find the value $V_t = V(t, X_t)$ at time t is by solving a partial integro-differential

equation (see **Partial Integro-differential Equations (PIDEs)**). If $V(t, x) \in C^{1,2}$ then the function $V(t, x)$ solves

$$\begin{aligned} rV(t, x) = & \gamma \frac{\partial}{\partial x} V(t, x) + \frac{\partial}{\partial t} V(t, x) \\ & + \int_{-\infty}^{+\infty} \left(V(t, x+y) - V(t, x) \right. \\ & \left. - y \frac{\partial}{\partial x} V(t, x) \right) \nu^Q(dy) \\ V(T, x) = & F(x) \end{aligned} \quad (12)$$

where $[\gamma, 0, \nu^Q(dy)]$ is the Lévy triplet of the NIG process under the risk-neutral measure Q .

Pricing through the Characteristic Functions

Pricing can also be done by using the characteristic function [4] (see **Fourier Transform**). Let α be a positive constant such that the α th moment of the stock price exists, then the value of the option is given by

$$\begin{aligned} V_0 = & \frac{\exp\{-\alpha \log(K)\}}{\pi} \\ & \times \int_0^{+\infty} \exp\{-iv \log(K)\} \varrho(v) dv \end{aligned} \quad (13)$$

where

$$\begin{aligned} \varrho(v) = & \frac{\exp\{-rT\} E \left[\exp\{i(v - (\alpha + 1)i) \log(S_T)\} \right]}{\alpha^2 + \alpha - v^2 + i(2\alpha + 1)v} \\ = & \frac{\exp\{-rT\} \varphi(v - (\alpha + 1)i)}{\alpha^2 + \alpha - v^2 + i(2\alpha + 1)v} \end{aligned} \quad (14)$$

and

$$\varphi(u) = E \left[\exp\{iu \log(S_T)\} \right] \quad (15)$$

Other methods for the valuation of European options by applying characteristic functions can be found in [7] and [8].

Monte Carlo Simulations

One can make use of the representation (6) to simulate an NIG process. In such a way, a sample path of the NIG is obtained by sampling a standard Brownian motion and an IG process. We refer to [5] for the details on the generation of an IG random number.

Origin

The NIG distribution was introduced in [1]. The potential applicability of the NIG distribution and Lévy process for the modeling and analysis of statistical data from turbulence and finance is discussed in [2] and [3]. See also [9–11] for the application of the NIG distribution in modeling logarithmic asset returns.

References

- [1] Barndorff-Nielsen, O.E. (1995). *Normal Inverse Gaussian Distributions and the Modelling of Stock Returns*, Research Report No. 300, Department of Theoretical Statistics, Aarhus University.
- [2] Barndorff-Nielsen, O.E. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling, *Scandinavian Journal of Statistics* **24**(1), 1–13.
- [3] Barndorff-Nielsen, O.E. (1998). Processes of normal inverse Gaussian type, *Finance and Stochastics* **2**, 41–68.
- [4] Carr, P. & Madan, D. (1998). Option valuation using the fast Fourier transform, *Journal of Computational Finance* **2**, 61–73.
- [5] Devroye, L. (1986). *Non-Uniform Random Variate Generation*, Springer.
- [6] Gerber, H.U. & Shiu, E.S.W. (1994). Option pricing by Esscher-transforms, *Transactions of the Society of Actuaries* **46**, 99–191.
- [7] Lee, R.W. (2004). Option pricing by transform methods: extensions, unification, and error control, *Journal of Computational Finance* **7**(3), 50–86.
- [8] Raible, S. (2000). *Lévy Processes in Finance: Theory, Numerics, and Empirical Facts*. PhD thesis, University of Freiburg, Freiburg.
- [9] Rydberg, T. (1996). *The Normal Inverse Gaussian Lévy Process: Simulations and Approximation*, Research Report No. 344, Department of Theoretical Statistics, Aarhus University.
- [10] Rydberg, T. (1996). *Generalized Hyperbolic Diffusions with Applications Towards Finance*. Research Report

- No. 342, Department of Theoretical Statistics, Aarhus University.
- [11] Rydberg, T. (1997). A note on the existence of unique equivalent martingale measures in a Markovian setting, *Finance and Stochastics* **1**, 251–257.
- [12] Schoutens, W. (2003). *Lévy Processes in Finance—Pricing Financial Derivatives*, John Wiley & Sons, Chichester.

Related Articles

Exponential Lévy Models; Fourier Transform; Partial Integro-differential Equations (PIDEs).

HENRIK JÖNSSON, VIKTORIYA MASOL &
WIM SCHOUTENS

Generalized Hyperbolic Models

Generalized hyperbolic (GH) Lévy motions constitute a subclass of Lévy processes that are generated by *GH distributions*. GH distributions were introduced in Barndorff-Nielsen [1] in connection with a project with geologists. The Lebesgue density of this five-parameter class can be given in the following form:

$$\begin{aligned} d_{\text{GH}(\lambda, \alpha, \beta, \delta, \mu)}(x) &= a(\lambda, \alpha, \beta, \delta, \mu) \\ &\times (\delta^2 + (x - \mu)^2)^{\left(\lambda - \frac{1}{2}\right)/2} \\ &\times K_{\lambda - \frac{1}{2}}\left(\alpha \sqrt{\delta^2 + (x - \mu)^2}\right) \\ &\times \exp(\beta(x - \mu)) \end{aligned} \quad (1)$$

with the norming constant

$$a(\lambda, \alpha, \beta, \delta, \mu) = \frac{(\alpha^2 - \beta^2)^{\lambda/2}}{\sqrt{2\pi} \alpha^{\lambda - \frac{1}{2}} \delta^\lambda K_\lambda(\delta \sqrt{\alpha^2 - \beta^2})} \quad (2)$$

K_ν denotes the modified Bessel function of the third kind with index ν . The parameters can be interpreted as follows: $\alpha > 0$ determines the shape, β with $0 \leq |\beta| < \alpha$ the skewness and $\mu \in \mathbb{R}$ the location. $\delta > 0$ serves for scaling, and $\lambda \in \mathbb{R}$ characterizes subclasses. It is essentially the weight in the tails that changes with λ . There are two alternative parameterizations that are scale- and location-invariant, that is, they do not change under affine transformations $Y = aX + b$ for $a \neq 0$, namely, $\zeta = \delta \sqrt{\alpha^2 - \beta^2}$, $\rho = \beta/\alpha$ and $\xi = (1 + \zeta)^{-1/2}$, $\chi = \xi\rho$. Since $0 \leq |\chi| < \xi < 1$, for a fixed λ the distributions parameterized by χ and ξ can be represented by the points of a triangle, the so-called shape triangle.

GH distributions arise in a natural way as variance–mean mixtures of normal distributions. Let d_{GIG} denote the density of a generalized inverse Gaussian distribution (*see Normal Inverse Gaussian Model*) with parameters $\delta > 0$, $\gamma > 0$, and $\lambda \in \mathbb{R}$,

that is,

$$\begin{aligned} d_{\text{GIG}(\lambda, \delta, \gamma)}(x) &= \left(\frac{\gamma}{\delta}\right)^\lambda \frac{1}{2K_\lambda(\delta\gamma)} x^{\lambda-1} \\ &\times \exp\left(-\frac{1}{2}\left(\frac{\delta^2}{x} + \gamma^2 x\right)\right) \mathbb{1}_{\{x>0\}} \end{aligned} \quad (3)$$

Then if $N(\mu + \beta y, y)$ denotes a normal distribution with mean $\mu + \beta y$ and variance y , one can easily verify that

$$\begin{aligned} d_{\text{GH}(\lambda, \alpha, \beta, \delta, \mu)}(x) &= \int_0^\infty d_{N(\mu + \beta y, y)}(x) \\ &\times d_{\text{GIG}(\lambda, \delta, \sqrt{\alpha^2 - \beta^2})}(y) dy \end{aligned} \quad (4)$$

Using maximum likelihood estimation, one can fit GH distributions to empirical return distributions from financial time series such as the daily stock or index prices. Figure 1 shows a fit to the daily closing prices of Telekom over a period of seven years.

Figure 2 shows the same densities on a log scale in order to make the fit in the tails visible. One recognizes the hyperbolic shape of the GH density in comparison to the parabolic shape of the normal density. The characteristic function of the GH distribution is

$$\begin{aligned} \varphi_{\text{GH}}(u) &= e^{iu\mu} \left(\frac{\alpha^2 - \beta^2}{\alpha^2 - (\beta + iu)^2} \right)^{\frac{\lambda}{2}} \\ &\times \frac{K_\lambda\left(\delta \sqrt{\alpha^2 - (\beta + iu)^2}\right)}{K_\lambda(\delta \sqrt{\alpha^2 - \beta^2})} \end{aligned} \quad (5)$$

and expectation and variance are

$$E[GH] = \mu + \frac{\beta \delta^2}{\zeta} \frac{K_{\lambda+1}(\zeta)}{K_\lambda(\zeta)} \quad (6)$$

$$\begin{aligned} \text{Var}(GH) &= \frac{\delta^2}{\zeta} \frac{K_{\lambda+1}(\zeta)}{K_\lambda(\zeta)} + \frac{\beta^2 \delta^4}{\zeta^2} \\ &\times \left(\frac{K_{\lambda+2}(\zeta)}{K_\lambda(\zeta)} - \frac{K_{\lambda+1}^2(\zeta)}{K_\lambda^2(\zeta)} \right) \end{aligned} \quad (7)$$

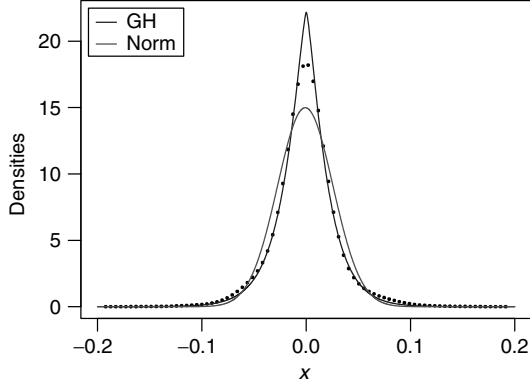


Figure 1 GH and normal density fitted to the daily Telekom returns

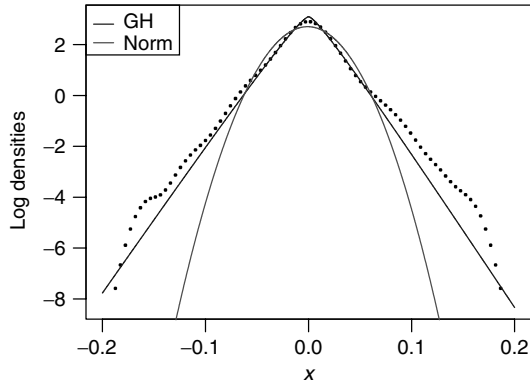


Figure 2 Fitted densities on a log scale

The moment-generating function exists for all u such that $-\alpha - \beta < u < \alpha - \beta$. Therefore, moments of all orders are finite.

There are two important subclasses. For $\lambda = 1$, one gets the class of *hyperbolic* distributions with density

$$d_{H(\alpha, \beta, \delta, \mu)}(x) = \frac{\sqrt{\alpha^2 - \beta^2}}{2\alpha\delta K_1(\delta\sqrt{\alpha^2 - \beta^2})} \times \exp\left(-\alpha\sqrt{\delta^2 + (x - \mu)^2} + \beta(x - \mu)\right) \quad (8)$$

whereas for $\lambda = -\frac{1}{2}$ one gets the class of normal inverse Gaussian (NIG) distributions with density

$$d_{\text{NIG}(\alpha, \beta, \delta, \mu)}(x) = \frac{\alpha}{\pi} \exp\left(\delta\sqrt{\alpha^2 - \beta^2} + \beta(x - \mu)\right) \times \frac{K_1\left(\alpha\delta\sqrt{1 + \left(\frac{x - \mu}{\delta}\right)^2}\right)}{\sqrt{1 + \left(\frac{x - \mu}{\delta}\right)^2}} \quad (9)$$

The latter one has a particularly simple characteristic function:

$$\varphi_{\text{NIG}}(u) = e^{iu\mu} \frac{\exp\left(\delta\sqrt{\alpha^2 - \beta^2}\right)}{\exp\left(\delta\sqrt{\alpha^2 - (\beta + iu)^2}\right)} \quad (10)$$

Many well-known distributions are limit cases of the class of GH distributions. For $\lambda > 0$ and $\delta \rightarrow 0$, one gets a variance-gamma distribution; in the special case of $\lambda = 1$ the result is a skewed and shifted Laplace distribution. Other limit cases are the Cauchy and the Student- t distribution as well as the gamma, the reciprocal gamma, and the normal distributions [4].

Exponential Lévy Models

GH distributions are infinitely divisible and therefore generate a Lévy process $L = (L_t)_{t \geq 0}$ such that the distribution of L_1 , $\mathcal{L}(L_1)$, is the given GH distribution. Analyzing the characteristic function in the Lévy–Khintchine form, one sees that the Lévy measure has an explicit density. There is no Gaussian component. Consequently the generated Lévy process is a process with purely discontinuous paths. The paths have infinite activity, which means that there are infinitely many jumps in any finite time interval (see **Jump Processes; Exponential Lévy Models**).

As a model for asset prices such as stock prices, indices, or foreign exchange rates, we take the exponential of the Lévy process L

$$S_t = S_0 \exp L_t \quad (11)$$

For hyperbolic Lévy motions, this model was introduced in [6], NIG Lévy processes were considered in [2], and the extension to GH Lévy motions

appeared in [3, 8]. The log returns from this model taken along time intervals of length 1 are $L_t - L_{t-1}$ and therefore they have exactly the GH distribution that generates the Lévy process. It was shown in [7] that the model (9) is successful in producing empirically correct distributions on other time horizons as well. This time consistency property can, for example, be used to derive the correct VaR estimates on a two-week horizon according to the Basel II rules. Equation (9) can be expressed by the following stochastic differential equation:

$$dS_t = S_{t-} (dL_t + e^{\Delta L_t} - 1 - \Delta L_t) \quad (12)$$

The price of a European option with payoff $f(S_T)$ is

$$V = e^{-rT} \mathbb{E} [f(S_T)] \quad (13)$$

where r is the interest rate and expectation is taken with respect to a risk-neutral (martingale) measure. As shown in [5], there are many equivalent martingale measures due to the rich structure of the driving process L . The simplest choice is the so-called Esscher transform, which was used in [6]. For the process L to be again a GH Lévy motion under an equivalent martingale measure (see **Equivalent Martingale Measures**), the parameters δ and μ have to be kept fixed [9]. Since the density of the distribution of S_T can be derived via inversion of the characteristic function, the expectation in equation (11) can be computed directly. A numerically much more efficient method based on two-sided Laplace transforms, which is applicable to a wide variety of options, has been developed in [9]. Assume that $e^{-Rx} f(e^{-x})$ is bounded and integrable for some R such that the moment-generating function of L_T is finite at $-R$. Write $g(x) = f(e^{-x})$ and $\psi_g(z) = \int_{\mathbb{R}} e^{-zx} g(x) dx$ for the bilateral Laplace transform of g . If $\zeta := -\log S_0$, then the option price V can be expressed in the following form:

$$V(\zeta) = \frac{e^{\zeta R - rT}}{2\pi} \int_{\mathbb{R}} e^{iu\zeta} \psi_g(R + iu) \varphi_{L_T}(iR - u) du \quad (14)$$

whenever the integral exists. φ_{L_T} denotes the characteristic function of the distribution of L_T .

References

- [1] Barndorff-Nielsen, O.E. (1977). Exponentially decreasing distributions for the logarithm of particle size, *Proceedings of the Royal Society of London A* **353**, 401–419.
- [2] Barndorff-Nielsen, O.E. (1998). Processes of normal inverse Gaussian type, *Finance and Stochastics* **2**(1), 41–68.
- [3] Eberlein, E. (2001). Application of generalized hyperbolic Lévy motions to finance, in *Lévy Processes. Theory and Applications*, O.E. Barndorff-Nielsen, T. Mikosch & S. Resnick, eds, Birkhäuser, pp. 319–336.
- [4] Eberlein, E. & von Hammerstein, E.A. (2004). Generalized hyperbolic and inverse Gaussian distributions: limiting cases and approximation of processes, in *Seminar on Stochastic Analysis, Random Fields and Applications IV*, R.C. Dalang, M. Dozzi & F. Russo, eds, Progress in Probability, Birkhäuser, Vol. 58, 221–264.
- [5] Eberlein, E. & Jacod, J. (1997). On the range of options prices, *Finance and Stochastics* **1**, 131–140.
- [6] Eberlein, E. & Keller, U. (1995). Hyperbolic distributions in finance, *Bernoulli* **1**(3), 281–299.
- [7] Eberlein, E. & Özkan, F. (2003). Time consistency of Lévy models, *Quantitative Finance* **3**, 40–50.
- [8] Eberlein, E. & Prause, K. (2002). The generalized hyperbolic model: financial derivatives and risk measures, in *Mathematical Finance – Bachelier Congress, 2000*, H. Geman, D. Madan, S. Pliska & T. Vorst, eds, Springer, Paris, pp. 245–267.
- [9] Raible, S. (2000). *Lévy Processes in Finance: Theory, Numerics, and Empirical Facts*. Ph.D. thesis, University of Freiburg.

Related Articles

Exponential Lévy Models; Fourier Methods in Options Pricing; Heavy Tails; Implied Volatility Surface; Jump-diffusion Models; Normal Inverse Gaussian Model; Partial Integro-differential Equations (PIDEs); Stochastic Exponential; Stylized Properties of Asset Returns.

ERNST EBERLEIN

Regime-switching Models

Many financial time series exhibit sudden changes in the structure of the data-generating process. Examples include financial crises, exchange rate swings, and jumps in the volatility. Sometimes, this sudden switch is due to a change in policy, for example, when moving from a fixed to a floating exchange rate regime. In other cases, the behavior of the series is influenced by an exogenous fundamental variable, such as the current position on the business or the credit cycle.

Regime-switching models attempt to capture this behavior by allowing the data-generating process to change in time, depending on an underlying, discrete but unobserved state variable. Typically, the functional form of the data-generating process remains the same across the different regimes with only the parameter values being state-dependent, as, for example, in a random walk equity return model where the drift and volatility change with the regime. However, it is feasible to set up models where the data-generating process itself changes, for example, moving from a deterministic fixed exchange to a stochastic floating one.

From a statistical point of view, regime-switching models will produce mixtures of distributions (*see Mixture of Distribution Hypothesis*), offering a very stylized and intuitive way of accommodating features such as fat tails, skewness, and volatility clustering (*see Stylized Properties of Asset Returns*). It is very easy to calibrate regime-switching models on historical data using maximum likelihood techniques, implementing what can be thought of as a discrete version of the Kalman filter (*see Filtering*). Virtually all economic and financial time series have been analyzed under the regime-switching framework, including interest and exchange rates, equity returns, commodity prices, energy prices, and credit spreads.

Derivative prices can be computed for regime-switching models by using transform methods (*see Hazard Rate*). The characteristic function of a regime-switching process can be computed in closed form if the characteristic functions conditional on each regime are available. This makes such processes a viable alternative to the stochastic volatility models, where one has to also resort to transform methods for pricing.

The Regime-switching Framework

A regime-switching model can be cast in either a discrete or continuous time setting. The model is built conditional to a Markov chain $s(t)$, the realization of which is not directly observed by economic agents. The chain can take a discrete set of values, and here we label them as $s(t) \in \{1, 2, \dots, N\} = \mathcal{S}$. The Markov chain is determined by its transition probability matrix in discrete time or its rate matrix in continuous time. In particular, in a discrete time setting, we write the transition probabilities $p_{i,j}$

$$P[S(t+1) = j | S(t) = i] = p_{i,j} \quad (1)$$

and we collect the elements $\{p_{i,j}\}$ in the transition probability matrix $\mathbf{P}$. The columns of $\mathbf{P}$ must sum up to 1, and all transition probabilities must be nonnegative. For a continuous time process, we define the transition rates $q_{i,j}$:

$$P[S(t+dt) = j | S(t) = i] = \mathbf{1}(i = j) + q_{i,j}dt \quad (2)$$

where $\mathbf{1}$ is the indicator function. The elements $\{q_{i,j}\}$ are collected in the rate matrix $\mathbf{Q}$. Note that by this definition, the columns of the rate matrix must sum up to 0, and the diagonal elements will be negative. The infinitesimal transition probability matrix is given as

$$\mathbf{P}(dt) = \mathbf{I} + \mathbf{Q}dt \quad (3)$$

where $\mathbf{I}$ is the unit ($N \times N$) matrix.

At each point in time, the data-generating process will vary, according to the regime $s(t)$ that prevails at that time. Thus, for a discrete time process, we can write

$$x(t) = g[t, s(t), \mathbf{y}(t), \epsilon(t); \boldsymbol{\beta}] \quad (4)$$

In the above expression, $\mathbf{y}(t)$ includes variables known at time t , including exogenous variables and lagged values of $x(t)$, and $\epsilon(t)$ represents the error term. In continuous time, we can write the stochastic differential equation:

$$dx(t) = \mu[t, s(t), \mathbf{y}(t); \boldsymbol{\beta}]dt + \sigma[t, s(t), \mathbf{y}(t); \boldsymbol{\beta}]dB(t) \quad (5)$$

Again, $\mathbf{y}(t)$ can include exogenous variables and the history of $x(t)$, while now $B(t)$ is a standard Brownian motion. The above equation can be extended to a multidimensional setting and can be generalized to include jumps or Lévy processes.

A standard simple example is a regime-dependent random walk process, where

$$\begin{aligned}\Delta x(t) &= \mu[s(t)] + \sigma[s(t)]\epsilon(t) \text{ in discrete time} \\ dx(t) &= \mu[s(t)]dt + \sigma[s(t)]dB(t) \\ &\text{in continuous time}\end{aligned}\quad (6)$$

The parameter set in this example is $\beta = \{\mu(1), \sigma(1), \mu(2), \sigma(2), \dots, \mu(N), \sigma(N)\}$.

Estimation from Historical Data

Given a set of historical values, the parameter vector β can be calibrated using maximum likelihood. As data are available over a discrete time grid, we focus on the calibration of the discrete time model. We give the switching random walk model with Gaussian noise as an example, but it should be straightforward to generalize this to more complex structures.

We denote the conditional density of $x(t)$ by $f_t(x|j) = P[x(t) \in dx|s(t) = j]$, given that the underlying Markov chain is in state j . In our example, this is a Gaussian density, with mean $\mu(j)$ and variance $\sigma^2(j)$. In addition, for future reference, we define the vector of conditional probabilities $\xi(t|t')$, with elements

$$\xi(t|t') = \{P[S(t) = j|\mathcal{F}(t')]\}_{j \in \mathcal{S}} \quad (7)$$

An important component of the calibration procedure is the vector of filtered probabilities $\xi_{t|t}$.

Given the parameter set β , the filtered probabilities can be computed based on the following recursion:

- Assume that the filtered probabilities are available up to time t .
- Compute the forecast probabilities $\xi(t+1|t) = \mathbf{P}\xi(t|t)$.
- The Bayes theorem yields the filtered probability:

- The numerator of the above expression is the product of the conditional density with the forecast probability for each state.
- The denominator, which is the sum of all numerators computed in the previous step, is also the conditional density of $P[x(t+1) \in dx|\mathcal{F}(t)]$. This is the likelihood function of the observation $x(t+1)$.

In the Kalman filtering terminology, the above computation can be compactly written in two steps:

$$\text{Prediction: } \xi(t+1|t) = \mathbf{P}\xi(t|t) \quad (9)$$

$$\begin{aligned}\text{Correction: } \xi(t+1|t+1) &= \frac{\mathbf{f}(t+1) \odot \xi(t+1|t)}{\mathbf{1}' \cdot (\mathbf{f}(t+1) \odot \xi(t+1|t))}\end{aligned}\quad (10)$$

The vector $\mathbf{f}(t)$ above collects the conditional distributions, across all possible regimes. In the Gaussian regime-switching model, the elements of $\mathbf{f}(t)$ would have the following elements:

$$\mathbf{f}(t) = \left\{ \frac{1}{\sqrt{2\pi}\sigma(j)} \exp\left(-\frac{(x(t) - \mu(j))^2}{2\sigma^2(j)}\right) \right\}_{j \in \mathcal{S}} \quad (11)$$

The symbol “ $\odot$ ” denotes element-by-element multiplication, and $\mathbf{1}$ is an $(N \times 1)$ vector of ones.

More details on estimation methods can be found in [18]. A small sample of the vast number of empirical applications that utilize the regime-switching framework is provided in [3, 15, 17, 19, 23, 27]. Generalizations include time-varying transition probabilities that depend on explanatory variables [14], switching generalized autoregressive conditionally heteroskedastic (GARCH)-type processes [16, 20], and models that resemble a multifractal setting [6]. A Bayesian alternative to maximum likelihood is sought in [1].

$$\begin{aligned}P[s(t+1) = j|\mathcal{F}(t+1)] &= P[s(t+1) = j|x(t+1), \mathcal{F}(t)] \\ &= P[x(t+1) \in dx|s(t+1) = j, \mathcal{F}(t)] \cdot \frac{P[s(t+1) = j|\mathcal{F}(t)]}{P[x(t+1) \in dx|\mathcal{F}(t)]} \\ &= \frac{P[x(t+1) \in dx|s(t+1) = j, \mathcal{F}(t)] \cdot P[s(t+1) = j|\mathcal{F}(t)]}{\sum_{\ell \in \mathcal{S}} P[x(t+1) \in dx|s(t+1) = \ell, \mathcal{F}(t)] \cdot P[s(t+1) = \ell|\mathcal{F}(t)]}\end{aligned}\quad (8)$$

Derivative Pricing under Regime Switching

Derivative pricing is typically carried out in a continuous time setting.^a For a vanilla payoff with maturity T , say $z(T) = h(x(T))$, the time-zero price is given by the risk neutral expectation:

$$z(0) = E^Q[D(T)z(T)] \quad (12)$$

where $D(t)$ is the discount factor.

In the regime-switching framework, pricing is routinely carried out using the Fourier inversion techniques (see **Fourier Transform**) outlined in [7]. In particular, if the log asset price $x(T)$ follows a regime-switching Brownian motion

$$dx(t) = \mu(s(t))dt + \sigma(s(t))dB(t) \quad (13)$$

then the characteristic function $\phi(u; T) = E \exp(iu x(T))$ is given by the matrix exponential

$$\phi(u; T) = e^{T \mathbf{A}(u)} \boldsymbol{\xi}(0|0) \quad (14)$$

where the $(N \times N)$ matrix $\mathbf{A}(u)$ has the following form:

$$a_{i,j}(u) = \begin{cases} q_{i,i} + g(u; i) & \text{if } i = j \\ q_{i,j} & \text{if } i \neq j \end{cases} \quad (15)$$

for $g(u; i) = iu\mu(i) - \frac{1}{2}u^2\sigma^2(i)$.

The first implementation that prices options where a two-regime process is present is that in [26]. In a more general setting with N regimes, vanilla call option prices can be easily retrieved using the Fast Fourier Transform (FFT) approach of Carr and Madan [7] or the fractional variant that allows explicit control of the discretization grids [10].

The above prototypical process can be extended in two directions. Rather than having switching Brownian motions that generate the conditional paths, one can consider switching Lévy processes (see **Lévy Processes; Exponential Lévy Models**) between the regimes (see [24], for the special two-regime case, and [11], for a more general setting). In that case, the function $g(u; i)$ in $\mathbf{A}(u)$ is replaced by the characteristic exponent of the Lévy process that is active in the i th regime. In addition, to introduce a correlation structure between the regime changes and the log-price changes, a jump in the log-price is introduced when the Markov chain switches. In that

case, the off-diagonal elements of $\mathbf{A}(u)$ are multiplied by the characteristic function of the jump size.

Pricing of American options can be done by setting up the continuation region [5] or by employing a variant of Carr's randomization procedure [4]. More exotic products can be handled by setting up a system of partial (integro-)differential equations (see **Partial Integro-differential Equations (PIDEs)**) or by explicitly using Fourier methods as in [22]. As the conditional distribution can be recovered from the characteristic function numerically, the density-based approach of [2] (see **Quadrature Methods**) can be a viable alternative.

Regime Switching as an Approximation

Rather than serving as the fundamental latent process, the Markov chain can serve as an approximation to more complex jump-diffusive dynamics. Then, one can use the regime-switching framework to tackle problem in a nonaffine (see **Affine Models**) setting, both in terms of calibration and derivative pricing. To achieve that, the number of regimes must be large, but the transition rates and conditional dynamics will be functions of a small number of parameters. The book by Kushner and Dupuis [25] outlines the convergence conditions for the approximation of generic diffusions and shows how one can implement the Markov chain approximation in practice.

Following this approach, many stochastic volatility problems can be cast as regime-switching ones. Chourdakis [8] shows how a generic stochastic volatility process can be approximated in that way, whereas Chourdakis [9] extends this method to produce the counterpart of the [21] stochastic volatility model (see **Heston Model**) in a regime-switching framework where the equity is driven by a Lévy noise.

End Notes

^aThe treatment in References [12, 13] are exceptions to this.

References

- [1] Albert, J. & Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts, *Journal of Business and Economic Statistics* **11**, 1–15.

- [2] Andricopoulos, A.D., Widdicks, M., Duck, P.W. & Newton, D.P. (2003). Universal option valuation using quadrature methods, *Journal of Financial Economics* **67**, 447–471.
- [3] Ang, A. & Bekaert, G. (2002). Regime switches in interest rates, *Journal of Business and Economic Statistics* **20**(2), 163–182.
- [4] Boyarchenko, S.I. & Levendorski, S.Z. (2006). *American Options in Regime-switching Models*. Manuscript available online at SSRN: 929215.
- [5] Buffington, J. & Elliott, R.J. (2002). American options with regime switching, *International Journal of Theoretical and Applied Finance* **5**, 497–514.
- [6] Calvet, L. & Fisher, A. (2004). How to forecast long-run volatility: regime-switching and the estimation of multifractal processes, *Journal of Financial Econometrics* **2**, 49–83.
- [7] Carr, P. & Madan, D. (1999). Option valuation using the Fast Fourier Transform, *Journal of Computational Finance* **3**, 463–520.
- [8] Chourdakis, K. (2004). Non-affine option pricing, *Journal of Derivatives* **11**(3), 10–25.
- [9] Chourdakis, K. (2005). Lévy processes driven by stochastic volatility, *Asia-Pacific Financial Markets* **12**, 333–352.
- [10] Chourdakis, K. (2005b). Option pricing using the Fractional FFT, *Journal of Computational Finance* **8**(2), 1–18.
- [11] Chourdakis, K. (2005c). *Switching Levy Models in Continuous Time: Finite Distributions and Option Pricing*. Manuscript available online at SSRN: 838924.
- [12] Chourdakis, K. & Tzavalis, E. (2000). *Option Pricing Under Discrete Shifts in Stock Returns*. Manuscript available online at SSRN: 252307.
- [13] Duan, J.-C., Popova, I. & Ritchken, P. (1999). *Option Pricing under Regime Switching*. Technical report, Hong Kong University of Science and Technology.
- [14] Filardo, A.J. (1994). Business-cycle phases and their transitional dynamics, *Journal of Business and Economic Statistics* **12**, 299–308.
- [15] Garcia, R., Luger, R. & Renault, E. (2003). Empirical assessment of an intertemporal option pricing model with latent variables, *Journal of Econometrics* **116**, 49–83.
- [16] Gray, S. (1996). Modeling the conditional distribution of interest rates as a regime-switching process, *Journal of Financial Economics* **42**, 27–62.
- [17] Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica* **57**, 357–384.
- [18] Hamilton, J.D. (1994). *Time Series Analysis*, Princeton University Press, Princeton, NJ.
- [19] Hamilton, J.D. (2005). What’s real about the business cycle? *Federal Reserve Bank of St. Louis Review* **87**(4), 435–452.
- [20] Hamilton, J.D. & Susmel, R. (1994). Autoregressive conditional heteroscedasticity and changes in regime, *Journal of Econometrics* **64**, 307–333.
- [21] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**, 327–344.
- [22] Jackson, K.R., Jaimungal, S. & Surkov, V. (2007). *Fourier Space Time-stepping for Option Pricing with Lévy Models*. Manuscript available online at SSRN: 1020209.
- [23] Jeanne, O. & Masson, P. (2000). Currency crises, sunspots, and Markov-switching regimes, *Journal of International Economics* **50**, 327–350.
- [24] Konikov, M. & Madan, D. (2002). Option pricing using Variance Gamma Markov chains, *Review of Derivatives Research* **5**(1), 81–115.
- [25] Kushner, H.J. & Dupuis, P.G. (2001). *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd Edition, Applications of Mathematics, Springer Verlag, New York, NY, Vol. 24.
- [26] Naik, V. (1993). Option valuation and hedging strategies with jumps in the volatility of asset returns, *The Journal of Finance* **48**, 1969–1984.
- [27] Weron, R., Bierbauer, M. & Trück, S. (2004). Modeling electricity prices: jump diffusion and regime switching, *Physica A* **336**, 39–48.

Related Articles

Exponential Lévy Models; Filtering; Fourier Methods in Options Pricing; Fourier Transform; Monte Carlo Simulation; Stochastic Volatility Models; Stylized Properties of Asset Returns; Variance-gamma Model.

KYRIAKOS CHOURDAKIS

Variance-gamma Model

The variance-gamma (VG) process is a stochastic process with independent stationary increments, which allows for flexible parameterization of skewness and kurtosis of increments. It has gained popularity, especially in option pricing, because of its analytical tractability. It is an example of a pure-jump Lévy process (*see Lévy Processes*).

The VG model is derived from the (symmetric) VG probability distribution, which is so named because it is the distribution of a random variable X that results from mixing a normal variable on its variance by a gamma distribution. Specifically, the conditional distribution of X is given by $X|W \sim \mathcal{N}(\mu, \sigma^2 W)$, $\mu \in \mathbb{R}$, $\sigma > 0$, where $W \sim \Gamma(\alpha, \alpha)$, $\alpha > 0$. The symbol “ $\sim$ ” stands for “is distributed as”. The symbol $\Gamma(\alpha, \lambda)$ indicates a gamma probability distribution, with probability density function (PDF), for $\alpha, \lambda > 0$,

$$f_{\Gamma}(w; \alpha, \lambda) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} w^{\alpha-1} e^{-\lambda w}, \quad w > 0; = 0 \text{ elsewhere} \quad (1)$$

The choice $\lambda = \alpha$ implies that $\mathbb{E}(W) = 1$ and $\mathbb{V}\text{ar}(W) = \frac{1}{\alpha}$. Thus, X is a random variable symmetrically distributed about its mean $\mathbb{E}(X) = \mu$, with $\mathbb{V}\text{ar}(X) = \sigma^2$ and a simple characteristic function (CF), so that when X is mean-corrected to $Y = X - \mu$, the CF is

$$\mathbb{E}(e^{iuY}) = \left(1 + \frac{\sigma^2 u^2}{2\alpha}\right)^{-\alpha} \quad (2)$$

A random variable X having (symmetric) VG distribution may also be viewed as

$$X \stackrel{\mathcal{D}}{=} \mu + \sigma W^{\frac{1}{2}} Z, \quad \mathbb{E}(W) = 1 \quad (3)$$

where the symbol $\stackrel{\mathcal{D}}{=}$ means “has the same distribution as”. Here $Z \sim \mathcal{N}(0, 1)$ and W is a positive nondegenerate random variable distributed independently of Z . In the case of the VG distribution, $W \sim \Gamma(\alpha, \alpha)$.

Log returns of financial assets

$$X_t = \log P_t - \log P_{t-1}, \quad t = 1, 2, \dots, N \quad (4)$$

reveal a kurtosis well in excess of 3, suggesting that the modeling of returns should be done by a symmetric distribution with heavier tails than the normal (*see Stylized Properties of Asset Returns; Heavy Tails*). For example, in 1972, Praetz [20] argued in favor of variance-dilation of the normal through variance-mixing and found that mixing according to $X|W \sim \mathcal{N}(\mu, \sigma^2 W)$, where W has reciprocal (inverse) gamma PDF with $\mathbb{E}W = 1$, gives the scaled t -distribution symmetric about μ for the returns. This is a slight generalization of the classical Student’s t -distribution, in that fractional degrees of freedom are permitted.

Influenced by Praetz’s work, Madan and Seneta [16] took the distribution of the mixing variable W itself to be gamma (rather than reciprocal gamma). This resulted in a continuous-time model, which is now known as the (*symmetric*) *variance-gamma model*.

The VG model may be placed within the context of a more general subordinator (*see Time Change*) model [10], which gives the price P_t of a risky asset over continuous time $t \geq 0$ as

$$P_t = P_0 \exp \{ \mu t + \sigma B(T_t) \} \quad (5)$$

where μ and $\sigma (> 0)$ are real constants. $\{T_t\}$, the (market) activity time, is a positive, increasing random process (with stationary differences $W_t = T_t - T_{t-1}$, $t = 1, 2, \dots$), which is independent of the standard Brownian motion $\{B(t)\}$. The corresponding returns are then given as

$$X_t = \log P_t - \log P_{t-1} = \mu + \sigma (B(T_t) - B(T_{t-1})) \quad (6)$$

We assume that $\mathbb{E}(W_t) < \infty$, and so without loss of generality that

$$\mathbb{E}(W_t) = 1 \quad (7)$$

to make the expected activity time change over unit calendar time equal to one unit, the scaling change in time being absorbed into σ , while noting that

$$X_t \stackrel{\mathcal{D}}{=} \mu + \sigma W_t^{\frac{1}{2}} B(1) \quad (8)$$

which is of form (3). The case $T_t = t$ of the model (5) is the classical geometric Brownian motion (GBM) model for the process $\{P_t\}$, with corresponding returns being independently $\mathcal{N}(\mu, \sigma^2)$ distributed.

In the VG model, $\{T_t\}$ for $t \geq 0$, is the gamma process, a process of stationary-independent increments. The distribution of an increment over a time interval of length t is $\Gamma(\alpha t, \alpha)$. It is a remarkable feature that the distributional form for any t is the same; this is inherited by the VG model for $\{\log(P_t/P_0)\}$, $t \geq 0$, which is a process of stationary-independent increments, with the distribution of an increment over any time period t having CF

$$e^{i\mu u t} \left(1 + \frac{\sigma^2 u^2}{2\alpha}\right)^{-\alpha t} \quad (9)$$

The corresponding distribution is also called a (*symmetric*) *VG distribution*. Its mean and variance are given by $\mathbb{E} \log(P_t/P_0) = \mu t$ and $\text{Var}(\log(P_t/P_0)) = \sigma^2 t$, respectively. The whole structure is redolent of Brownian motion, to which the VG process reduces in the limit as $\alpha \rightarrow \infty$.

An important consequence of the VG distributional form of an increment over any time interval of length t is that, irrespective of the size of unit of time between successive data readings, returns have a VG distribution.

The CF (2), clearly the CF of an infinitely divisible distribution, is also the CF of a difference of two independently and identically distributed (i.i.d.) gamma random variables, which reflects the fact shown in [16] that the process $\{\log(P_t/P_0) - t\mu\}$, $t \geq 0$, is the difference of two i.i.d. gamma processes.

The VG model is a pure-jump process [16] (*see Jump Processes*) reflecting this feature of a gamma process. This is seen from the Lévy–Khinchin representation (*see Lévy Processes*).

The analytical simplicity of the VG model and its pure-jump nature make it a leading candidate for modeling historical financial data. Further, the VG distribution's PDF has explicit structural form (see below), which is tractable for maximum-likelihood estimation of parameters from returns data.

Returns $\{X_t\}$, $t = 0, 1, 2, \dots$, considered in isolation, need not be taken to be i.i.d. as in the preceding discussion, but to form (more generally) a strictly stationary sequence, to which moment estimation methods, for example, will still apply [25]. The symmetric scaled t -distribution continues to enjoy favor as a model for the distribution of returns because of its power-law (Pareto-type) probability tails, a property manifested (in contrast to the VG) in the nonexistence of higher moments. For some data sets, empirical

investigations [10, 12] suggest the nonexistence of higher moments in a model for returns (*see Heavy Tails; Stylized Properties of Asset Returns*) and hence the scaled t -distribution.

On the other hand, other investigations [9] suggest that it is virtually impossible to distinguish between the symmetric scaled t and VG distributions in regard to distributional tail structure by taking compatible parameter values in the two distributions. In fact, the PDFs of the two distributions reveal that the concentration of probability near the point of symmetry μ and in the middle range of the distributions is qualitatively and quantitatively different. The VG distribution tends to increase the probability near μ and in the tails, at the expense of the middle range. The different natures in regard to shape are most significantly revealed by the Cauchy distribution as a special case of the t -distribution and the Laplace (two-sided exponential) distribution as a special case of the VG distribution.

The first monograph to include a study of the VG model was [4]. Since then it has found a place in monographs such as [22], where it is treated in the general context of Lévy processes (*see Lévy Processes*).

Both the VG distribution and the scaled t -distribution are extreme cases of the generalized hyperbolic (*see Generalized Hyperbolic Models*) distribution [23, 25].

Allowing for Skewness

A generalized normal mean–variance-mixing distribution is the distribution of X , where the conditional distribution of X is given as

$$X|W \sim \mathcal{N}(\mu + \theta W, \sigma^2 W + d^2) \quad (10)$$

Here, μ , θ , d , and $\sigma (> 0)$ are real numbers, and W is a nondegenerate positive random variable. The distribution is skew if $\theta \neq 0$, and it is symmetric otherwise. Press [21] studied a continuous-time model with this distribution for returns, where $W \sim \text{Poisson}(\lambda)$. This is a process of stationary-independent increments, resulting from adding a compound Poisson process of normal shocks to a Brownian motion, and has both continuous and jump components. Some special cases of equation (10) as a returns distribution, with focus on the estimation of parameters by the method of moments, are considered in [25].

A random variable X is said to have a normal variance–mean mixture (NVM) distribution [1] if equation (10) holds with $d = 0$.

The symmetric VG and scaled t -distributions are instances of equation (10), with $d = \theta = 0$.

The skew VG distribution, as introduced in [15], is the case of NVM where W is described by equation (1) with $\mathbb{E}W = 1$ as in the symmetric case. (The skewed scaled t -distribution is defined analogously by taking W to have a reciprocal gamma distribution.)

The skew VG distribution has PDF

$$f_{\text{VG}}(x) = \sqrt{\frac{2}{\pi}} \frac{\alpha^\alpha e^{\frac{(x-\mu)\theta}{\sigma^2}}}{\sigma \Gamma(\alpha)} \left(\frac{|x-\mu|}{\sqrt{\theta^2 + 2\alpha\sigma^2}} \right)^{\alpha-\frac{1}{2}} \times K_{\alpha-\frac{1}{2}} \left(\frac{|x-\mu|\sqrt{\theta^2 + 2\alpha\sigma^2}}{\sigma^2} \right), \quad x \in \mathbb{R} \quad (11)$$

and CF

$$\phi_{\text{VG}}(u) = \mathbb{E}(e^{iux}) = e^{i\mu u} \left(1 - \frac{1}{\alpha} \left(iu\theta - \frac{\sigma^2 u^2}{2} \right) \right)^{-\alpha} \quad (12)$$

$K_\eta(\omega)$ for $\eta \in \mathbb{R}$ and $\omega > 0$, given as

$$K_\eta(\omega) = \frac{1}{2} \int_0^\infty z^{\eta-1} e^{-\frac{\omega}{2} \left(z + \frac{1}{z} \right)} dz \quad (13)$$

is a modified Bessel function of the third kind with index η ($K_\eta(\omega)$ is referred to as a *modified Bessel function* of the *second* kind in some texts).

An equivalent representation is

$$X \stackrel{\mathcal{D}}{=} \mu + \theta W + \sigma W^{\frac{1}{2}} Z, \quad \mathbb{E}W = 1 \quad (14)$$

where Z and W are independently distributed, $Z \sim \mathcal{N}(0, 1)$, $W \sim \Gamma(\alpha, \alpha)$ as mentioned before. This distributional structure is consistent with the continuous-time model for prices

$$P_t = P_0 \exp \{ \mu t + \theta T_t + \sigma B(T_t) \} \quad (15)$$

where $\{T_t\}$, $t \geq 0$, is a gamma process, exactly as before. The process of independent stationary increments $\{\log(P_t/P_0)\}$, $t \geq 0$, with the distribution of returns described by equations (11)–(13), is also

called the *variance-gamma model*. Its properties are extensively studied in [15].

Dependence and Estimation

The VG model described above is a Lévy process (see **Lévy Processes**)—a stochastic process in continuous time with stationary independent increments—whose increments are independent and VG distributed. To discuss dependence, we consider the model for returns:

$$X_t = \log P_t - \log P_{t-1} = \mu + \theta W_t + \sigma W_t^{1/2} Z_t, \quad t = 1, 2, \dots \quad (16)$$

where Z_t , $t = 1, 2, \dots$, are identically distributed $\mathcal{N}(0, 1)$ random variables, independent of the strictly stationary process $\{W_t\}$, $t = 1, 2, \dots$. Here $\theta, \sigma (> 0)$ are constants as before.

When $\theta = 0$, this discrete-time model is equivalent in distribution to that described by the subordinator model of Heyde [10] given by equations (5)–(8). Note that $\text{Cov}(X_t, X_{t+k}) = 0$, while $\text{Cov}(X_t^2, X_{t+k}^2) \neq 0$, $k = 1, 2, \dots$. This is an important feature inasmuch as many asset returns display a sample autocorrelation function plot characteristic of white noise, but no longer do so in a sample autocorrelation plot of squared returns and of absolute values of returns [10, 12, 17].

McLeish [17] considered the distribution of individual $W_t \sim \Gamma(\alpha, \lambda)$, which gives the distribution of individual X_t as (symmetric) VG, which he regarded as a robust alternative to the normal. He suggested a number of ways of introducing the dependence in the process $\{W_t\}$, $t = 1, 2, \dots$.

The continuous-time subordinator model was expanded in [11] to allow for scaled t -distributed returns. Their specification of the activity time process in continuous time $\{T_t\}$ incorporated self-similarity (a scaling property) and long-range dependence (LRD) in the stochastic process of squared returns. (LRD in the Allen sense is expressed as divergence of the sequence of ultimately nonnegative autocorrelations of a discrete stationary process.)

The general form of the continuous-time model for prices over continuous time $t \geq 0$ as

$$P_t = P_0 \exp \{ \mu t + \theta T_t + \sigma B(T_t) \} \quad (17)$$

for which the returns are equivalent in distribution to equation (16) was given in [23] as a generalization of the subordinator model that allows for skewness in the distribution of returns in the same way as in [15], but the returns inherit the postulated strict stationarity of the sequence $\{W_t\}, t = 1, 2, \dots$. Following on from [11], Finlay and Seneta [5, 6] studied in detail and in parallel the continuous time structure of the skew VG model and the skew t -model, with focus on skewness, asymptotic self-similarity, and LRD.

Maximum-likelihood estimation for independent readings from a symmetric VG distribution is discussed in [17] and in [23], which however proposes moment estimation in the presence of dependence. Moment estimation, allowing for dependence, is further developed in [25], along with goodness of fit of various models for several sets of asset data. A method of simulating data from long-range dependent processes with skew VG or t -distributed increments is described in [7], and various estimation procedures (method of moments, product-density maximum likelihood, and nonstandard minimum χ^2 in particular) are tested on the data to assess their performance. In the simulations considered, the product-density maximum-likelihood method performs favorably. The conclusion, within the limited testing carried out, indicates then that, in practice, ordinary product density maximum-likelihood estimation is satisfactory even in the presence of LRD. This is tantamount to saying that one may treat such data on returns as i.i.d. This entails an enormous simplification in estimation procedures in fitting the skew VG and skew t -distributions.

Option Pricing Applications

Our discussion is based on the difference of gamma (DG) models for real-world (historical) data:

$$P_t = P_0 e^{\mu t + G_1(t; a, b) - G_2(t; c, d)} \quad (18)$$

where $\{G(t; \alpha, \beta)\}$ is a gamma process, and so $G(t; \alpha, \beta) \sim \Gamma(t\alpha, \beta)$ for any given t , and the two gamma processes are independent of each other. For each t , the returns (4) for P_t have the following CF:

$$\begin{aligned} \phi_{\text{DG}}(u; \mu, a, b, c, d) \\ = e^{i\mu u} \left(1 - \frac{i u}{b}\right)^{-a} \left(1 + \frac{i u}{d}\right)^{-c} \end{aligned} \quad (19)$$

Comparing equations (12) and (19), it is clear that choosing $c = a$ results in a (skew) VG distribution, with PDF (11) parameters $\alpha = a$, $\theta = a\left(\frac{1}{b} - \frac{1}{d}\right)$, and $\sigma^2 = \frac{2a}{bd}$. The further simplification $b = d$ results in the symmetric VG process for returns.

Using this model for option pricing requires imposing parameter restrictions to ensure that $\{e^{-rt} P_t\}$ is a martingale, where r is the interest rate. This amounts to ensuring that $\mathbb{E}(e^{-rt} P_t | \mathcal{F}_s) = e^{-rs} P_s$, where $\mathcal{F}_s$ represents information available to time $s \leq t$. In the case of the DG process,

$$\begin{aligned} \mathbb{E}(e^{-rt} P_t | \mathcal{F}_s) \\ = e^{-rs} P_s \times e^{(\mu-r)(t-s)} \left(\frac{b}{b-1}\right)^{a(t-s)} \left(\frac{d}{d+1}\right)^{c(t-s)} \end{aligned} \quad (20)$$

so that imposing the restriction

$$\mu = r - a \log\left(\frac{b}{b-1}\right) - c \log\left(\frac{d}{d+1}\right) \quad (21)$$

with $b > 1$ results in $\{e^{-rt} P_t\}$, which is a martingale with four free parameters: a , b , c , and d . We label it *MDG*.

The (skew) VG special case is obtained by choosing $c = a$. Relabeling the parameters as above, $\alpha = a$, $\theta = a\left(\frac{1}{b} - \frac{1}{d}\right)$, and $\sigma^2 = \frac{2a}{bd}$, results in a martingale that is a (skew) VG process. The mean constraint (21) now translates to

$$\mu = r + \alpha \log\left(1 - \frac{\theta + \frac{1}{2}\sigma^2}{\alpha}\right) \quad (22)$$

where we take $\alpha > \theta + \frac{1}{2}\sigma^2$. This martingale now has only three free parameters. We label it *MVG*. This corresponds to the labeling “VG” in [22] and is the martingale used in [15].

Both MDG and MVG are, in the terminology of the work by Schoutens [22], “mean-correcting martingales”, since the restriction (21), (22) is on the mean (μ) to produce a martingale.

Another way of producing a martingale from a (skew) VG process is to begin by noting from [5] and equation (17) that irrespective of the distribution of T_t ,

$$\begin{aligned} & \mathbb{E}(e^{-rt} P_t | \mathcal{F}_s) \\ &= e^{-rs} P_s \times e^{(\mu-r)(t-s)} \mathbb{E}\left(e^{(\theta + \frac{1}{2}\sigma^2)(T_t - T_s)} | \mathcal{F}_s\right) \end{aligned} \quad (23)$$

where the sequence $\{W_t\}$, where $W_t = T_t - T_{t-1}$, is strictly stationary.

Thus, if we take $\mu = r$ and $\theta = -\frac{1}{2}\sigma^2$ in (23), the right-hand side of equation (23) becomes $e^{-st} P_s$, and we have a martingale. This construction of a martingale, simple and quite general, is slightly restrictive, however, in that two parameters, μ and θ , are constrained. We shall refer to this construction as a *skew-correcting martingale*, since θ , the parameter that determines skewness, is constrained. We denote this martingale model by *MSK*. Out of the “external” parameters μ , θ , and σ , the only parameter that is retained is σ , which is called the *historical volatility* in the Black–Scholes (BS) context, which is a special case when $T_t = t$. Any additional parameters in the martingale (risk-neutral) process will be those emanating from the nature of $\{T_t\}$, which will need to be specified for any examination of estimation and goodness of fit.

When the CF of the risk-neutral distribution of price is of the closed form, option prices may be calculated using Fourier methods (*see Fourier Methods in Options Pricing*) as in [3, 14]. Specifically, for $C(\Upsilon, k)$, the price of a European call option with time to maturity Υ and strike price K and $k = \log(K)$, let $q_\Upsilon(p)$ be the risk-neutral density of $\log(P_\Upsilon)$, with CF $\phi_\Upsilon(u)$, at time Υ . Thus,

$$\begin{aligned} C(\Upsilon, k) &= e^{-r\Upsilon} \mathbb{E}(P_\Upsilon - K)^+ \\ &= \int_k^\infty e^{-r\Upsilon} (e^p - e^k) q_\Upsilon(p) dp \end{aligned} \quad (24)$$

Define the modified call price as

$$c(\Upsilon, k) = e^{\gamma k} C(\Upsilon, k) \quad (25)$$

for some γ such that $\mathbb{E}(P_\Upsilon^{\gamma+1}) < \infty$. The Fourier transform of $c(\Upsilon, k)$ is then given as

$$\begin{aligned} \Psi_\Upsilon(x) &= \int_{-\infty}^\infty e^{ikx} c(\Upsilon, k) dk \\ &= \frac{e^{-r\Upsilon} \phi_\Upsilon(x - (\gamma + 1)i)}{\gamma^2 + \gamma - x^2 + ix(2\gamma + 1)} \end{aligned} \quad (26)$$

Taking the inverse Fourier transform and using the fact that $c(\Upsilon, k)$ is real, and using equation (25) gives

$$\begin{aligned} C(\Upsilon, k) &= \frac{e^{-\gamma k}}{2\pi} \int_{-\infty}^\infty e^{-ixk} \Psi_\Upsilon(x) dx \\ &= \frac{e^{-\gamma k}}{\pi} \int_0^\infty \Re\{e^{-ixk} \Psi_\Upsilon(x)\} dx \end{aligned} \quad (27)$$

In fact, we shall use a modified version of equation (27) suggested in [14]:

$$C(\Upsilon, k) = R_\gamma + \frac{e^{-\gamma k}}{\pi} \int_0^\infty \Re\{e^{-ixk} \Psi_\Upsilon(x)\} dx \quad (28)$$

where $R_\gamma = \phi_\Upsilon(-i)$ for $-1 < \gamma < 0$. The choice of γ generally impacts on the error generated by the numerical approximation of equation (28). Finally, the option price (28) is computed *via* numerical integration.

The option price in this procedure is given simply by the sum of a number of function evaluations. Lee [14] shows that with a judicious choice of tuning parameters, one can calculate the option price up to 99.99% accuracy with less than 100, and in some cases less than 10, function evaluations. This CF-based pricing method lends itself easily to the fast Fourier transform, which allows for a very fast calculation of a range of option prices.

To numerically illustrate the method and the empirical performance of MVG against some competitors, we [8] use the data set in [22], Appendix C, which contains 77 call option prices on the S&P 500 index at the close of market on April 18, 2002. Fundamentally, each data point consists of the triple: strike, option price, and expiry date.

Fitting models involves estimating model parameters. To do this, we follow [22], p. 7, by minimizing with respect to the model parameters, the root-mean-square error (RMSE):

$$\begin{aligned} \text{RMSE} &= \sqrt{\left(\sum_{\text{options}} \frac{(\text{market price} - \text{model price})^2}{\text{number of options}} \right)} \end{aligned} \quad (29)$$

and then comparing the values of the minimized errors between models. If a model perfectly described

the asset price process, the RMSE value would be zero, with all model prices matching market prices, given the single true set of parameters.

The estimates of model parameters produced for a given model correspond to the current market status of that model. The procedure is thus, for a given model, a calibration procedure. No historical data are used in this procedure. The use of this data set for comparison of several different models in this way as already done in [22] allows for easy comparison of goodness of fit. We used the tuning parameter value $\gamma = -\frac{1}{2}$; the other nonmodel constants, q, r , were as in [22], Appendix C.

The RMSE surface for the MDG was reported in [8] to be quite flat, with a number of different parameter values giving essentially the same RMSE value of 2.24. The parameter values that gave the lowest value by 0.001 were as follows: $a = 4.35, b = 240.86, c = 9.79 \times 10^{-6}, d = 2.65 \times 10^{-7}$. The four-parameter MDG model thus did better than the four-parameter CGMY and GH models reported in [22], p. 83, and shown in Table 1.

The VG (skew) model fit reported in [22], p. 83, corresponded to the parameter values $\alpha (= a = c) = 5.4296 \times 10^{-3}, b = 14.2699, d = 5.8704$. The recalculation of RMSE with these parameter values reported in [8] gave the value 3.57. Optimization of RMSE reported in [8] with starting values $a = c = 0.01, b = d = 10$ resulted in the parameter estimates and RMSE as in [22]. Thus in Table 1, the RMSE values reported under VG and MVG are the same, 3.56.

As expected, this three-parameter model (MVG/VG) does not perform quite as well as the four-parameter models. The VG model is a special case of the GH model, so this is not unexpected.

Finlay and Seneta [8] discuss fitting an MSK martingale model, which allows for LRD in the historical data. This introduces two parameters in addition to the “historical volatility” parameter σ , namely, a parameter α corresponding to the gamma distribution with mean 1, as before, and a “Hurst” parameter H associated with dependence. The fit of MSK produces an RMSE of 6.35 and an estimate of $\sigma = 0.012$. There is almost no improvement on the BS situation reported in Table 1, which is the standard BS

martingale model; its RMSE and σ -estimate values are reported in [22], pp. 40-41, and are 6.73 and 0.011, respectively. This apparent insensitivity of the MSK model to departure from BS, possibly due to the skewness parameter being constrained in the martingale construction, is overcome, as reported in [8], by a four-parameter (“lack of static arbitrage” model: [2]) model, which is termed as C3. This model, though not a martingale model, gives an RMSE = 0.76, and its parameter estimates conform with estimates from historical data for an LRD VG model [7].

Thus, for a given maturity, the three-parameter MVG model and its associated (skew) VG model for historical data perform reasonably well in fitting option prices. If a four-parameter martingale model is to be used, the parent model of MVG that should be used is the MDG, in which the gamma process continues to play a fundamental role.

Historical Notes

In the case where $\sigma^2 = 2\alpha$ in the CF (2), the corresponding PDF (11) (with $\mu = \theta = 0$) already appears in [18], p. 184, equation (xlⁱⁱ), and is the theme of [19], where it is shown to be the distribution of difference of two i.i.d. gamma random variables, an idea clarified in [13]. The definition of the Bessel function $K_\eta(\omega)$ used differs from equation (13). Teichroew [24] obtained the PDF (11) (with $\mu = \theta = 0$), in terms of a Hankel function, from the normal variance-mixing structure of the distribution of X , using form (1) for the PDF of the mixing variable W . These themes are taken up by McLeish [17] as a starting point.

The skew VG distribution with $\alpha = 2n$, where n is a positive integer, and $-1 < \theta/\sigma^2 < 1$ appears in [26], a paper generalizing [19], which was also published in 1932.

Acknowledgments

Many thanks are due to Richard Finlay for his help.

References

- [1] Barndorff-Nielsen, O.E., Kent, J. & Sørensen, M. (1982). Normal variance-mean mixtures and z distributions, *International Statistical Review* **50**, 145–159.

Table 1 Fit of models to Schoutens [22] option data

Model	MDG	CGMY	GH	VG	MVG	BS
RMSE	2.24	2.76	2.88	3.56	3.56	6.73

- [2] Carr, P., Geman, H., Madan, D. & Yor, M. (2003). Stochastic volatility for Lévy processes, *Mathematical Finance* **13**, 345–382.
- [3] Carr, P. & Madan, D. (1999). Option valuation using the fast Fourier transform, *Journal of Computational Finance* **2**, 61–73.
- [4] Epps, T.W. (2000). *Pricing Derivative Securities*, World Scientific, Singapore.
- [5] Finlay, R. & Seneta, E. (2006). Stationary-increment Student and Variance-Gamma processes, *Journal of Applied Probability* **43**, 441–453.
- [6] Finlay, R. & Seneta, E. (2007). A gamma activity time process with noninteger parameter and self-similar limit, *Journal of Applied Probability* **44**, 950–959.
- [7] Finlay, R. & Seneta, E. (2008a). Stationary-increment Variance-Gamma and t -models: simulation and parameter estimation, *International Statistical Review* **76**, 167–186.
- [8] Finlay, R. & Seneta, E. (2008b). Option pricing with VG-like models, *International Journal of Theoretical and Applied Finance* **11**, 943–955.
- [9] Fung, T. & Seneta, E. (2007). Tailweight, quantiles and kurtosis. A study of competing distributions, *Operations Research Letters* **35**, 448–454.
- [10] Heyde, C.C. (1999). A risky asset model with strong dependence through fractal activity time, *Journal of Applied Probability* **36**, 1234–1239.
- [11] Heyde, C.C. & Leonenko, N.N. (2005). Student processes, *Advances in Applied Probability* **37**, 342–365.
- [12] Heyde, C.C. & Liu, S. (2001). Empirical realities for a minimal description risky asset model. The need for fractal features, *Journal of the Korean Mathematical Society* **38**, 1047–1059.
- [13] Kullback, S. (1936). The distribution laws of the difference and quotient of variables independently distributed in Pearson type III laws, *Annals of Mathematical Statistics* **7**, 51–53.
- [14] Lee, R. (2004). Option pricing by transform methods: extensions, unification and error control, *Journal of Computational Finance* **7**, 51–86.
- [15] Madan, D.B., Carr, P.P. & Chang, E.C. (1998). The Variance-Gamma process and option pricing, *European Finance Review* **2**, 79–105.
- [16] Madan, D.B. & Seneta, E. (1990). The Variance-Gamma (V.G.) model for share market returns, *Journal of Business* **63**, 511–524.
- [17] McLeish, D.L. (1982). A robust alternative to the normal distribution, *Canadian Journal of Statistics* **10**, 89–102.
- [18] Pearson, K., Jeffery, G.B. & Elderton, E.M. (1929). On the distribution of the first product-moment coefficient in samples drawn from an indefinitely large normal population, *Biometrika* **21**, 164–201.
- [19] Pearson, K., Stouffer, S.A. & David, F.N. (1932). Further applications in statistics of the $T_m(x)$ Bessel function, *Biometrika* **24**, 316–343.
- [20] Praetz, P.D. (1972). The distribution of share price changes, *Journal of Business* **45**, 49–55.
- [21] Press, S.J. (1967). A compound events model for security prices, *Journal of Business* **40**, 317–335.
- [22] Schoutens, W. (2003). *Lévy Processes in Finance. Pricing Financial Derivatives*, Wiley, Chichester.
- [23] Seneta, E. (2004). Fitting the Variance-Gamma model to financial data, in *Stochastic Methods and Their Applications (C.C. Heyde Festschrift)*, J. Gani & E. Seneta, eds, Journal of Applied Probability, Vol. 41A, pp. 177–187.
- [24] Teichroew, D. (1957). The mixture of normal distributions with different variances, *Annals of Mathematical Statistics* **28**, 510–512.
- [25] Tjetjep, A. & Seneta, E. (2006). Skewed normal variance-mean models for asset pricing and the method of moments, *International Statistical Review* **74**, 109–126.
- [26] Wishart, J. & Bartlett, M.S. (1932). The distribution of second order moment statistics in a normal system, *Proceedings of the Cambridge Philosophical Society* **28**, 455–459.

Further Reading

Seneta, E. (2007). The early years of the Variance-Gamma process, in *Advances in Mathematical Finance (Dilip B. Madan Festschrift)*, M.C. Fu, R.A. Jarrow, J.-Y.J. Yen, & R.J. Elliott, eds, Birkhäuser, Boston, pp. 3–19.

Related Articles

Exponential Lévy Models; Generalized Hyperbolic Models; Hazard Rate; Heavy Tails; Lévy Processes; Stylized Properties of Asset Returns; Tempered Stable Process.

EUGENE SENETA

Jump-diffusion Models

Jump-diffusion (JD) option pricing models are particular cases of exponential Lévy models (*see Exponential Lévy Models*) in which the frequency of jumps is finite. They can be considered as prototypes for a large class of more complex models such as the stochastic volatility plus jumps model of Bates (*see Bates Model*).

Consider a market with a riskless asset (the bond) and one risky asset (the stock) whose price at time t is denoted by S_t . In a JD model, the SDE for the stock price is given as

$$dS_t = \mu S_{t-} dt + \sigma S_{t-} dZ_t + S_{t-} dJ_t \quad (1)$$

where Z_t is a Brownian motion and

$$J_t = \sum_{i=1}^{N_t} Y_i \quad (2)$$

is a compound Poisson process where the jump sizes Y_i are independent and identically distributed with distribution F and the number of jumps N_t is a Poisson process with intensity λ . The asset price S_t thus follows geometric Brownian motion between jumps. Monte Carlo simulation of the process can be carried out by first simulating the number of jumps N_t , the jump times, and then simulating geometric Brownian motion on intervals between jump times.

The SDE (1) has the exact solution:

$$S_t = S_0 \exp\{\mu t + \sigma Z_t - \sigma^2 t/2 + J_t\} \quad (3)$$

Merton [5] considers the case where the jump sizes Y_i are normally distributed.

Risk-neutral Drift

If the above model is used as a pricing model, the drift μ in equation (1) is given by the risk-neutral drift $\hat{\mu}$, which contains a jump compensator μ_J :

$$\mu = \hat{\mu} + \mu_J \quad (4)$$

To identify μ_J , taking expectations of equation (1) and from the definition of $\hat{\mu}$,

$$\begin{aligned} \mathbb{E}[dS_t] &= \hat{\mu} S_t dt = \mu S_t dt \\ &+ \lambda \left\{ \int \xi F(d\xi) \right\} S_t dt \end{aligned} \quad (5)$$

where F is the risk-neutral jump distribution. The jump compensator is then given as

$$\mu_J = -\lambda \int \xi F(d\xi) \quad (6)$$

To simplify the presentation, we henceforth assume zero dividends so that $\hat{\mu} = r$, the risk-free rate.

Characteristic Function

Define the forward price $F := S_0 e^{rt}$. If $x_t := \log S_t/F$ is a Lévy process (*see Lévy Processes*), its characteristic function $\phi_T(u) := \mathbb{E}[e^{iux_T}]$ has the Lévy–Khintchine representation

$$\begin{aligned} \phi_T(u) &= \exp \left\{ i u (\mu_J - \sigma^2/2) T - u^2 \sigma^2/2 T \right. \\ &\quad \left. + T \int [e^{i u \xi} - 1] \nu(\xi) d\xi \right\} \end{aligned} \quad (7)$$

Typical assumptions for the distribution of jump sizes are as follows: normal as in the original paper by Merton [5] and double exponential as in [3] (*see Kou Model*). In the Merton model, the Lévy density $\nu(\cdot)$ is given as

$$\nu(\xi) = \frac{\lambda}{\sqrt{2\pi} \delta} \exp \left\{ -\frac{(\xi - \alpha)^2}{2\delta^2} \right\} \quad (8)$$

where α is the mean of the log-jump size $\log J$ and δ the standard deviation of jumps. This leads to the explicit characteristic function

$$\begin{aligned} \phi_T(u) &= \exp \left\{ i u \omega T - \frac{1}{2} u^2 \sigma^2 T \right. \\ &\quad \left. + \lambda T (e^{i u \alpha - u^2 \delta^2/2} - 1) \right\} \end{aligned} \quad (9)$$

with

$$\omega = -\frac{1}{2} \sigma^2 - \lambda (e^{\alpha + \delta^2/2} - 1) \quad (10)$$

In the double-exponential case (see **Kou Model**)

$$\nu(\xi) = \lambda \left\{ p \alpha_+ e^{-\alpha_+ \xi} \mathbf{1}_{\xi \geq 0} + (1-p) \alpha_- e^{-\alpha_- |\xi|} \mathbf{1}_{\xi < 0} \right\} \quad (11)$$

where α_+ and α_- are the expected positive and negative jump sizes, respectively, and p is the relative probability of a positive jump. This gives the explicit characteristic function:

$$\phi_T(u) = \exp \left\{ i u \omega T - \frac{1}{2} u^2 \sigma^2 T + \lambda T \times \left(\frac{p}{\alpha_+ - i u} - \frac{1-p}{\alpha_- + i u} \right) \right\} \quad (12)$$

with

$$\omega = -\frac{1}{2} \sigma^2 - \lambda \left(\frac{p}{\alpha_+ + 1} - \frac{1-p}{\alpha_- - 1} \right) \quad (13)$$

Pricing of European Options

Given a characteristic function, European call options can be priced using Fourier methods (see **Fourier Methods in Options Pricing**), as in [4]:

$$C(S, K, T) = e^{-rT} \left\{ F - \sqrt{FK} \frac{1}{\pi} \int_0^\infty \frac{du}{u^2 + \frac{1}{4}} \times \operatorname{Re} \left[e^{-iuk} \phi_T(u - i/2) \right] \right\} \quad (14)$$

where the log-strike $k := \log(K/F)$.

Valuation Equation

Assuming the process (1) and a constant risk-free rate r and further supposing that the market is complete, the value $V(S, t)$ of a European-style option satisfies

$$\frac{\partial V}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + r S \frac{\partial V}{\partial S} - r V + \lambda \int F(d\xi) \left\{ V(S e^\xi, t) - V(S, t) - (\xi - 1) S \frac{\partial V}{\partial S} \right\} = 0 \quad (15)$$

where for ease of notation, V denotes $V(S, t)$. Equation (15) is a partial integro-differential equation (PIDE) (see **Partial Integro-differential Equations (PIDEs)**), which can be solved using finite-difference methods [1].

A Valuation Formula for European Options

Merton [5] derived an exact solution of the valuation equation (15) for a European-style call option with strike K and time to expiration T , which has the form of an infinite sum of Black–Scholes-like terms:

$$C(S, K, T) = \sum_{n=0}^{\infty} \frac{e^{-\lambda T} (\lambda T)^n}{n!} \int F_n(d\xi) \times C_{BS}(S e^{\xi} e^{\mu_J T}, K, r, \sigma, T) \quad (16)$$

where F_n is the distribution of the sum of n -independent jumps and $C_{BS}(\cdot)$ denotes the Black–Scholes solution, which is given as

$$C_{BS}(S, K, r, \sigma, T) = S N(d_1) - K e^{-rT} N(d_2) \quad (17)$$

with

$$d_1 = \frac{\log(S/K) + rT + \frac{\sigma \sqrt{T}}{2}}{\sigma \sqrt{T}} \quad d_2 = \frac{\log(S/K) + rT - \frac{\sigma \sqrt{T}}{2}}{\sigma \sqrt{T}} \quad (18)$$

Jump to Ruin

In the case where $e^{Y_i} = 0$ with probability 1, $\mu_J = \lambda$ and equation (16) simplifies to

$$C(S, K, T) = e^{-\lambda T} C_{BS}(S e^{\lambda T}, K, r, \sigma, T) \quad (19)$$

which is the Black–Scholes formula with a shifted interest rate $r \rightarrow r + \lambda$. This special case of the JD model where the stock price jumps to zero (or ruin) whenever there is a jump is the simplest possible model of default. Equation (19) for the option price in the jump-to-ruin model may also be derived from a Black–Scholes style replication argument using stock and bonds of the issuer of the stock; upon default of the issuer, both stock and bonds jump to zero. The

cost of funding stock with bonds of the issuer is $r + \lambda$ in this picture, which explains the simple form (19) of the solution.

Local Volatility

There is a particularly simple expression for Dupire local volatility in the jump-to-ruin model. It is given by Gatheral [2]:

$$\sigma_{\text{loc}}^2(K, T; S) = \sigma^2 + 2\lambda\sigma\sqrt{T} \frac{N(d_2)}{N'(d_2)} \quad (20)$$

with

$$d_2 = \frac{\log(S/K) + \lambda T}{\sigma\sqrt{T}} - \frac{\sigma\sqrt{T}}{2} \quad (21)$$

As $K \rightarrow \infty$, $d_2 \rightarrow -\infty$ and the correction term vanishes, and as $K \rightarrow 0$, the correction term explodes. In addition, as the *hazard rate* λ increases, so does d_2 , increasing the local volatility for low strikes K relative to high strikes.

Normally Distributed Jumps

Merton [5] also shows that if jumps are normally distributed with $Y_i \sim N(\alpha, \delta)$, equation (16) again simplifies considerably to give

$$C(S, K, T) = \sum_{n=0}^{\infty} \frac{e^{-\lambda' T} (\lambda' T)^n}{n!} C_{\text{BS}}(S, K, r_n, \sigma_n, T) \quad (22)$$

with

$$\lambda' = \lambda e^{\alpha + \delta^2/2} \quad (23)$$

$$\sigma_n^2 T = \sigma^2 T + n \delta^2 \quad (24)$$

$$r_n T = (r + \mu_J) T + n (\alpha + \delta^2/2) \quad (25)$$

Each term $C_{\text{BS}}(S, K, r_n, \sigma_n, T)$ in equation (22) is the value of the option conditional on there being exactly n jumps during its life.

Implied Volatility Smile

If there were no jumps in this model, the implied volatility smile would be flat. Jumps in the JD stock price process induce an implied volatility smile whose short time limit (see, e.g., [2]) is given as

$$K \frac{\partial}{\partial K} \sigma_{\text{BS}}^2(K, T) \rightarrow -2\mu_J \text{ as } T \rightarrow 0 \quad (26)$$

The greater the μ_J , because jumps are either more frequent or more negatively skewed, the more negative is the implied volatility skew.

References

- [1] Cont, R. & Voltchkova, E. (2005). A finite difference scheme for option pricing in jump-diffusion and exponential Lévy models, *SIAM Journal on Numerical Analysis* **43**(4), 1596–1626.
- [2] Gatheral, J. (2006). *The Volatility Surface*, John Wiley & Sons, Hoboken, Chapter 5.
- [3] Kou, S. (2002). A jump-diffusion model for option pricing, *Management Science* **48**, 1086–1101.
- [4] Lewis, A.L. (2000). *Option Valuation under Stochastic Volatility with Mathematica Code*, Finance Press, Newport Beach, CA.
- [5] Merton, R.C. (1976). Option pricing when underlying stock returns are discontinuous, *Journal of Financial Economics* **3**, 125–144.

Related Articles

Bates Model; Exponential Lévy Models; Fourier Methods in Options Pricing; Implied Volatility Surface; Kou Model; Partial Integro-differential Equations (PIDEs).

JIM GATHERAL

Time-changed Lévy Process

If L is a Lévy process (*see Lévy Processes*) and $(T_t)_{t \geq 0}$ is a positive increasing process, $X_t = L(T_t)$ is called a *time-changed Lévy process* with time change $(T_t)_{t \geq 0}$. When the time change is independent from L , many properties of X can be derived from those of L .

Many **Lévy processes** have appeared as time-changed Brownian motions (*see Time Change*), so one might ask why one should time change them yet again. In this regard, we note that Lévy processes are by construction processes of independent identically distributed increments and hence all distributional parameters such as variances, skewness, kurtosis, and possibly correlations are constant. Yet all these and possibly more entities are stochastic in actual economies, and this randomness may be important for particular questions of interest. These considerations led in the first instance to the construction of processes displaying stochastic volatility with the local innovations of a Lévy process. Analytical tractability of characteristic functions motivated models with exponential affine characteristic functions (*see Affine Models*) and through the work of Duffie *et al.* [4] it became well known that this would be the case if the infinitesimal generator of the resulting Markov process was linear in state variables. The recipe for constructing these models was, therefore, clear.

A number of models in this direction were presented by Carr *et al.* [2]. Three Lévy processes were selected for being time changed and they were the normal inverse Gaussian (NIG) (*see Normal Inverse Gaussian Model*), the variance gamma (VG) (*see Variance-gamma Model*), and the CGMY model (*see Tempered Stable Process*). The first two were already known to be time changes of Brownian motion with drift. Cont and Tankov [3, Prop. 4.1] show that the CGMY also has such a representation. Characteristic exponents $\psi(u; \cdot)$ are critical to the development of the exponential affine representations (*see Affine Models*) involved, which are given here by the logarithm of the characteristic functions taken at unit time. For these three models, the characteristic exponents are given as (subscript indicates

the model name)

$$\begin{aligned} \psi_{\text{NIG}}(u; \sigma, \nu, \theta) \\ = \sigma \left(\frac{\nu}{\theta} - \sqrt{\frac{\nu^2}{\theta^2} - 2 \frac{\theta i u}{\sigma^2} + u^2} \right), \sigma, \nu > 0, \theta \in \mathbb{R} \end{aligned} \quad (1)$$

$$\begin{aligned} \psi_{\text{VG}}(u; \sigma, \nu, \theta) \\ = \frac{1}{\nu} \log \left(1 - i u \theta \nu - \frac{\sigma^2 \nu}{2} u^2 \right), \sigma, \theta, \nu \in \mathbb{R} \end{aligned} \quad (2)$$

$$\begin{aligned} \psi_{\text{CGMY}}(u; C, G, M, Y) \\ = C \Gamma(-Y) ((M - i u)^Y - M^Y + (G + i u)^Y \\ - G^Y), C, G, M > 0, 0 < Y < 2. \end{aligned} \quad (3)$$

The details of the associated Lévy processes are given, which would assist in various applications. The NIG and VG processes can be written as Brownian motion with drift θ and volatility σ time changed by an inverse Gaussian process and a gamma process, respectively. The inverse Gaussian process T_t^ν is the time taken by an independent Brownian motion with drift ν to reach the level t , while the gamma process G_t^ν is an increasing process with independent identically distributed increments where the increments over unit time have a gamma distribution with unit mean and volatility ν . Both the NIG and VG are pure jump processes with Lévy measures $k_{\text{NIG}}(x) dx$, $k_{\text{VG}}(x) dx$ defined as

$$k_{\text{NIG}}(x) = \sqrt{\frac{2}{\pi}} \sigma \alpha^2 \frac{e^{\beta x} K_1(|x|)}{|x|} \quad (4)$$

$$\beta = \frac{\theta}{\sigma^2}, \quad \alpha^2 = \frac{\nu^2}{\sigma^2} + \frac{\theta^2}{\sigma^4} \quad (5)$$

$$k_{\text{VG}}(x) = \frac{C}{|x|} \exp\left(\frac{G-M}{2}x\right) \exp\left(-\frac{G+M}{2}|x|\right) \quad (6)$$

$$\begin{aligned} C = \frac{1}{\nu}; G = \left(\sqrt{\frac{\theta^2 \nu^2}{4} + \frac{\sigma^2 \nu}{2}} - \frac{\theta \nu}{2} \right)^{-1}, \\ M = \left(\sqrt{\frac{\theta^2 \nu^2}{4} + \frac{\sigma^2 \nu}{2}} + \frac{\theta \nu}{2} \right)^{-1} \end{aligned} \quad (7)$$

where $K_\alpha(x)$ is the Bessel K function.

The CGMY process was defined in terms of its Lévy measure $k_{\text{CGMY}}(x) dx$ with

$$k_{\text{CGMY}}(x) = \frac{C}{|x|^{1+Y}} \exp\left(\frac{G-M}{2}x\right) \times \exp\left(-\frac{G+M}{2}|x|\right) \quad (8)$$

It was shown in [3, Proposition 4.1] (see also [6]) that the CGMY process can be represented as Brownian motion with drift $(G-M)/2$ time changed by a shaved stable $\frac{Y}{2}$ subordinator with shaving function

$$f(y) = e^{-\frac{(B^2 - A^2)y}{2}} E \left[e^{-\frac{B^2 y}{2} \frac{\gamma_{Y/2}}{\gamma_{1/2}}} \right],$$

$$A = \frac{G-M}{2}, \quad B = \frac{G+M}{2} \quad (9)$$

where $\gamma_{Y/2}, \gamma_{1/2}$ are independent gamma variates.

One may explicitly evaluate in terms of Hermite functions:

$$E \left[e^{-\frac{B^2 y}{2} \frac{\gamma_{Y/2}}{\gamma_{1/2}}} \right] = \frac{\Gamma(Y)}{\Gamma\left(\frac{Y}{2}\right) 2^{\frac{Y}{2}-1}} h_{-Y}(B\sqrt{y}) \quad (10)$$

where

$$h_\nu(z) = \frac{1}{\Gamma(-\nu)} \int_0^\infty e^{-y^2/2 - yz} y^{-\nu-1} dy \quad (\nu < 0) \quad (11)$$

A Continuous Time Change

We can introduce stochastic volatility along with a clustering of volatility by time changing these Lévy processes by the integral of the square root process $y(t)$, where

$$dy = \kappa(\eta - y) dt + \lambda\sqrt{y} dW \quad (12)$$

for an independent Brownian motion $(W(t), t > 0)$. For a candidate Lévy process $X(t)$, we consider as a model for the uncertainty driving the stock the composite process

$$Z(t) = X(Y(t)) \quad (13)$$

where

$$Y(t) = \int_0^t y(u) du \quad (14)$$

The characteristic function for the composite process is easily derived from the characteristic function of $Y(t)$ as

$$E[e^{iuY(t)}] = \phi(u, t, y(0), \kappa, \eta, \lambda) = A(t, u) \exp(B(t, u)y(0)) \quad (15)$$

$$A(t, u) = \frac{\exp\left(\frac{\kappa^2 \eta t}{\lambda^2}\right)}{\left(\cosh\left(\frac{\gamma t}{2}\right) + \frac{\kappa}{\gamma} \sinh\left(\frac{\gamma t}{2}\right)\right) \frac{2\kappa\eta}{\lambda^2}} \quad (16)$$

$$B(t, u) = \frac{2iu}{\kappa + \gamma \coth\left(\frac{\gamma t}{2}\right)} \quad (17)$$

$$\gamma = \sqrt{\kappa^2 - 2\lambda^2 iu} \quad (18)$$

It follows that

$$E[\exp(iuZ(t))] = \phi(-i\psi_X(u), t, y(0), \kappa, \eta, \lambda) \quad (19)$$

The Stock Price Model

There are two approaches to model the stock price $S(t)$. The first approach takes the exponential of the composite process corrected to get the correct forward price, whereby we define

$$S_1(t) = S(0) \frac{\exp(Z(t))}{E[\exp(Z(t))]} \quad (20)$$

In this case, the stock price has the right forward and the resulting option prices are free of static arbitrage. However, there may be the possibility of dynamic arbitrage in the model and this is an issue if the model is being used continuously to quote on options with constant parameters through time. To exclude dynamic arbitrage in the model, one could form a martingale model for the forward stock price by modeling it as the stochastic exponential of the martingale:

$$n(t) = Z(t) - \int_0^t \int_{-\infty}^\infty xy(t)k_X(x) dx ds \quad (21)$$

In the second approach, one writes the stock price process $S_2(t)$ as

$$S_2(t) = S(0) \exp((r - q)t) \exp(Z(t) - Y(t)\psi_X(-i)) \quad (22)$$

For the first approach, the log characteristic function for the logarithm of the stock price is given as

$$\begin{aligned} E[\exp(iu \log S_1(t))] \\ = \exp(iu(\log(S(0) + (r - q)t))) \\ \times \frac{\phi(-i\psi_X(u), t, y(0); \kappa, \eta, \lambda)}{\phi(-i\psi_X(-i), t, y(0); \kappa, \eta, \lambda)^{iu}} \end{aligned} \quad (23)$$

The second approach leads to the following characteristic function:

$$\begin{aligned} E[\exp(iu \log S_2(t))] \\ = \exp(iu(\log(S(0) + (r - q)t))) \\ \times \phi(-i\psi_X(u) - u\psi_X(-i), t, y(0); \kappa, \eta, \lambda) \end{aligned} \quad (24)$$

The models of the first approach are termed *NIGSA*, *VGSA*, and *CGMYSA* for NIG, VG, and CGMY with a stochastic arrival rate of jumps adapted to the level of the process $y(t)$. The models of the second approach are martingale models and are termed *NIGSAM*, *VGSA*, and *CGMYSAM*, respectively. It is observed in calibrations that the first approach generally fits the option price data better.

Some Discontinuous Time Changes

One can replace the continuous stochastic process for the arrival rate of jump activity $y(t)$ by a discontinuous process that now only has upward jumps. We call this process $y^J(t)$ for discontinuous jump arrival rates. Given a background driving Lévy process (BDLP) $U(t)$ with only positive jumps, we define

$$dy^J(t) = -\kappa y^J(t) dt + dU(t) \quad (25)$$

The composite process now permits some direct dependence between arrival rate jumps and the underlying uncertainty:

$$Z^J(t) = X(Y^J(t)) + \rho U(t) \quad (26)$$

$$Y^J(t) = \int_0^t y^J(s) ds \quad (27)$$

We suppose that the background driving Lévy process has the following characteristic function:

$$E[\exp(iuU(t))] = \exp(t\psi_U(u)) \quad (28)$$

The characteristic function of the composite process $Z^J(t)$ may be developed in terms of the joint characteristic function of $Y^J(t)$, $U(t)$ as

$$\Phi_t(a, b) = E[\exp(iaY^J(t) + ibU(t))] \quad (29)$$

We may show that

$$E[\exp(iuZ^J(t))] = \Phi_t(-i\psi_X(u), u\rho) \quad (30)$$

We have that

$$\begin{aligned} \Phi_t(a, b) = \exp\left(iaY(0)\frac{1 - e^{-\kappa t}}{\kappa}\right) \\ \times \exp\left(\int_L^U \frac{\psi_U(v)}{a + \kappa b - \kappa v} dv\right) \end{aligned} \quad (31)$$

$$L = b \quad (32)$$

$$U = b + a \frac{1 - e^{-\kappa t}}{\kappa} \quad (33)$$

The characteristic functions for the logarithm of the stock price for the exponential model are now

$$\begin{aligned} E[\exp(iu \log(S_1^J(t)))] \\ = \exp(iu(\log(S(0) + (r - q)t))) \\ \times \Phi_t(-i\psi_X(u), \rho u) \\ \times \exp(-iu \log(\Phi_t(-i\psi_X(-i), -i\rho))) \end{aligned} \quad (34)$$

For the stochastic exponential, the result is given as

$$\begin{aligned} E[\exp(iu \log(S_2^J(t)))] \\ = \exp(iu(\log(S(0)) + (r - q)t - \psi_U(-i\rho)t)) \\ \times \Phi_t(-i\psi_X(u) - u\psi_X(-i), \rho u) \end{aligned} \quad (35)$$

Some explicit examples for $\psi_U(u)$, for which we may obtain exact expressions for $\Phi_t(a, b)$, remain to be determined.

Examples for $\psi_U(u)$ and $\Phi_t(a, b)$

Three explicit models for ψ_U were developed. These are SG for stationary gamma, IG for inverse Gaussian, and SIG for stationary inverse Gaussian.

The SG Case

In this case, the Lévy density for jumps in the process $U(t)$ is

$$k_U(x) = \frac{\lambda}{\zeta} e^{-x/\zeta} \quad (36)$$

The log characteristic function of the BDLP is

$$\psi_U(u) = \frac{i u \lambda}{1/\zeta - i u} \quad (37)$$

The IG Case

The Laplace transform for inverse Gaussian (*see Normal Inverse Gaussian Model*) time with drift ν for the Brownian motion is

$$E \left[\exp(-\lambda T_1^\nu) \right] = \exp \left(\nu - \sqrt{\nu^2 + 2\lambda} \right) \quad (38)$$

and the log characteristic function is

$$\psi_U(u) = \nu - \sqrt{\nu^2 - 2iu} \quad (39)$$

The SIG Case

For this case, Barndorff-Nielsen and Shephard [1] show that the Lévy density is

$$k_U(x) = \frac{1}{2\sqrt{2\pi}} x^{-3/2} (1 + \nu^2 x) \exp \left(-\frac{\nu^2 x}{2} \right) \quad (40)$$

The log characteristic function is

$$\psi_U(u) = \frac{i u}{\sqrt{\nu^2 - 2iu}} \quad (41)$$

For these three cases, the construction of $\Phi_t(a, b)$ is completed on determining the integral

$$\int_L^U \frac{\psi_Z(v)}{a + \kappa b - \kappa v} dv = \Psi(U, a, b) - \Psi(L, a, b) \quad (42)$$

and we have analytic expressions for $\Psi(x, a, b)$ in the SG, IG, and SIG cases that are as follows:

$$\begin{aligned} \Psi_{SG}(x, a, b; \kappa, \lambda, \zeta) \\ = \log \left[\left(x + \frac{i}{\zeta} \right)^{\frac{\lambda}{\kappa - i\zeta(a + \kappa b)}} \right. \\ \left. \times (a + \kappa b - \kappa x)^{\frac{\lambda(a + \kappa b)\zeta}{\kappa((a + \kappa b)\zeta + i\kappa)}} \right] \quad (43) \end{aligned}$$

$$\begin{aligned} \Psi_{IG}(x, a, b; \kappa, \nu) \\ = \frac{2\sqrt{\nu^2 - 2ix}}{\kappa} + \frac{2\sqrt{\nu^2 \kappa - 2i(a + \kappa b)}}{\kappa^{3/2}} \\ \times \operatorname{arctanh} \left[\frac{\sqrt{\kappa} \sqrt{\nu^2 - 2ix}}{\sqrt{\nu^2 \kappa - 2i(a + \kappa b)}} \right] \\ - \frac{\nu \log(a + \kappa b - \kappa x)}{\kappa} \quad (44) \end{aligned}$$

$$\begin{aligned} \Psi_{SIG}(x, a, b; \kappa, \nu) \\ = \frac{\sqrt{\nu^2 - 2ix}}{\kappa} - \frac{2i(a + \kappa b)}{\kappa^{3/2} \sqrt{\nu^2 \kappa - 2i(a + \kappa b)}} \\ \times \operatorname{arctanh} \left[\frac{\sqrt{\kappa} \sqrt{\nu^2 - 2ix}}{\sqrt{\nu^2 \kappa - 2i(a + \kappa b)}} \right] \quad (45) \end{aligned}$$

Correlation in VGSA or VGCSA

We consider the introduction of correlation in VGSA along the following lines. We define the correlated uncertainty as

$$Z^C(t) = X(Y(t)) + \rho y(t) \quad (46)$$

The characteristic function now follows from the joint characteristic function of $Y(t)$, $y(t)$:

$$E \left[e^{iuZ^C(t)} \right] = E \left[e^{Y(t)\psi_X(u) + iu\rho y(t)} \right] \quad (47)$$

Let

$$\Phi_t^C(a, b, x) = E \left[\exp(iaY(t) + iby(t)) | y(0) = x \right] \quad (48)$$

We have

$$\phi_{Z^C}(u) = \Phi_t^C(-i\psi_X(u), \rho u) \quad (49)$$

We recall the solution for $\Phi_t(a, b, x)$ from [5, 7] as

$$\begin{aligned} \Phi_t^C(a, b, x) &= A^C(t, a, b) \exp(B^C(t, a, b)x) \\ A^C(t, a, b) &= \end{aligned} \quad (50)$$

$$\begin{aligned} &= \frac{\exp \left(\frac{\kappa^2 \eta t}{\lambda^2} \right)}{\left[\cosh \left(\frac{\gamma t}{2} \right) + \frac{(\kappa - ib\lambda^2)}{\gamma} \sinh \left(\frac{\gamma t}{2} \right) \right]^{\frac{2\kappa\eta}{\lambda^2}}} \quad (51) \end{aligned}$$

$$B^C(t, a, b) = \frac{ib \left[\gamma \cosh\left(\frac{\gamma t}{2}\right) - \kappa \sinh\left(\frac{\gamma t}{2}\right) \right] + 2ia \sinh\left(\frac{\gamma t}{2}\right)}{\gamma \cosh\left(\frac{\gamma t}{2}\right) + (\kappa - ib\lambda^2) \sinh\left(\frac{\gamma t}{2}\right)} \quad (52)$$

$$\gamma = \sqrt{\kappa^2 - 2\lambda^2 ia} \quad (53)$$

We get the characteristic function for the model VGCSA, where the letter C denotes correlated stochastic arrival by exponentiation as

$$\begin{aligned} E[\exp(iu \log(S_1^C(t)))] \\ = \exp(iu(\log(S(0)) + (r - q)t)) \\ \times \Phi_t^C(-i\psi_X(u), \rho u) \\ \times \exp(-iu \log(\Phi_t^C(-i\psi_X(-i), -i\rho))) \quad (54) \end{aligned}$$

Exciting the Jumps by the Level of Activity that Is Also a Heston Type of Correlated Volatility

In this class of models, we introduce stochastic volatility and allow jump arrival rates to respond to the volatility on each side with separate sensitivities. This will give rise to stochastic skewness as well as to volatility. The model for the logarithm of the stock price $H(t) = \log(S(t))$ is now as follows:

$$\begin{aligned} H(t) = H(0) + (r - q)t - \int_0^t \frac{y(u)}{2} du \\ - (c_p t + s_p Y(t)) \int_0^\infty (e^x - 1 - x) k_p(x) dx \\ - (c_n t + s_n Y(t)) \int_{-\infty}^0 (e^x - 1 - x) k_n(x) dx \\ + x * (\mu - \nu) + \int_0^t \sqrt{y(u)} dW_S(u) \quad (55) \end{aligned}$$

$$Y(t) = \int_0^t y(u) du \quad (56)$$

$$dy = \kappa(\eta - y) dt + \lambda \sqrt{y} dW_y(t) \quad (57)$$

$$dW_y dW_S = \rho dt \quad (58)$$

$$\begin{aligned} \nu(dx, dt) = (c_p + s_p y(t)) k_p(x) \mathbf{1}_{x>0} dx \\ + (c_n + s_n y(t)) k_n(x) \mathbf{1}_{x<0} dx \quad (59) \end{aligned}$$

The growth rate of the stock price is at the risk neutral level of $(r - q)$. The coefficients c_p, c_n are the Lévy jump response components. The sensitivities of jumps to volatility are captured by the two slope coefficients s_p, s_n for the positive and the negative sides. The logarithm of the stock price is a continuous martingale with stochastic volatility plus a compensated jump martingale that has jumps responding to volatility with log price drift set to fix the stock drift at $r - q$.

The joint characteristic function of the log of the stock price, the level of the terminal variance, and the remaining integrated variance is

$$\begin{aligned} \Psi(t, H(t), y(t)) \\ = E_t \left[\exp(iaH(T) + iby(T) + ic \int_t^T y(u) du) \right] \quad (60) \end{aligned}$$

We have a closed form for Ψ in this model given as

$$\Psi(t, H(t), y(t)) = A(\tau) \exp(iaH(t) + \gamma(\tau)y(t)) \quad (61)$$

$$A(\tau) = \exp \left(\left(ia(r - q) + c_p u_p + c_n u_n + (\kappa - \lambda \rho ia) \frac{\kappa \eta}{\lambda^2} \right) \tau \right) \left(\frac{\cosh(D)}{\cosh\left(D - \frac{\tau}{2} \xi\right)} \right)^{\frac{2\kappa \eta}{\lambda^2}} \quad (62)$$

$$\gamma(\tau) = \frac{\kappa - \lambda \rho i a}{\lambda^2} + \frac{\xi}{\lambda^2} \tanh\left(D - \frac{\xi}{2} \tau\right) \quad (63)$$

$$\xi = \sqrt{(\kappa - \lambda \rho i a)^2 + \lambda^2 [a^2 + i(a - 2c) - 2(s_p u_p + s_n u_n)]} \quad (64)$$

$$u_p = \int_0^\infty (e^x - 1 - i a x) k_p(x) dx - i a \int_0^\infty (e^x - 1 - x) k_p(x) dx \quad (65)$$

$$u_n = \int_{-\infty}^0 (e^x - 1 - i a x) k_n(x) dx - \int_{-\infty}^0 (e^x - 1 - x) k_n(x) dx \quad (66)$$

$$D = \tanh^{-1}\left(\frac{i b \lambda^2}{\xi} - \frac{\kappa - \lambda \rho i a}{\xi}\right) \quad (67)$$

On setting $b = c = 0$, we obtain the characteristic function of the log of the final stock price and this yields the models: SVADNE, SVAVG, and SVAC-CGMY.

We note that for DNE

$$u_p = (1 - i a) \frac{1}{\beta_p - 1} \quad (68)$$

$$u_n = -(1 - i a) \frac{1}{\beta_n + 1} \quad (69)$$

The corresponding calculations for VG in the CGM parameterization are

$$u_p = \log\left(\frac{M}{M - 1}\right) \quad (70)$$

$$u_n = \log\left(\frac{G}{G + 1}\right) \quad (71)$$

For CCGMY, we have the following result:

$$u_p = \Gamma(-y_p) ((M - 1)^{y_p} - M^{y_p}) \quad (72)$$

$$u_n = \Gamma(-y_n) ((G + 1)^{y_n} - G^{y_n}) \quad (73)$$

References

- [1] Barndorff-Nielsen, O.E. (1998). Processes of normal inverse Gaussian type, *Finance and Stochastics* **2**, 41–68.
- [2] Carr, P., Geman, H., Madan, D. & Yor, M. (2003). Stochastic volatility for Levy processes, *Mathematical Finance* **13**, 345–382.

- [3] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, Series in Financial Mathematics, CRC Press.
- [4] Duffie, D., Filipovic, D. & Schachermayer, W. (2003). Affine processes and applications in finance, *Annals of Applied Probability* **13**, 984–1053.
- [5] Lambertson, D. & Lapeyre, B. (1996). *Introduction to Stochastic Calculus Applied to Finance*, Chapman and Hall, New York.
- [6] Madan, D. & Yor, M. (2008). Representing the CGMY and Meixner Levy processes as time changed Brownian motions, *Journal of Computational Finance* Fall, 27–47.
- [7] Pitman, J. & Yor M. (1982). A decomposition of Bessel Bridges, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **59**, 425–457.

Further Reading

- Carr, P., Geman, H., Madan, D. & Yor, M. (2002). The fine structure of asset returns: an empirical investigation, *Journal of Business*, **75**(2), 305–332.
- Madan, D., Carr, P. & Chang, E. (1998). The variance gamma process and option pricing, *European Finance Review* **2**, 79–105.
- Madan, D.B. & Seneta, E. (1990). The Variance Gamma (VG) model for share market returns, *Journal of Business* **63**, 511–524.

Related Articles

Affine Models; Barndorff-Nielsen and Shephard (BNS) Models; Exponential Lévy Models; Heston Model; Lévy Processes; Normal Inverse Gaussian Model; Squared Bessel Processes; Stochastic Exponential; Tempered Stable Process; Time Change; Variance-gamma Model.

DILIP B. MADAN

Kou Model

It is well known that empirically asset return distributions have heavier tails (*see* **Heavy Tails**) than those of normal distributions, in contrast to the classical Black–Scholes model (*see* **Black–Scholes Formula**). Jump-diffusion models are among the most popular alternative models proposed to address this issue, and they are especially useful to price options with short maturities (*see* **Exponential Lévy Models**). However, analytical tractability is one of the challenges faced by many alternative models. More precisely, although many alternative models can lead to analytical solutions for European call and put options, unlike the Black–Scholes model, it is difficult to do so for path-dependent options such as lookback (*see* **Lookback Options**), barrier (*see* **Barrier Options**), and American options, which are treated using numerical methods (*see* **Partial Integro-differential Equations (PIDEs)**). For example, the convergence rates of binomial trees and Monte Carlo simulation for path-dependent options are typically much slower than those for call and put options; *see* Boyle *et al.* [3].

The double exponential jump-diffusion model is a jump-diffusion model in which the jump size distribution follows a two-sided exponential distribution. It was introduced to further extend the analytical tractability of models with jumps.

In jump-diffusion models under the physical probability measure P , the asset price, $S(t)$, is modeled as

$$\frac{dS(t)}{S(t-)} = \mu dt + \sigma dW(t) + d\left(\sum_{i=1}^{N(t)} (V_i - 1)\right) \quad (1)$$

where $W(t)$ is a standard Brownian motion, $N(t)$ is a Poisson process with rate λ , and $\{V_i\}$ is a sequence of independent identically distributed (i.i.d.) nonnegative random variables. All the sources of randomness, $N(t)$, $W(t)$, and V 's, are assumed to be independent. Solving the stochastic differential equation (1) gives the dynamics of the asset price:

$$S(t) = S(0) \exp \left\{ \left(\mu - \frac{1}{2} \sigma^2 \right) t + \sigma W(t) \right\} \prod_{i=1}^{N(t)} V_i \quad (2)$$

In the Merton model [28], $Y = \log(V)$ has a normal distribution. In the double exponential jump-diffusion model [23] $Y = \log(V)$ has an asymmetric double exponential distribution with the density

$$f_Y(y) = p \cdot \eta_1 e^{-\eta_1 y} 1_{\{y \geq 0\}} + q \cdot \eta_2 e^{\eta_2 y} 1_{\{y < 0\}}, \quad \eta_1 > 1, \eta_2 > 0 \quad (3)$$

where $p, q \geq 0$, $p + q = 1$, represent the probabilities of upward and downward jumps. The requirement $\eta_1 > 1$ is needed to ensure that $E(V) < \infty$ and $E(S(t)) < \infty$; it essentially means that the average upward jump cannot exceed 100%, which is quite reasonable [2].

As pointed out in [23], the jump part of the double exponential jump-diffusion model can be interpreted as the market response to outside developments; and the heavier tail and higher peak (in comparison to the standard normal distribution) of the double exponential distribution attempt to model market overreaction and underreaction, respectively. Ramezani and Zeng [29] independently proposed the double exponential jump-diffusion model from an econometric viewpoint as a way of improving the empirical fit of Merton's normal jump-diffusion model to stock price data.

Such models lead to incomplete markets in which the replication of an option payoff is impossible. The monograph by Cont and Tankov [10] discusses hedging issues for jump-diffusion models and resulting pricing measures. Alternatively, one can use the rational expectations in [27] and [32] to choose a risk-neutral measure to price derivative as in [23].

The double exponential jump-diffusion model belongs to the class of exponential Lévy models (*see* **Exponential Lévy Models**). There is a large literature on Lévy processes in finance, including several excellent books, for example, the books by Cont and Tankov [10] and Kijima [22].

Analytical Tractability

The main advantage of the double exponential jump-diffusion model is that it offers a rare case where we can derive the analytical solution of the joint distribution of the first passage time and $X(t) = \log(S(t)/S(0))$, thereby making it possible to price path-dependent options such as lookback, barrier, and perpetual American options. An intuitive explanation for this follows.

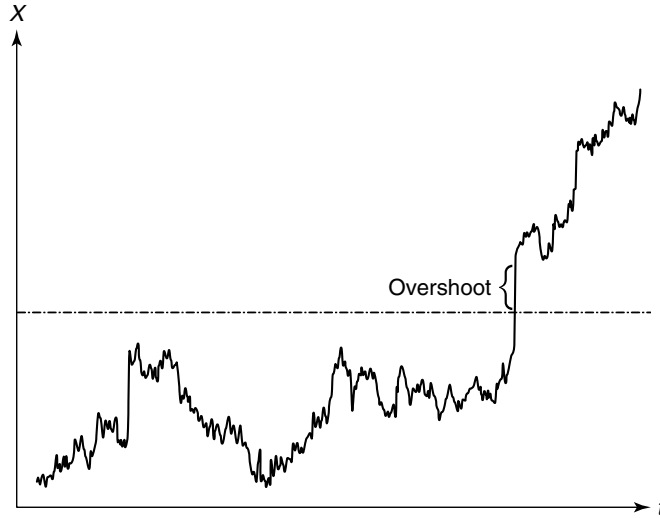


Figure 1 A simulated sample path with the overshoot problem

To price lookback, barrier, and perpetual American options, it is pivotal to study the first passage times τ_b when the process crosses a flat boundary with a level b . Without loss of generality, assume that $b > 0$. When a jump-diffusion process crosses the boundary, sometimes it hits the boundary exactly and sometimes it incurs an “overshoot”, $X_{\tau_b} - b$, over the boundary as shown in Figure 1. The overshoot presents several problems if one wants to compute the distribution of the first passage time analytically. First, one needs the exact distribution of the overshoot, $X_{\tau_b} - b$; particularly, $P(X_{\tau_b} - b = 0)$ and $P(X_{\tau_b} - b > x)$, $x > 0$. Second, one needs to know the dependence structure between the overshoot, $X_{\tau_b} - b$, and the first passage time τ_b .

These difficulties may be resolved under the assumption that the jump size Y has a double exponential distribution. Mathematically, this is because the exponential function has some very nice properties, such as the product of exponential functions is still an exponential function, and the derivatives of exponential functions are still exponential function. These nice properties enable us to solve related ordinary integro-differential equations (OIDE) explicitly, leading to analytical solutions for the marginal and joint distributions of the first passage times, and ultimately, analytical tractability for pricing lookback, barrier, and perpetual American options. More precisely, the infinitesimal generator of the return

process $X(t)$ is given by

$$\begin{aligned} \mathcal{L}u(x) = & \frac{1}{2}\sigma^2 u''(x) + \tilde{\mu}u'(x) \\ & + \lambda \int_{-\infty}^{\infty} [u(x+y) - u(x)] f_Y(y) dy \end{aligned} \quad (4)$$

for all twice continuously differentiable functions $u(x)$. When studying the first passage time, we encounter an OIDE with discontinuous regions as follows:

$$\begin{cases} (\mathcal{L}u)(x) = \alpha u(x), & x < x_0 \\ u(x) = g(x), & x \geq x_0 \end{cases} \quad (5)$$

where $\alpha > 0$ and $g(x)$ is a given function. Many times x_0 is a fixed number, but in the case of American options, x_0 is a parameter that needs to be determined by solving a free boundary problem. Note that $u(x)$ solves the OIDE not for all $x \in \mathbb{R}$ but only for $x < x_0$. However, $u(x)$ does involve the information on $x > x_0$, as the integral inside the generator (4) depends on the function $g(x)$, thereby making itself more complicated. This OIDE can be solved explicitly under the double exponential jump-diffusion model, thereby leading to an analytical solution of the joint distribution of the first passage time τ_b and X_{τ_b} ; see [25, 26], and [24].

In addition to pricing options related to the first passage times, the double exponential jump-diffusion models have been studied in many papers. What is detailed below is only a snapshot of some interesting results.

1. In terms of computational issues, see [11] and [12] for numerical methods *via* solving partial integro-differential equations (*see* **Partial Integro-differential Equations (PIDEs)**); Feng and Linetsky [17] and Feng *et al.* [16] showed how to price path-dependent options numerically *via* extrapolation and variational methods.
2. In terms of applications, see the references in [18] for applications in fixed income derivatives and term structure models, and the references in [9] for applications in credit risk and credit derivatives.
3. Double-barrier options (with both upper and lower barriers) are studied in [30] and [4].
4. Statistical inference and econometric analysis for Lévy processes are discussed in [31].

Volatility Clustering Effect

In addition to the leptokurtic feature, returns distributions also have an interesting dependent structure, called the *volatility clustering effect*; see [14]. More precisely, the volatility of returns (which are related to the squared returns) are correlated, but asset returns themselves have almost no autocorrelation. In other words, a large movement in asset prices, either upward or downward, tends to generate large movements in the future asset prices, although the direction of the movements is unpredictable.

In particular, any model for stock returns with independent increments (such as Lévy processes) cannot incorporate the volatility clustering effect. However, one can combine jump-diffusion processes with other processes [1, 13] or consider time-changed Brownian motion and Lévy processes (*see* **Time Change**) to incorporate the volatility clustering effect. More precisely, if $\tau(t)$ contains a diffusion component (i.e., not a subordinator), then $W(\tau(t))$ and $X(\tau(t))$ may have dependent increments and no longer be Lévy processes; see [6, 7], and [8].

Hyper-Exponential Jumps

Although the main empirical motivation for using Lévy processes in finance comes from the fact that asset return distributions tend to have tails heavier than those of normal distribution, it is not clear how heavy the tail distributions are, as some people favor power-type distributions and others exponential-type distributions, although, as pointed out in [23, p. 1090], the power-type right tails cannot be used in models with continuous compounding as they lead to infinite expectation for the asset price. We stress that, quite surprisingly, it is very difficult to distinguish power-type tails from exponential-type tails and from empirical data unless one has extremely large sample size perhaps in the order of tens of thousands or even hundreds of thousands [19]. Therefore, it is very difficult to choose a good model based on the limited empirical data alone.

A good intuition may be obtained by simply looking at the quantiles for both standardized Laplace (with a symmetric density $f(x) = \frac{1}{2}e^{-x}I_{[x>0]} + \frac{1}{2}e^xI_{[x<0]}$) and standardized t distributions with mean 0 and variance 1. The right quantiles for the Laplace and normalized t densities with degrees of freedom (DOF) from 3 to 7 are given in Table 1.

This table shows that the Laplace distributions may have higher tail probabilities than t distributions, even if asymptotically the Laplace distributions should have lighter tails than t distributions. For example, regardless of the sample size, the Laplace distribution may appear to be heavier tailed than a t -distribution with DOF 6 or 7, up to the 99.9th percentile. To distinguish the distributions it is necessary to use quantiles with very low p values and correspondingly large sample sizes for statistical inference. If the true quantiles have to be estimated from data, then the problem is even worse, as the sample standard deviations need to be considered, resulting in

Table 1 The right quantiles of the Laplace and normalized t -distributions

Prob.	Laplace	t_7	t_6	t_5	t_4	t_3
1%	2.77	2.53	2.57	2.61	2.65	2.62
0.1%	4.39	4.04	4.25	4.57	5.07	5.90
0.01%	6.02	5.97	6.55	7.50	9.22	12.82
0.001%	7.65	8.54	9.82	12.04	16.50	27.67

sample sizes typically in the tens of thousands or even hundreds of thousands necessary to distinguish power-type tails from exponential-type tails. For further discussion, see [20], in which is also discussed the implication in terms of risk measured.

The difficulty in distinguishing tail behavior motivated Cai and Kou [5] to extend the double exponential jump-diffusion model to a hyperexponential jump-diffusion model, in which the jump size $\{Y_i := \log(V_i) : i = 1, 2, \dots\}$ is a sequence of i.i.d. hyperexponential random variables with density

$$f_Y(x) = \sum_{i=1}^m p_i \eta_i e^{-\eta_i x} I_{\{x \geq 0\}} + \sum_{j=1}^n q_j \theta_j e^{\theta_j x} I_{\{x < 0\}} \quad (6)$$

where $p_i > 0$ and $\eta_i > 1$ for all $i = 1, \dots, m$, $q_j > 0$ and $\theta_j > 0$ for all $j = 1, \dots, n$, and $\sum_{i=1}^m p_i + \sum_{j=1}^n q_j = 1$. Here the condition that $\eta_i > 1$, for all $i = 1, \dots, m$, is imposed to ensure that the stock price S_t has a finite expectation.

The hyperexponential distribution is general enough to provide a link between various heavy-tail distributions, no matter which ones we prefer. In particular, any completely monotone distribution, for example, with a density $f(x)$ satisfying the condition that all derivatives of $f(x)$ exist and $(-1)^n f^{(n)}(x) \geq 0$ for all x and $n \geq 1$, can be approximated by hyperexponential distributions as closely as possible in the sense of weak convergence. Many distributions with tails heavier than those of the normal distribution are completely monotone. Here are some examples of completely monotone distributions frequently used in finance:

1. *Gamma distribution.* The density of Gamma (α, β) is $x^{\alpha-1} e^{-\beta x}$, where $\alpha, \beta > 0$. When $\alpha < 1$, the distribution is completely monotone.
2. *Weibull distribution.* The cumulative distribution function of Weibull (c, d) is given by $1 - e^{-(x/d)^c}$, where $c, d > 0$. When $c < 2$, it has heavier tails than the normal distribution.
3. *Pareto distribution.* The distribution of Pareto (a, b) is given by $1 - (1 + bx)^{-a}$, where $a, b > 0$.
4. *Pareto mixture of exponential distribution (PME).* The density of PME (a, b) is given by $\int_0^{+\infty} f_{a,b}(y) y^{-1} e^{-x/y} dy$, where $f_{a,b}$ is the density of the Pareto (a, b) .

In summary, many heavy-tail distributions used in finance can be approximated arbitrarily closely

by the hyperexponential distribution. Feldmann and Whitt [15] develop a numerical algorithm to approximate completely monotone distributions by the hyperexponential distribution.

Cai and Kou [5] show that the hyperexponential jump-diffusion model can lead to analytical solutions for popular path-dependent options, such as lookback, barrier, and perpetual American options. These analytical solutions are made possible mainly because we solve several high-order integro-differential equations related to first passage time problems and optimal stopping problems explicitly. Solving the high-order integro-differential equations is the main technical contribution of [5], which is achieved by discovering a connection between integro-differential equations and homogeneous ordinary differential equations in the case of the hyperexponential jump-diffusion generator.

Multivariate Version

A significant drawback of most of the Levy processes discussed in the literature is that they are one dimensional, whereas many options traded in markets have several underlying assets. To overcome this, Huang and Kou [21] introduced a multivariate jump-diffusion model in which, under the physical measure P , the following stochastic differential equation is proposed to model the asset prices $S(t)$:

$$\frac{dS(t)}{S(t-)} = \mu dt + \sigma dW(t) + d \left(\sum_{i=1}^{N(t)} (V_i - 1) \right) \quad (7)$$

where $W(t)$ is an n -dimensional standard Brownian motion, $\sigma \in R^{n \times n}$ with the covariance matrix $\Sigma = \sigma \sigma^T$. The rate of the Poisson process $N(t)$ process is $\lambda = \lambda_c + \sum_{k=1}^n \lambda_k$; in other words, there are two types of jumps, common jumps for all assets with jump rate λ_c and individual jumps with rate λ_k , $1 \leq k \leq n$, only for the k th asset.

The logarithms of the common jumps have an m -dimensional asymmetric Laplace distribution $\mathcal{AL}_n(m_c, J_c)$, where $m_c = (m_{1,c}, \dots, m_{n,c})' \in R^n$ and $J_c \in R^{n \times n}$ is positive definite. For the individual jumps of the k th asset, the logarithms of the jump sizes follow a one-dimensional asymmetric Laplace distribution, $\mathcal{AL}_1(m_k, v_k^2)$. In summary,

$$Y = \log(V) \sim \begin{cases} \mathcal{AL}_n(m_c, J_c), & \text{with prob. } \lambda_c/\lambda \\ \underbrace{(0, \dots, 0, \mathcal{AL}_1(m_k, v_k^2))}_{k-1}, \underbrace{0, \dots, 0}_{n-k}, & \text{with prob. } \lambda_k/\lambda, \quad 1 \leq k \leq n \end{cases} \quad (8)$$

The sources of randomness, $N(t)$, $W(t)$ are assumed to be independent of the jump sizes V_i . Jumps at different times are assumed to be independent. Note that in the univariate case, the above model degenerates to the double exponential jump-diffusion model [23] but with $p\eta_1 = q\eta_2$.

In the special case of a two-dimensional model, the two-dimensional jump-diffusion return process $(X_1(t), X_2(t))$, with $X_i(t) = \log(S_i(t)/S(0))$, is given by

$$\begin{aligned} X_1(t) &= \mu_1 t + \sigma_1 W_1(t) + \sum_{i=1}^{N(t)} Y_i^{(1)} \\ X_2(t) &= \mu_2 t + \sigma_2 \left[\rho W_1(t) + \sqrt{1 - \rho^2} W_2(t) \right] \\ &\quad + \sum_{i=1}^{N(t)} Y_i^{(2)} \end{aligned} \quad (9)$$

Here all the parameters are risk-neutral parameters; $W_1(t)$ and $W_2(t)$ are two independent standard Brownian motions; and $N(t)$ is a Poisson process with rate $\lambda = \lambda_c + \lambda_1 + \lambda_2$. The distribution of the logarithm of the jump sizes Y_i is given by

$$\begin{aligned} Y_i &= (Y_i^{(1)}, Y_i^{(2)})' \\ &\sim \begin{cases} \mathcal{AL}_2(m_c, J_c), & \text{with prob. } \lambda_c/\lambda \\ (\mathcal{AL}_1(m_1, v_1^2), 0)', & \text{with prob. } \lambda_1/\lambda \\ (0, \mathcal{AL}_1(m_2, v_2^2))', & \text{with prob. } \lambda_2/\lambda \end{cases} \end{aligned} \quad (10)$$

where the parameters for the common jumps are

$$m_c = \begin{pmatrix} m_{1,c} \\ m_{2,c} \end{pmatrix} \quad \text{and} \quad J_c = \begin{pmatrix} v_{1,c}^2 & cv_{1,c}v_{2,c} \\ cv_{1,c}v_{2,c} & v_{2,c}^2 \end{pmatrix} \quad (11)$$

The infinitesimal generator of $\{X_1(t), X_2(t)\}$ is given by

$$\begin{aligned} \mathcal{L}u &= \mu_1 \frac{\partial u}{\partial x_1} + \mu_2 \frac{\partial u}{\partial x_2} \\ &\quad + \frac{1}{2} \sigma_1^2 \frac{\partial^2 u}{\partial x_1^2} + \frac{1}{2} \sigma_2^2 \frac{\partial^2 u}{\partial x_2^2} + \rho \sigma_1 \sigma_2 \frac{\partial^2 u}{\partial x_1 \partial x_2} \\ &\quad + \lambda_c \int_{y_2=-\infty}^{\infty} \int_{y_1=-\infty}^{\infty} [u(x_1+y_1, x_2+y_2) - u(x_1, x_2)] \\ &\quad \times f_{(Y^{(1)}, Y^{(2)})}^c(y_1, y_2) dy_1 dy_2 \\ &\quad + \lambda_1 \int_{y_1=-\infty}^{\infty} [u(x_1+y_1, x_2) - u(x_1, x_2)] f_{Y^{(1)}}(y_1) dy_1 \\ &\quad + \lambda_2 \int_{y_2=-\infty}^{\infty} [u(x_1, x_2+y_2) - u(x_1, x_2)] \\ &\quad \times f_{Y^{(2)}}(y_2) dy_2 \end{aligned} \quad (12)$$

for all continuous twice differentiable function $u(x_1, x_2)$, where $f_{(Y^{(1)}, Y^{(2)})}^c(y_1, y_2)$ is the joint density of correlated common jumps $\mathcal{AL}_2(m_c, J_c)$, and $f_{Y^{(i)}}(y_i)$ is the individual jump density of $\mathcal{AL}_1(m_i, J_i)$, $i = 1, 2$.

One difficulty in studying the generator is that the joint density of the asymmetric Laplace distribution has no analytical expression. Therefore, the calculation related to the joint density and generator becomes complicated. See [21] for change of measures from a physical measure to a risk-neutral measure, analytical solutions for the first passage times, and pricing formulae for barrier and exchange options.

References

- [1] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck based models and some of their uses in financial economics (with discussion), *Journal of Royal Statistical Society, Series B* **63**, 167–241.
- [2] Boyarchenko, S. & Levendorskii, S. (2002). *Non-Gaussian Merton-Black-Scholes Theory*, World Scientific, Singapore.

- [3] Boyle, P., Broadie, M. & Glasserman, P. (1997). Monte Carlo methods for security pricing, *Journal of Economic Dynamics and Control* **21**(89), 1267–1321.
- [4] Cai, N., Chen, N. & Wan, X. (2008). *Pricing Double Barrier Options Under a Flexible Jump Diffusion Model*, Hong Kong University of Science and Technology. Preprint.
- [5] Cai, N. & Kou, S.G. (2008). *Option Pricing Under a HyperExponential Jump Diffusion Model*, Columbia University. Preprint.
- [6] Carr, P., Geman, H., Madan, D. & Yor, M. (2002). The fine structure of asset returns: an empirical investigation, *Journal of Business* **75**, 305–332.
- [7] Carr, P., Geman, H., Madan, D. & Yor, M. (2003). Stochastic volatility for Lévy processes, *Mathematical Finance* **13**, 345–382.
- [8] Carr, P. & Wu, L. (2004). Time-changed lévy processes and option pricing, *Journal of Financial Economics* **71**, 113–141.
- [9] Chen, N. & Kou, S.G. (2005). Credit spreads, optimal capital structure, and implied volatility with endogenous default and jump risk, *Mathematical Finance* Preprint, Columbia University. To appear.
- [10] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, 2nd Printing, Chapman & Hall/CRC Press, London.
- [11] Cont, R. & Voltchkova, E. (2005). Finite difference methods for option pricing in jump-diffusion and exponential Lévy models, *SIAM Journal of Numerical Analysis* **43**, 1596–1626.
- [12] d’Halluin, Y., Forsyth, P.A. & Vetzal, K.R. (2003). *Robust Numerical Methods for Contingent Claims under Jump-diffusion Processes*, Working paper, University of Waterloo.
- [13] Duffie, D., Pan, J. & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions, *Econometrica* **68**, 1343–1376.
- [14] Engle, R. (1995). *ARCH: Selected Readings*, Oxford University Press.
- [15] Feldmann, A. & Whitt, W. (1998). Fitting mixtures of exponentials to long-tail distributions to analyze network performance models, *Performance Evaluation* **31**, 245–279.
- [16] Feng, L., Kovalov, P., Linetsky, V. & Marozzi, M. (2007). Variational methods in derivatives pricing, in *Handbook of Financial Engineering*, J. Birge & V. Linetsky, eds, Elsevier, Amsterdam.
- [17] Feng, L. & Linetsky, V. (2008). Pricing options in jump-diffusion models: an extrapolation approach, *Operations Research* **52**, 304–325.
- [18] Glasserman, P. & Kou, S.G. (2003). The term structure of simple forward rates with jump risk, *Mathematical Finance* **13**, 383–410.
- [19] Heyde, C.C. & Kou, S.G. (2004). On the controversy over tailweight of distributions, *Operations Research Letters* **32**, 399–408.
- [20] Heyde, C.C., Kou, S.G. & Peng, X.H. (2008). *What is a Good Risk Measure: Bridging the Gaps Between Robustness, Subadditivity, Prospect Theory, and Insurance Risk Measures*, Columbia University. Preprint.
- [21] Huang, Z. & Kou, S.G. (2006). *First Passage Times and Analytical Solutions for Options on Two Assets with Jump Risk*, Columbia University. Preprint.
- [22] Kijima, M. (2002). *Stochastic Processes with Applications to Finance*, Chapman & Hall, London.
- [23] Kou, S.G. (2002). A jump-diffusion model for option pricing, *Management Science* **48**, 1086–1101.
- [24] Kou, S.G., Petrella, G. & Wang, H. (2005). Pricing path-dependent options with jump risk via Laplace transforms, *Kyoto Economic Review* **74**, 1–23.
- [25] Kou, S.G. & Wang, H. (2003). First passage time of a jump diffusion process, *Advances in Applied Probability* **35**, 504–531.
- [26] Kou, S.G. & Wang, H. (2004). Option pricing under a double exponential jump-diffusion model, *Management Science* **50**, 1178–1192.
- [27] Lucas, R.E. (1978). Asset prices in an exchange economy, *Econometrica* **46**, 1429–1445.
- [28] Merton, R.C. (1976). Option pricing when underlying stock returns are discontinuous, *Journal of Financial Economics* **3**, 125–144.
- [29] Ramezani, C.A. and Zeng, Y. (2002). *Maximum Likelihood Estimation of Asymmetric Jump-Diffusion Process: Application to Security Prices*, Working Paper, Department of Mathematics and Statistics, University of Missouri, Kansas City.
- [30] Sepp, A. (2004). Analytical pricing of double-barrier options under a double exponential jump diffusion process: applications of Laplace transform, *International Journal of Theoretical and Applied Finance* **7**, 151–175.
- [31] Singleton, K. (2006). *Empirical Dynamic Asset Pricing*, Princeton University Press.
- [32] Stokey, N.L. & Lucas, R.E. (1989). *Recursive Methods in Economic Dynamics*, Harvard University Press.

Further Reading

Hull, J. (2005). *Options, Futures, and Other Derivatives*, Prentice Hall.

Related Articles

Barrier Options; Exponential Lévy Models; Jump Processes; Lookback Options; Partial Integro-differential Equations (PIDEs); Wiener–Hopf Decomposition.

STEVEN KOU

Exponential Lévy Models

Exponential Lévy models generalize the classical Black and Scholes model by allowing the stock prices to jump while preserving the independence and stationarity of returns. There are ample reasons for introducing jumps in financial modeling. First, asset prices exhibit jumps, and the associated risks cannot be handled within continuous-path models. Second, the well-documented phenomenon of *implied volatility smile* in option markets shows that the risk-neutral returns are non-Gaussian and leptokurtic, all the more so for short maturities, a clear indication of the presence of jumps. In continuous-path models, the law of returns for shorter maturities becomes closer to the Gaussian law, whereas in reality and in models with jumps, returns actually become less Gaussian as the horizon becomes shorter. Finally, jump processes correspond to genuinely incomplete markets, whereas all continuous-path models are either complete or can be made so with a small number of additional assets. This fundamental incompleteness makes it possible to carry out a rigorous analysis of the hedging errors in discontinuous models and find ways to improve the hedging performance using additional instruments such as liquid European options.

Lévy Processes

Lévy processes (*see Fundamental Theorem of Asset Pricing*) [1, 3, 17] are stochastic processes with stationary and independent increments. The only Lévy process with continuous trajectories is the Brownian motion with drift; all others have paths with discontinuities in finite or (countably) infinite number. The simplest example of a Lévy process is the Poisson process (*see Poisson Process*): the increasing piecewise constant process with jumps of size 1 only and exponential waiting times between jumps. If (τ_i) is a sequence of independent exponential random variables with intensity λ and $T_k := \sum_{i=1}^k \tau_i$, then the process

$$Z_t := \sum_i 1_{T_i \leq t} \quad (1)$$

is called a *Poisson process* with intensity λ . A piecewise constant Lévy process with arbitrary jump sizes

is called *compound Poisson* and can be written as

$$X_t = \sum_{k=1}^{Z_t} Y_k \quad (2)$$

where Z is a Poisson process and (Y_i) is an i.i.d. sequence of random variables. In general, the number of jumps of a Lévy process in a given interval need not be finite, and the process can be represented as a sum of a Brownian motion with drift and a limit of processes of the form in equation (2):

$$X_t = \gamma t + B_t + N_t + \lim_{\varepsilon \downarrow 0} M_t^\varepsilon \quad (3)$$

where B is a d -dimensional Brownian motion, $\gamma \in \mathbb{R}^d$, N is a compound Poisson process that includes the jumps of X with $|\Delta X_t| > 1$, and M_t^ε is a compensated compound Poisson process (compound Poisson minus its expectation) that includes the jumps of X with $\varepsilon < |\Delta X_t| \leq 1$. The law of a Lévy process is completely identified by its characteristic triplet—the positive definite matrix A (unit covariance of B), the vector γ (drift), and the measure ν on $\mathbb{R}^d$, called the *Lévy measure*, which determines the intensity of jumps of different sizes. $\nu(A)$ is the expected number of jumps on the time interval $[0, 1]$, whose sizes fall in A . The Lévy measure satisfies the integrability condition

$$\int_{\mathbb{R}^d} 1 \wedge \|x\|^2 \nu(dx) < \infty \quad (4)$$

and $\nu(\mathbb{R}) < \infty$ if the process has finite jump intensity. The law of X_t at all times t is determined by the triplet and, in particular, the Lévy–Khintchine formula gives the characteristic function $E[e^{iuX_t}] = \exp[t\psi(u)]$ with

$$\begin{aligned} \psi(u) = & i \langle \gamma, u \rangle + \frac{1}{2} \langle Au, u \rangle + \int_{\mathbb{R}^d} (e^{i \langle u, x \rangle} - 1 \\ & - i \langle u, x \rangle 1_{\|x\| \leq 1}) \nu(dx) \end{aligned} \quad (5)$$

Conversely, any infinitely divisible law (*see Infinite Divisibility*) has a Lévy–Khintchine representation as above, so modeling with Lévy processes allows to pick any infinitely divisible distribution for the law (say, at time $t = 1$) of the process.

Exponential Lévy Models

The Black–Scholes model

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t \quad (6)$$

can be equivalently rewritten in the exponential form $S_t = S_0 e^{(\mu - \sigma^2/2)t + \sigma W_t}$. This gives us two possibilities to construct an exponential Lévy model starting from a (one-dimensional) Lévy process X , using the stochastic differential equation

$$\frac{dS_t}{S_{t-}} = dX_t \quad (7)$$

or using the ordinary exponential $S_t = S_0 e^{X_t}$. The solution to equation (7) with initial condition $S_0 = 1$ is called the *stochastic exponential* of X . It can become negative if the process X has a big negative jump: $\Delta X_s < -1$ for $s \leq t$. However, if X does not have jumps of size smaller than -1 , then its stochastic exponential is positive, and the stochastic and the ordinary exponential yield the same class of positive processes. Given this result and the fact that ordinary exponentials are more tractable (in particular, we have the Lévy–Khintchine representation), they are more often used for modeling financial time series than the stochastic ones. In the rest of this article, we focus on the exponential Lévy model

$$S_t = S_0 e^{rt + X_t} \quad (8)$$

where X is a one-dimensional Lévy process with characteristic triplet (σ^2, ν, γ) and r denotes the interest rate.

Examples

Exponential Lévy models fall into two categories. In the first category, called *jump-diffusion* models, the “normal” evolution of prices is given by a diffusion process, punctuated by jumps at random intervals. Here the jumps represent rare events—crashes and large drawdowns. Such an evolution can be represented by a Lévy process with a nonzero Gaussian component and a jump part with finitely many jumps:

$$X_t = \gamma t + \sigma W_t + \sum_{i=1}^{N_t} Y_i \quad (9)$$

In the *Merton model* (see **Jump-diffusion Models**) [16], which is the first model of this type, suggested in the literature, jumps in the log price X are assumed to have a Gaussian distribution: $Y_i \sim N(\mu, \delta^2)$. In the risk-neutral version (i.e., with the choice of drift such that e^X becomes a martingale), the characteristic exponent of the log stock price takes the following form:

$$\begin{aligned} \psi(u) = & -\frac{\sigma^2 u^2}{2} + \lambda \{e^{-\delta^2 u^2/2 + i\mu u} - 1\} \\ & - iu \left(\frac{\sigma^2}{2} + \lambda(e^{\delta^2/2 + \mu} - 1) \right) \end{aligned} \quad (10)$$

In the *Kou model* (see **Kou Model**) [13], jump sizes are distributed according to an asymmetric Laplace law with a density of the form

$$\nu_0(x) = [p\lambda_+ e^{-\lambda_+ x} 1_{x>0} + (1-p)\lambda_- e^{-\lambda_- |x|} 1_{x<0}] \quad (11)$$

with $\lambda_+ > 0$, $\lambda_- > 0$ governing the decay of the tails for the distribution of positive and negative jump sizes and $p \in [0, 1]$ representing the probability of an upward jump. The probability distribution of returns in this model has semiheavy (exponential) tails.

The second category consists of models with an infinite number of jumps in every interval, which we call *infinite activity* or *infinite intensity* models. In these models, one does not need to introduce a Brownian component since the dynamics of jumps is already rich enough to generate nontrivial small time behavior [4].

There are several ways to define a parametric Lévy process with infinite jump intensity. The first approach is to obtain a Lévy process by subordinating a Brownian motion with an independent increasing Lévy process (called *subordinator*). Two examples of models from this class are the variance gamma process and the normal inverse Gaussian process. The variance gamma process (see **Variance-gamma Model**) [5, 15] is obtained by time changing a Brownian motion with a gamma subordinator and has the characteristic exponent of the form

$$\psi(u) = -\frac{1}{\kappa} \log \left(1 + \frac{u^2 \sigma^2 \kappa}{2} - i\theta \kappa u \right) \quad (12)$$

The density of the Lévy measure of the variance gamma process is given by

$$\nu(x) = \frac{c}{|x|} e^{-\lambda_- |x|} 1_{x < 0} + \frac{c}{x} e^{-\lambda_+ x} 1_{x > 0} \quad (13)$$

where $c = 1/\kappa$, $\lambda_+ = \frac{\sqrt{\theta^2 + 2\sigma^2/\kappa}}{\sigma^2} - \frac{\theta}{\sigma^2}$ and $\lambda_- = \frac{\sqrt{\theta^2 + 2\sigma^2/\kappa}}{\sigma^2} + \frac{\theta}{\sigma^2}$.

The normal inverse Gaussian process (see **Normal Inverse Gaussian Model**) [2] is the result of time changing a Brownian motion with the inverse Gaussian subordinator and has the characteristic exponent

$$\psi(u) = \frac{1}{\kappa} - \frac{1}{\kappa} \sqrt{1 + u^2 \sigma^2 \kappa - 2i\theta u \kappa} \quad (14)$$

The second approach is to specify the Lévy measure directly. The main example of this category is the tempered stable process (see **Tempered Stable Process**), introduced by Koponen [12] and also known under the name of *CGMY model* [4]. This process has a Lévy measure with density of the form

$$\nu(x) = \frac{c_-}{|x|^{1+\alpha_-}} e^{-\lambda_- |x|} 1_{x < 0} + \frac{c_+}{x^{1+\alpha_+}} e^{-\lambda_+ x} 1_{x > 0} \quad (15)$$

with $\alpha_+ < 2$ and $\alpha_- < 2$.

The third approach is to specify the density of increments of the process at a given time scale, say Δ , by taking an arbitrary infinitely divisible distribution. Generalized hyperbolic processes (see **Generalized Hyperbolic Models**) [10] can be constructed in this way. In this approach, it is easy to simulate the increments of the process at the same time scale and to estimate parameters of the distribution if data are sampled with the same period Δ , but, unless this distribution belongs to some parametric class closed under convolution, we do not know the law of the increments at other time scales.

Market Incompleteness and Option Pricing

The exponential Lévy models correspond, in general, to arbitrage-free incomplete markets, meaning that options cannot be replicated exactly and,

consequently, their price is not uniquely determined by the law of the underlying. This is good news: this means that the pricing model can be adjusted to take into account both the historical dynamics of the underlying and the market-quoted prices of European call and put options, a procedure known as *model calibration* (see **Model Calibration**). Once the risk-neutral measure Q is calibrated, one can price an exotic option with payoff H_T at time T by taking the discounted expectation

$$P_0 = e^{-rT} E^Q[H_T] \quad (16)$$

Fourier Transform Methods for Option Pricing and Model Calibration

In exponential Lévy models, and in all models where the characteristic function of the log stock price $\Phi_t(u) = E[e^{iuX_t}]$ is known explicitly, Fourier inversion provides a very efficient algorithm for pricing European options. This method was introduced in [5] and later improved and generalized in [14].

Consider a financial model of the form $S_t = S_0 e^{rt + X_t}$, where X is stochastic process whose characteristic function is known explicitly. To compute the price of a call option,

$$C(k) = S_0 E[(e^{X_T} - e^k)^+] \quad (17)$$

where $k = \log(K/S_0) - rT$ is the log forward moneyness, we would like to express its Fourier transform in terms of the characteristic function of X_T and then find the prices for a range of strikes by Fourier inversion. However, the Fourier transform of $C(k)$ is not well defined because this function is not integrable, so we subtract the Black–Scholes call price with nonzero volatility Σ to obtain a function that is both integrable and smooth:

$$z_T(k) = C(k) - C_{BS}^\Sigma(k) \quad (18)$$

If X is a stochastic process such that $E[e^{X_T}] = 1$ and $E[e^{(1+\alpha)X_T}] < \infty$ for some $\alpha > 0$, then the Fourier transform of $z_T(k)$ is given by

$$\zeta_T(v) = S_0 \frac{\Phi_T(v-i) - \Phi_T^\Sigma(v-i)}{iv(1+iv)} \quad (19)$$

where $\Phi_T^\Sigma(v) = \exp(-\frac{\Sigma^2 T}{2}(v^2 + iv))$ is the characteristic function of log stock price in the Black–

Scholes model with volatility Σ . The exact value of Σ is not very important, and one can take, for example, $\Sigma = 0.2$ for practical calculations.

Option prices are computed by evaluating numerically the inverse Fourier transform of ζ_T :

$$z_T(k) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-ivk} \zeta_T(v) dv \quad (20)$$

This integral can be efficiently computed for a range of strikes using the fast Fourier transform algorithm.

The Fourier-based fast deterministic algorithms for European option pricing can be used to calibrate exponential Lévy models to market-quoted option prices by penalized least squares as in [7]. Exponential Lévy models perform well for calibrating market option prices for a range of strikes and a single maturity, but fail to calibrate the entire implied volatility surface containing many maturities. This is due to the fact that the law of a Lévy process is completely determined by its distribution at a given date, so that if we know option prices for many strikes and a single maturity, we can readily reconstruct the law of the process at all dates, which may be incompatible with the other observations we may have. In particular, the implied volatility smile in exponential Lévy model flattens too fast for long-dated options (see Figure 1). Usually, a jump component can be included in a model to calibrate the short-maturity prices, and a stochastic volatility component is used to calibrate the skew at longer maturities.

PIDE Methods for Exotic Options

For contracts with barriers or American-style exercise, partial integro-differential equation (PIDE) methods provide an efficient alternative to Monte Carlo simulation. In diffusion models, the price of an option with payoff $h(S_T)$ at time T solves the Black–Scholes partial differential equation (PDE)

$$\begin{aligned} \frac{\partial P}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 P}{\partial S^2} &= rP - rS \frac{\partial P}{\partial S} \\ P(T, S) &= h(S) \end{aligned} \quad (21)$$

In an exponential Lévy model, there is a similar equation for the option price

$$P(t, S) = e^{-r(T-t)} E^Q[h(S_T) | S_t = S] \quad (22)$$

but due to the presence of jumps, an integral term appears in addition to the partial derivatives (see **Partial Integro-differential Equations (PIDEs)**):

$$\begin{aligned} \frac{\partial P}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 P}{\partial S^2} - rP + rS \frac{\partial P}{\partial S} \\ + \int_{\mathbb{R}} v(dz) \left\{ P(t, Se^z) - P(t, S) \right. \\ \left. - S(e^z - 1) \frac{\partial P}{\partial S}(t, S) \right\} = 0, \quad P(T, S) = h(S) \end{aligned} \quad (23)$$

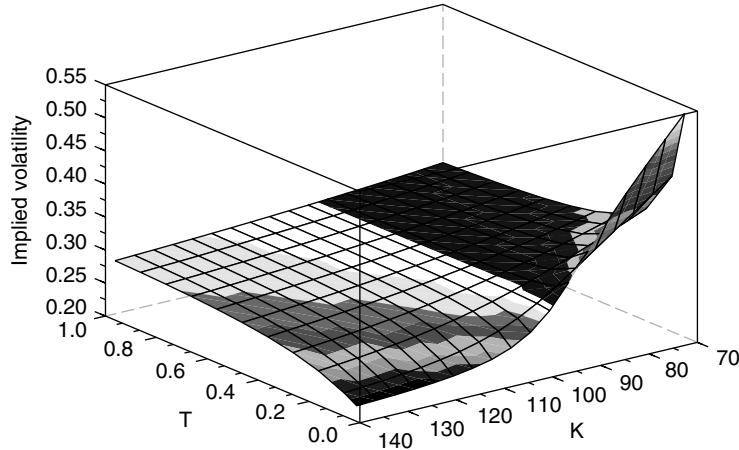


Figure 1 Implied volatility surface in the Kou model with diffusion volatility $\sigma = 0.2$ and only negative jumps with intensity $\lambda = 10$ and average size $\frac{1}{\lambda_-} = 0.05$

Different path-dependent characteristics of the payoff are translated into the boundary conditions of the equation: for example, for a down-and-out option with barrier B , we would impose $P(t, S) = 0$ for $S \leq B$ and all t . This equation and its numerical solution using finite differences is discussed in detail in [9] (see **Partial Integro-differential Equations (PIDEs)**).

Hedging

In the Black–Scholes model, delta hedging is known to completely eliminate the risk of an option position. In the presence of jumps, delta hedging is no longer optimal: to hedge a jump of a given size, one should use the sensitivity to fluctuations of this particular size rather than the sensitivity to infinitesimal movements. Since the jump size is not known in advance, the risk associated with jumps cannot be hedged away completely. The model given by equation (8) therefore corresponds to an incomplete market except for the following two cases:

- no jumps in the stock price ($\nu \equiv 0$, the Black–Scholes case) and
- no diffusion component ($\sigma = 0$) and only one possible jump size ($\nu = \delta_{z_0}(z)$). In this case, the optimal hedging strategy is

$$\phi_t = \frac{P(S_t e^{z_0}) - P(S_t)}{S_t(e^{z_0} - 1)} \quad (24)$$

In all other cases, the hedging becomes an approximation problem: instead of *replicating* an option, one tries to *minimize* the residual hedging error. Many authors (see, e.g. [8, 11]) studied the quadratic hedging, where the optimal strategy is obtained by minimizing the expected squared hedging error. A particularly simple situation is when this error is computed under the martingale probability. The optimal hedge is then a weighted sum of the sensitivity of option price to infinitesimal stock movements, and the average sensitivity to jumps:

$$\begin{aligned} & \phi^*(t, S_t) \\ &= \frac{\sigma^2 \frac{\partial P}{\partial S} + \frac{1}{S_t} \int \nu(dz)(e^z - 1)(P(t, S_t e^z) - P(t, S_t))}{\sigma^2 + \int (e^z - 1)^2 \nu(dz)} \end{aligned} \quad (25)$$

For jump diffusions, if jumps are small, the Taylor decomposition of this formula gives

$$\begin{aligned} \phi_t &\approx \frac{\partial P}{\partial S} + \frac{S_t}{2\Sigma^2} \frac{\partial^2 P}{\partial S^2} \int \nu(dz)(e^z - 1)^3 \\ \Sigma^2 &= \sigma^2 + \int (e^z - 1)^2 \nu(dz) \end{aligned} \quad (26)$$

Therefore, the optimal strategy can be seen as a small and typically negative (since the jumps are mostly negative) correction to delta hedging. For pure-jump processes such as variance gamma, $(\partial^2 P / \partial S^2)$ may not be defined and the correction may be big.

Numerical studies of the performance of hedging strategies in the presence of jumps show that

- If the jumps are small, delta hedging works well and its performance is close to optimal.
- In the presence of a strong jump component, the optimal strategy is superior to delta hedging both in terms of hedge stability and residual error.
- If jumps are strong, the residual hedging error can be further reduced by adding options to the hedging portfolio.

To eliminate the remaining hedging error, a possible solution is to use liquid options as hedging instruments. Optimal quadratic hedge ratios in the case when the hedging portfolio may contain options can be found in [8].

Additional Reading

For a more in-depth treatment, the reader may refer to the monographs [6, 18].

References

- [1] Appelbaum, D. (2004). *Lévy Processes and Stochastic Calculus*, Cambridge University Press.
- [2] Barndorff-Nielsen, O. (1998). Processes of normal inverse Gaussian type, *Finance and Stochastics* **2**, 41–68.
- [3] Bertoin, J. (1996). *Lévy Processes*, Cambridge University Press, Cambridge.
- [4] Carr, P., Geman, H., Madan, D. & Yor, M. (2002). The fine structure of asset returns: an empirical investigation, *Journal of Business* **75**, 305–332.
- [5] Carr, P. & Madan, D. (1998). Option valuation using the fast Fourier transform, *Journal of Computational Finance* **2**, 61–73.

- [6] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, Chapman & Hall/CRC Press.
- [7] Cont, R. & Tankov, P. (2006). Retrieving Lévy processes from option prices: regularization of an ill-posed inverse problem, *SIAM Journal on Control and Optimization* **45**, 1–25.
- [8] Cont, R., Tankov, P. & Voltchkova, E. (2007). Hedging with options in models with jumps. *Proceedings of the 2005 Abel Symposium in Honor of Kiyosi Itô*, F.E. Benth, G. Di Nunno, T. Lindstrom, B. Øksendal & T. Zhang, eds, Springer, pp. 197–218.
- [9] Cont, R. & Voltchkova E. (2005). A finite difference scheme for option pricing in jump-diffusion and exponential Lévy models, *SIAM Journal on Numerical Analysis* **43**, 1596–1626.
- [10] Eberlein, E. (2001). Applications of generalized hyperbolic Lévy motion to Finance, in *Lévy Processes—Theory and Applications*, O. Barndorff-Nielsen, T. Mikosch & S. Resnick, eds, Birkhäuser, Boston, pp. 319–336.
- [11] Kallsen, J., Hubalek, F. & Krawczyk, L. (2006). Variance-optimal hedging for processes with stationary independent increments, *The Annals of Applied Probability* **16**, 853–885.
- [12] Koponen, I. (1995). Analytic approach to the problem of convergence of truncated Lévy flights towards the Gaussian stochastic process, *Physical Review E* **52**, 1197–1199.
- [13] Kou, S. (2002). A jump-diffusion model for option pricing, *Management Science* **48**, 1086–1101.
- [14] Lee, R.W. (2004). Option pricing by transform methods: extensions, unification and error control, *Journal of Computational Finance* **7**, 51–86.
- [15] Madan, D., Carr, P. & Chang, E. (1998). The variance gamma process and option pricing, *European Finance Review* **2**, 79–105.
- [16] Merton, R. (1976). Option pricing when underlying stock returns are discontinuous, *Journal Financial Economics* **3**, 125–144.
- [17] Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge.
- [18] Schoutens, W. (2003). *Lévy Processes in Finance: Pricing Financial Derivatives*, Wiley, New York.

Related Articles

Barndorff-Nielsen and Shephard (BNS) Models; Fourier Transform; Infinite Divisibility; Jump Processes; Jump-diffusion Models; Kou Model; Partial Integro-differential Equations (PIDEs); Tempered Stable Process; Time-changed Lévy Process; Tempered Stable Process.

PETER TANKOV

Uncertain Volatility Model

Black–Scholes and Realized Volatility

What happens when a trader uses the Black–Scholes ((BS) in the sequel) formula to dynamically hedge a call option at a given constant volatility while the realized volatility is not constant?

It is not difficult to show that the answer is the following: if the realized volatility is lower than the managing volatility, the corresponding profit and loss (P&L) will be nonnegative. Indeed, a simple, yet, clever application of Itô's formula shows us that the instantaneous P&L of being short a delta-hedged option reads

$$P\&L_t = \frac{1}{2} \Gamma S_t^2 \left[\sigma_t^2 dt - \left(\frac{dS_t}{S_t} \right)^2 \right] \quad (1)$$

where Γ is the gamma of the option (the second derivative with respect to the underlying, which is positive for a call option), and σ_t the spot volatility, for example, the volatility at which the option was sold and $\left(\frac{dS_t}{S_t} \right)^2$ represents the realized variance over the period $[t, t + dt]$. Note that this holds without any assumption on the realized volatility, which will certainly turn out to be nonconstant. This result is fundamental in practice: it allows traders to work with neither exact knowledge of the behavior of the volatility nor a more complex toolbox than the plain BS formula; an upper bound of the realized volatility is enough to grant a profit (conversely, a lower bound for option buyers). This way of handling the realized volatility with the BS formula is of historical importance in the option market. El Karoui, Jeanblanc, and Shreve have formalized it masterfully in [5].

Superhedging and the Uncertain Volatility Model (UVM)

The UVM Framework

Assume that you perform the previous strategy. You are certainly not alone in the market, and you wish you have the lowest possible selling price compatible with your risk aversion. In practice, on the derivatives desk (this is a big difference with the insurance

world where the risk is distributed among a large enough number of buyers), the risk aversion is total, meaning that your managing policy will aim at yielding a nonnegative P&L whatever the realized path. This approach is what is called the superhedging strategy (or superstrategy) approach to derivative pricing. Of course, the larger the set of the underlying scenarios (or paths) for which you want to have the superhedging property (*see Superhedging*), the higher the initial selling price. The first set that comes to mind is the set of paths associated with an unknown volatility, say between two boundary values $\sigma_{\min}$ and $\sigma_{\max}$. In other words, we look for the cheapest price at which we can sell and manage an option without any assumption on the volatility except that it lies in the $[\sigma_{\min}, \sigma_{\max}]$ range. This framework is the uncertain volatility model (UVM) introduced by Avellaneda *et al.* [2].

If you take a call option (or more generally a European option with convex payoff), the BS price at volatility $\sigma_{\max}$ is a good candidate. Indeed, it yields a superhedging strategy by result (1). And should the realized volatility be constantly $\sigma_{\max}$, then your P&L will be 0. It is easy to conclude from this that the BS $\sigma_{\max}$ price is the UVM selling price for an option with a convex payoff.

Now very often traders use strategies (butterflies, callspreads, etc.) which are not convex any longer. It is not at all easy to find a superstrategy in such cases. There is one exception; if you hedge at the selling time and do not rebalance your hedge before maturity, the cheapest price associated to such a strategy will be the value at the initial underlying value of the concave envelope of the payoff function. It is easy to see that this value corresponds to the total uncertainty case, or to the $[0, \infty]$ case in the UVM model. For a call option it will be the value of the underlying.

Black–Scholes–Barenblatt Equation

There come into play the seminal work [2] and independently [7]: Going back to equation (1), we are looking for a model with the property that the managing volatility is $\sigma_{\min}$ when the gamma is nonnegative, and $\sigma_{\max}$ in the converse situation. Should such a model exist, it will yield an optimal solution to the superhedging problem.

An easy way to approximate the optimal solution is to consider a tree (a trinomial tree, for instance) where the dependence upon the volatility lies in

the node probabilities and not in the tree grid. In the classical backward pricing scheme one can then choose the managing volatility according to the local convexity (since it is a trinomial tree, each node has three offshoot and so a convexity information) of the immediately forward price. Of course, it is not the convexity of the current price since we are calculating it, but the related error of replacing the current convexity by the forward one will certainly go to zero when the time step goes to zero.

The related continuous-time object is the Black–Scholes partial differential equation (PDE) where the second-order term is replaced by the following non-linear one

$$\frac{1}{2} S_t^2 (\sigma_{\max}^2 \Gamma^+ - \sigma_{\min}^2 \Gamma^-)$$

where, as usual, x^+ and x^- denote the positive and negative parts. This PDE has been named *Black–Scholes–Barenblatt* since it looks like the Barenblatt PDE occurring in porosity theory. More precisely, in case of no arbitrage, assume that the stock price dynamics satisfy $dS_t = S_t (r dt + \sigma_t dW_t)$, where W_t is a standard Brownian motion and r is the risk-free interest rate. This is valid under the class $\mathcal{P}$ of all the probability measures such that $\sigma_{\min} \leq \sigma_t \leq \sigma_{\max}$. Let Π_t denote the value of a derivative at time t written on S_t with maturity T and final payoff $\Phi(S_T)$; then at any time $0 \leq t \leq T$, we must have $W^-(t, S_t) \leq \Pi_t \leq W^+(t, S_t)$ where

$$\begin{aligned} W^-(t, S_t) &= \inf_{P \in \mathcal{P}} \mathbb{E}_t^P [e^{-r(T-t)} \Phi(S_T)] \\ W^+(t, S_t) &= \sup_{P \in \mathcal{P}} \mathbb{E}_t^P [e^{-r(T-t)} \Phi(S_T)] \end{aligned} \quad (2)$$

The two bounds satisfy the following nonlinear PDE, called the *Black–Scholes–Barenblatt equation* (which reduces to the classical BS one in the case $\sigma_{\min} = \sigma_t = \sigma_{\max}$):

$$\begin{aligned} \frac{\partial W^\pm}{\partial t} + r \left(\frac{\partial W^\pm}{\partial S} S - W^\pm \right) \\ + \frac{1}{2} \Sigma \left(\frac{\partial^2 W^\pm}{\partial S^2} \right) S^2 \frac{\partial^2 W^\pm}{\partial S^2} = 0 \end{aligned} \quad (3)$$

with the terminal condition

$$W^\pm(S, T) = \Phi(S_T) \quad (4)$$

where

$$\Sigma \left(\frac{\partial^2 W^+}{\partial S^2} \right) = \begin{cases} \sigma_{\max}^2 & \text{if } \frac{\partial^2 W^+}{\partial S^2} \geq 0 \\ \sigma_{\min}^2 & \text{if } \frac{\partial^2 W^+}{\partial S^2} < 0 \end{cases} \quad (5)$$

and

$$\Sigma \left(\frac{\partial^2 W^-}{\partial S^2} \right) = \begin{cases} \sigma_{\max}^2 & \text{if } \frac{\partial^2 W^-}{\partial S^2} \leq 0 \\ \sigma_{\min}^2 & \text{if } \frac{\partial^2 W^-}{\partial S^2} > 0 \end{cases} \quad (6)$$

Observe that in case Φ is convex, the BS price at volatility $\sigma_{\max}$ is convex for any time t , so that it solves the Black–Scholes–Barenblatt equation. Conversely, if Φ is concave, so is its BS price at volatility $\sigma_{\max}$ for any time t , which yields the unique solution to the Black–Scholes–Barenblatt equation.

Superstrategies and Stochastic Control

Note that this PDE is also a classical Hamilton–Jacobi–Bellman equation occurring in stochastic control theory. Indeed a related object of interest is the supremum of the risk-neutral prices over all the dynamics of volatility that satisfy the range property:

$$\sup_{P \in \mathcal{P}} \mathbb{E}^P f$$

where $\mathcal{P}$ is the set of risk-neutral probabilities, each of which corresponds to a volatility process with value at each time in $[\sigma_{\min}, \sigma_{\max}]$. In fact, such an object is not that easy to define in the classical probabilistic modeling framework, since two different volatility processes will typically yield mutually singular probability measures on the set of possible paths. A convenient framework is the stochastic control framework. In such a framework, the managing volatility being interpreted as a control, one tries to optimize a given expectation—the risk-neutral price in this case. It turns out that stochastic optimal control will yield the optimal superstrategy price.

Nevertheless, the connection between the superstrategy problem and stochastic control is not that obvious, and these need to be spelled out carefully in this respect. Recall that the stochastic control problem is the maximization of an expectation over a set of processes, whereas the superstrategy problem is

the almost sure domination of the option payoff at maturity by a hedging strategy.

Note that even in the UVM case, there are still plenty of open questions. In fact, a neat formulation of the superhedging problem is not a piece of cake. The issue is avoided in [2], handled partially in [7], and more formally in [8], where the model uncertainty is specified as a set of martingale probabilities on the canonical space, and also in [6]. Once this is done, a natural theoretical problem, given such a “model set”, is to find out a formula for the cheapest superhedging price. The supremum of the risk-neutral prices over all the probabilities of the set will in general be strictly smaller than the cheapest price, even if they match in the UVM setting. The precise property of the “model set” that makes this equality remains to be clarified. Some partial results in this direction, with progresses towards a general theorem, are available in [4], where the case of path-dependent payoffs in the UVM framework is also solved.

Lagrangian UVM

In practice, the UVM approach is easy to implement for standard options by using the tree scheme described above, for example. It can be extended in the same way for path-dependent options. Nevertheless, when the price pops up, the usual reaction of the trader or risk officer is that the price is too high, especially too high to explain the observed market price.

The fact that the price is high is a direct consequence of the total aversion approach in the superstrategy formulation, and also of the fact that the price corresponds to the worst-case scenario where the gamma changes signs exactly when the volatility switches regimes. This is a highly unlikely situation. To lower the price and fit in the traditional setting where one wants to fit the observed market price of liquid European calls and puts (so-called vanillas), Avellaneda, Levy, and Paras propose a constrained extension of the UVM model where the price of the complex products of the trader is handled within the UVM framework with the additional constraint of fitting the vanilla prices. By duality, this reduces to computing the UVM price for a portfolio parameterized by a Lagrangian multiplier and then minimizing the dual value function over the Lagrangian parameter. Mathematically speaking, let us consider an asset S_t and a payoff $\Phi(S_T)$. m European options with

payoffs $F_1(S_{T_1}), \dots, F_m(S_{T_m})$ with maybe different strikes and maturities are available for hedging; let $f_1, \dots, f_m$ be their respective market prices at the time of the valuation $t \leq \min(T, T_1, \dots, T_m)$. Consider now an agent who buys quantities $\lambda_1, \dots, \lambda_m$ of each option. His total cost of hedging then reads

$$\begin{aligned} \Pi(t, S_t, \lambda_1, \dots, \lambda_m) \\ = \sup_{P \in \mathcal{P}} \left\{ e^{-r(T-t)} \Phi(S_T) - \sum_{i=1}^m \lambda_i e^{-r(T_i-t)} F_i(S_{T_i}) \right\} \\ + \sum_{i=1}^m \lambda_i f_i \end{aligned} \quad (7)$$

where the supremum (sup) is calculated within the UVM framework as presented above, and we must specify a range $\Lambda_i^+ \leq \lambda_i \leq \Lambda_i^-$ ($\Lambda_i^\pm$ represent the quantities available on the market). The optimal hedge Π is then defined as the solution to the problem

$$\Pi^*(t, S_t) = \inf_{\lambda_1, \dots, \lambda_m} \Pi(t, S_t, \lambda_1, \dots, \lambda_m) \quad (8)$$

In fact, the first-order conditions read $\frac{\partial \Pi}{\partial \lambda_i} = \sum_{i=1}^m f_i - \mathbb{E}^{P^*}(e^{-r(T_i-t)} F_i(S_{T_i})) = 0$, where P^* realizes the sup above. These conditions exactly fit the model to observed market prices. The convexity of $\Pi(t, S_t, \lambda_1, \dots, \lambda_m)$ with respect to λ_i ensures that if a minimum exists, then it is unique.

This approach is very attractive from a theoretical point of view, but it is much harder to implement. The consistency of observed vanilla prices is a crucial step that is rarely met in practice. Even if numerous robust algorithms exist to handle the dual problem, their implementation is quite tricky. In fact, this constrained formulation implies a calibration property of the model, and the design of a stable and robust calibration algorithm is one of the greatest challenges in the field of financial derivatives.

The Curse of Nonlinearity

Another issue for a practitioner is the inherent nonlinearity of the UVM formulation. Most traditional models like BS, Heston, or Lévy-based models are linear models. The fact that an option price should depend on the whole portfolio of the trader is a no-brainer for risk officers, but this nonlinearity is a challenge for the modularity and the flexibility of

pricing systems. This is very often a no-go feature in practice.

The complexity of evaluating a portfolio in the UVM framework is real, as studied thoroughly by Avellaneda and Buff in [1]. Following [1], let us consider a portfolio with n options with payoffs $f_1, \dots, f_n$ and maturities $t_1, \dots, t_n$. The computational problem becomes tricky when the portfolio consists of barrier options. Indeed, this means that, at any time step, the portfolio we are trying to value might be different (in case the stock price has reached the barrier of any option) from the one at the previous time step. Because of the nonlinearity, a PDE specific to this portfolio has to be solved in this case. Avellaneda and Buff [1] addressed this very issue: a naive implementation would require solving the $2^n - 1$ nonlinear PDEs, each representing a subportfolio. They provide an algorithm to build the minimal number N_n of subportfolios (i.e., of nonlinear PDEs to solve) and show the following:

- If the initial portfolio consists of barrier (single or double) and vanilla options, then $N_n \leq \frac{n(n+1)}{2}$
- If the initial portfolio only consists of single barrier options (n_u up-and-out ones and $n_d = n - n_u$ down-and-out ones), then $N_n = n_d + n_u + n_d n_u$. This assumes that all the barriers are different. If some are identical, then the number of required computations decreases.

Numerically speaking, the finite-difference pricing is done on a lattice, matching almost exactly all the barriers. Nevertheless in [3], an optimal construction of the lattice to solve the PDEs is provided.

References

- [1] Avellaneda, M. & Buff, R. (1999). Combinatorial implications of nonlinear uncertain volatility models: the case of barrier options, *Applied Mathematical Finance* **1**, 1–18.
- [2] Avellaneda, M., Levy, A. & Paras, A. (1995). Pricing and hedging derivative securities in markets with uncertain volatilities, *Applied Mathematical Finance* **2**, 73–88.
- [3] Avellaneda, M. & Paras, A. (1996). Managing the volatility risk of portfolios of derivative securities: the Lagrangian uncertain volatility model, *Applied Mathematical Finance* **3**, 21–52.
- [4] Denis, L. & Martini, C. (2006). A theoretical framework for the pricing of contingent claims in the presence of model uncertainty, *Annals of Applied Probability* **16**(2), 827–852.
- [5] El Karoui, N., Jeanblanc, M. & Shreve, S. (1998). Robustness of the Black and Scholes formula, *Mathematical Finance* **8**(2), 92–126.
- [6] Frey, R. (2000). Superreplication in stochastic volatility models and optimal stopping, *Finance and Stochastics* **4**(2), 161–187.
- [7] Lyons, T.J. (1995). Uncertain volatility and the risk-free synthesis of derivatives, *Applied Mathematical Finance* **2**, 117–133.
- [8] Martini, C. (1997). Superreplications and stochastic control, *IIIrd Italian Conference on Mathematical Finance*, Trento.

Related Articles

Black–Scholes Formula; Models; Stochastic Control.

CLAUDE MARTINI & ANTOINE JACQUIER

Implied Volatility: Market Models

The market model approach for implied volatility consists in taking implied volatilities as the quantities one wishes to model. In many options exchanges and over-the-counter markets, implied volatility is the way an option is quoted and hence plays the role of a price.

The market model approach for implied volatilities is inspired by the corresponding market model approach for interest rates, the so-called Heath–Jarrow–Morton (HJM) approach to interest rate modeling (*see Heath–Jarrow–Morton Approach*). In interest rate modeling, it is simple to characterize the dynamics of the entire family of instantaneous forward rates in such a way that the corresponding family of bond prices is arbitrage free. It is also simple to give examples of such dynamics and to price interest rate sensitive contingent claims in such models.

Correspondingly, the market model approach to implied volatility seeks to characterize the dynamics of the entire implied volatility surface in such a way that the corresponding family of option prices is arbitrage free. It also seeks simple examples of such dynamics and ideally practical means of computing prices of exotic options on the underlying in such models.

Despite numerous attempts and recent progress in this area, it is fair to say that this approach has unfortunately not delivered the same elegant and useful results as the HJM approach has in interest rate modeling. The market approach to implied volatility can be traced back to the works [11, 12, 20].

As in the HJM approach, the no-arbitrage condition for implied volatilities takes the form of a drift restriction. In other words, the drift must be constrained for option prices to be local martingales under the pricing measure.

Drift Restriction

To continue the discussion, we need the following definitions. Let $(\mathbf{W}_t)_{t \geq 0}$ be an n -dimensional Wiener process that models the uncertainty in the economy. We shall use boldface letters for vectors, $\cdot$ for the

usual scalar product, and $|\cdot|$ for the Euclidean norm. We assume that the probability measure is risk neutral, that is, discounted price processes are local martingales. We assume for simplicity that interest rates and dividends are zero.

The traded asset on which options are written is denoted by S_t and its volatility vector by σ_t , that is,

$$\frac{dS_t}{S_t} = \sigma_t \cdot d\mathbf{W}_t \quad (1)$$

The no-arbitrage constraint for the implied volatility $\Sigma_t(T, K)$ for the option with strike K and maturity T implies the following drift restriction in their dynamics

$$\begin{aligned} \Sigma_t(T, K) &= \Sigma_0(T, K) - \int_0^t \left[\frac{|\sigma_s - \ln(S_s/K) \xi_s|^2 - \Sigma_s^2}{2\Sigma_s(T-s)} \right. \\ &\quad \left. + \frac{1}{2} \Sigma_s \sigma_s \cdot \xi_s - \frac{1}{8} \Sigma_s^3 (T-s) |\xi_s|^2 \right] (T, K) ds \\ &\quad + \int_0^t \Sigma_s(T, K) \xi_s(T, K) \cdot d\mathbf{W}_s \end{aligned} \quad (2)$$

where $\xi_t(T, K)$ is the implied volatility's volatility vector. The corresponding call option price dynamics is

$$\begin{aligned} C_t(T, K) &= C_0(T, K) + \int_0^t S_s (\Phi(d_1) \sigma_s \\ &\quad + \Sigma_s \sqrt{T-s} \varphi(d_1) \xi_s)(T, K) \cdot d\mathbf{W}_s \end{aligned} \quad (3)$$

where, as usual, $d_1 = (\ln(S_t/K)/\Sigma_t(T, K)\sqrt{T-t}) + \frac{1}{2} \Sigma_t(T, K)\sqrt{T-t}$ and Φ and φ denote, respectively, the cumulative distribution and density probability function of a standard Gaussian random variable.

The above equation (2) is the equivalent of the HJM equation for implied volatilities. However, unlike the HJM equation, the drift does not solely involve the volatility vector ξ_t but also depends on S_t and σ_t .

The Spot Volatility Specification

Equation (2) has interesting properties. In particular, when we consider the infinite system of equation (2)

for a fixed K and all $T > t$, it alone specifies the spot volatility σ_t . This phenomenon is directly related to the convergence of option prices to the option payoff at expiry. Equivalently, the solution to equation (2) should not blow up too fast near expiry. It is called *no-bubble restriction* in [17], whereas [5] calls it the *feedback condition* and traces it back to [8]. It is also called the *volatility specification* in [19] and [6]. It reads

$$\Sigma_t(t, K) = \left| \sigma_t - \ln \left(\frac{S_t}{K} \right) \xi_t(t, K) \right| \quad (4)$$

For a proof under proper assumptions, see [13].

The case where we let $K = S_t$ in equation (4) says $\Sigma_t(t, S_t) = |\sigma_t|$. In other words, the current value of the spot volatility can be exactly recovered from the implied volatility smile. This very much parallels the fact that the instantaneous forward rate with infinitely small tenor is the short rate in the HJM approach to interest rates.

It is shown in [14] that the relation $\Sigma_t(t, S_t) = |\sigma_t|$ holds in great generality even when jumps in the spot and/or its volatility are present. It turns out to be a consequence of the central limit theorem for martingales.

Equation (4) has an interesting connection to the work of Berestycki *et al.* [3]. In a time homogeneous stochastic volatility model, [3] shows that the implied volatility in the short maturity limit can be expressed using the geodesic distance associated with the generator of the bivariate diffusion (x_t, y_t) , where x_t is the log-moneyness and y_t is the spot volatility ($|\sigma_t|$ in our notation). Keeping their notation, we denote by $d(x, y)$ the signed geodesic distance from (x, y) to $(0, y)$, and obtain

$$\Sigma_t(t, K) = \frac{\ln(S_t/K)}{d(\ln(S_t/K), |\sigma_t|)} \quad (5)$$

By comparing equations (4) and (5), it becomes clear that the geodesic distance associated with the generator of the stochastic volatility model and the implied volatility's volatility vector are strongly related.

The Case of a Single Strike

We first deal with the problem studied by [1, 4, 16, 17, 19] where only a single option is considered. The goal is to set up conditions under which the infinite system of equation (2) for a fixed K and

all $T > t$ together with equation (1) admits a unique solution. The best results were obtained by [4] and [19]. Without loss of generality, one can assume that equation (1) is driven by the first Wiener process only and that $\sigma_t = (\sigma_t, 0, \dots, 0)$. Assume that ξ_t has the functional form,

$$\xi_t(T, K) = \frac{1}{2} \int_t^T \frac{X_t(u, K)}{X_t(T, K)} \mathbf{V}_t(u, K) du \quad (6)$$

where $X_t(T, K) = \frac{\partial}{\partial T}(\Sigma_t(T, K)^2(T - t))$ is the square of the *forward implied volatility* and where $\mathbf{V}$ has the form

$$\mathbf{V}_t(T, K) = \mathbf{V}(t, T, K, \Sigma_t(T, K), \Sigma_t(t, K), S_t) \quad (7)$$

for a deterministic function $\mathbf{V}$ satisfying technical positivity, growth, and Lipschitz conditions [19]. Assume also that the spot volatility has the functional form $\sigma_t = \sigma(t, K, \Sigma_t(t, K), S_t)$, where the deterministic function σ is determined by equation (4). Then, the infinite system of equation (2) for a fixed K and all $T > t$ together with equation (1) admits a unique solution.

The Case of Several Strikes

The infinite system of equation (2) for all K and all $T > t$ together with equation (1) is more complicated and conditions on ξ_t under which it admits a unique solution are still poorly understood. One advantage of dealing with all strikes K at once is that one can remove the dependence on S in equation (2) by changing the parameterization of the surface from K to moneyness K/S_t . The dynamics of the implied volatility surface in these coordinates are obtained by applying the Itô–Wentzell formula as in [5]. One of the difficulties of the multistrike case is that the solution to the infinite system in equation (2) must satisfy some shape restrictions at each time t . These are consequences of the well-known static arbitrage restrictions that we now recall.

Static Arbitrage Restrictions

Static arbitrage relations lead to constraints on the shape of the implied volatility surface. The fact that calendar spreads have positive values leads to

$$\frac{\partial \Sigma_t}{\partial T} + \frac{\Sigma_t}{2(T - t)} \geq 0 \quad (8)$$

The fact that call values are a decreasing function of the strike leads to

$$\frac{-\Phi(-d_1)}{\varphi(d_1)\sqrt{T-t}} \leq K \frac{\partial \Sigma_t}{\partial K} \leq \frac{\Phi(d_2)}{\varphi(d_2)\sqrt{T-t}} \quad (9)$$

Finally, the fact that butterfly spreads have positive values, or that calls are convex functions of the strike leads to

$$\begin{aligned} & \left(1 - \frac{\ln(K/S_t)}{\Sigma_t} K \frac{\partial \Sigma_t}{\partial K}\right)^2 - \frac{(T-t)^2}{4} \Sigma_t^2 \left(K \frac{\partial \Sigma_t}{\partial K}\right)^2 \\ & + (T-t) \Sigma_t \left(K \frac{\partial \Sigma_t}{\partial K} + K^2 \frac{\partial^2 \Sigma_t}{\partial K^2}\right) \geq 0 \end{aligned} \quad (10)$$

where $d_2 = d_1 - \Sigma_t(T, K)\sqrt{T-t}$. These restrictions must hold at each time t and at each point (T, K) of the implied volatility surface.

Deterministic Models

Practitioners [10] have proposed two simple models for implied volatility surfaces movements: the sticky strike model and the sticky delta model. The sticky strike model supposes that between date s and $t \geq s$, the implied volatility surface evolves as

$$\Sigma_t(T, K) = \Sigma_s(T, K) \quad (11)$$

whereas the sticky delta model supposes that

$$\Sigma_t(T, K) = \Sigma_s\left(T, \frac{S_s}{S_t} K\right) \quad (12)$$

In a sticky strike model, an option with a given strike has constant implied volatility. This contrasts with a sticky delta model where options with same moneyness have same implied volatilities. In other words, the implied volatility surface moves in perfect sync with the spot. In reality, implied volatilities move in a more complicated fashion but these two extreme cases are useful stylized benchmarks.

The sticky strike and sticky delta models, in fact, imply strong restrictions on the possible spot dynamics. Balland [2] showed that a sticky delta occurs if and only if the underlying asset price is the exponential of a process with independent increments (i.e., a Lévy process, *see Exponential Lévy Models*) under the pricing measure, and that a sticky strike situation occurs in the Black–Scholes model only!

Empirical Models

To overcome the obvious shortcomings of the sticky strike and sticky delta models, Cont and da Fonseca [9] have proposed to write down a model for the future evolution of the surface as an infinite system where each point of the surface is driven by a few common factors. These dynamics allow for easy calibration using principal component analysis [9] and can be useful for risk management and scenarios simulation. It is difficult, however, to check whether such specifications satisfy arbitrage restrictions, which prevents them from being used to price exotic options.

The Spot Volatility Dynamics from the Implied Volatility Surface

In the HJM approach (*see Heath–Jarrow–Morton Approach*) to interest rate modeling, the HJM equation can be used to write down the short rate dynamics starting from the forward rate dynamics. The parallel result in the case of implied volatility was obtained in [13]. The statement is the following: there exists a scalar Wiener process $W^\perp$ adapted to the filtration generated by $(\mathbf{W}_t)_{t \geq 0}$ such that

$$\begin{aligned} |\sigma_t|^2 = & |\sigma_0|^2 + \int_0^t \left[4 |\sigma_s| \frac{\partial \Sigma_s}{\partial T}(s, S_s) + 6 |\sigma_s|^2 \right. \\ & \times \left(S_s \frac{\partial \Sigma_s}{\partial K}(s, S_s) \right)^2 + 2 |\sigma_s|^3 S_s^2 \frac{\partial^2 \Sigma_s}{\partial K^2}(s, S_s) \Big] ds \\ & + \int_0^t 4 |\sigma_s| \frac{\partial \Sigma_s}{\partial K}(s, S_s) dS_s + \int_0^t 2 |\sigma_s|^2 \xi_s^\perp dW_s^\perp \end{aligned} \quad (13)$$

where

$$\begin{aligned} (\xi_t^\perp)^2 = & -\frac{2}{|\sigma_t|} \frac{d}{dt} \left\langle S_t, \frac{\partial \Sigma_t}{\partial K}(t, S_t) \right\rangle \\ & + 2 \left(S_t \frac{\partial \Sigma_t}{\partial K}(t, S_t) \right)^2 - |\sigma_t| S_t \frac{\partial \Sigma_t}{\partial K}(t, S_t) \\ & - 3 |\sigma_t| S_t^2 \frac{\partial^2 \Sigma_t}{\partial K^2}(t, S_t) \end{aligned}$$

Moreover, the two local martingales appearing in the decomposition are orthogonal in the sense that

$$\left\langle \int_0^t 4 |\sigma_s| \frac{\partial \Sigma_s}{\partial K}(s, S_s) dS_s; \int_0^t 2 |\sigma_s|^2 \xi_s^\perp dW_s^\perp \right\rangle = 0 \quad (14)$$

This result actually has a converse, which allows one to get a very precise idea of the implied volatility of a given spot model. It indeed allows to compute the first terms of the Taylor expansion of the implied volatility surface for short maturity and around the money [13].

Other Approaches

Modeling implied volatilities is equivalent to modeling option prices; as seen in equation (3), it is merely a parameterization of the options' volatilities. The difficulties in modeling implied volatilities have led researchers to look for other and possibly more tractable parameterizations. We mention them here, although these approaches depart from the strict study of implied volatilities.

First, following a program started in [7, 10] model option prices by modeling Dupire local volatility as a random field. They are able to find explicit drift conditions as well as some examples of such dynamics. The Dupire local volatility surface also specifies the spot volatility in the short maturity limit but does not have complicated static arbitrage restrictions like equations 8–10.

Another way of parameterizing option prices consists in modeling its intrinsic value, that is, the difference between the option price and the payoff if the option was exercised today. This is the approach taken by [15] in a very general semimartingale framework. Exactly as with implied volatilities, this approach yields a spot specification when options are close to maturity.

Finally, let us mention the recent work [18], where the authors introduce new quantities: the “local implied volatilities” and “price level” to parameterize option prices. These have nicer dynamics and naturally satisfy the static arbitrage conditions. They derive existence results for the infinite system of equations driving these quantities.

References

- [1] Babbar, K. (2001). *Aspects of Stochastic Implied Volatility in Financial Markets*. PhD thesis, Imperial College, London.
- [2] Bolland, P. (2002). Deterministic implied volatility models, *Quantitative Finance* **2**(2), 31–44.
- [3] Berestycki, H., Busca, J. & Florent, I. (2004). Computing the implied volatility in stochastic volatility models, *Communications on Pure and Applied Mathematics* **57**(10), 1352–1373.
- [4] Brace, A., Fabbri, G. & Goldys, B. (2007). *An Hilbert Space Approach for A Class of Arbitrage Free Implied Volatilities Models*, Technical report, Department of Statistics, University of New South Wales, at <http://arxiv.org/abs/0712.1343>.
- [5] Brace, A., Goldys, B., Klebaner, F. & Womersley, R. (2001). *Market Model for Stochastic Implied Volatility with Application to the BGM Model*, Technical report, Department of Statistics, University of New South Wales.
- [6] Carmona, R. (2007). HJM: a unified approach to dynamic models for fixed income, credit and equity markets, in *Paris-Princeton Lectures on Mathematical Finance 2004*, Lecture Notes in Mathematics, Springer, Vol. 1919.
- [7] Carmona, R. & Nadtochiy, S. (2009). Local volatility dynamic models, *Finance and Stochastics* **13**(1), 1–48.
- [8] Carr, P. (2000). *A Survey of Preference Free Option Valuation with Stochastic Volatility*, Risk's 5th annual European derivatives and risk management congress, Paris.
- [9] Cont, R. & Da Fonseca, J. (2002). Dynamics of implied volatility surfaces, *Quantitative Finance* **2**(2), 45–60.
- [10] Derman, E. (1999). Regimes of volatility, *Risk* (4), 55–59.
- [11] Derman, E. & Kani, I. (1998). Stochastic implied trees: arbitrage pricing with stochastic term and strike structure of volatility, *International Journal of Theoretical Applied Finance* **1**(1), 61–110.
- [12] Dupire, B. (1993). Model art, *Risk* **6**(9), 118–124.
- [13] Durrleman, V. (2004). *From Implied to Spot Volatilities*, PhD thesis, Department of Operations Research & Financial Engineering, Princeton University, at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1162425 to appear in *Finance and Stochastics*.
- [14] Durrleman, V. (2008). Convergence of at-the-money implied volatilities to the spot volatility, *Journal of Applied Probability* **45**, 542–550.
- [15] Jacod, J. and Protter, P. (2006). *Risk Neutral Compatibility with Option Prices*, Technical report, Université Paris VI and Cornell University, at <http://people.orie.cornell.edu/protter/WebPapers/JP-OptionPrice.pdf>.
- [16] Lyons, T. 1995. Uncertain volatility and the risk-free synthesis of derivatives, *Applied Mathematical Finance* (2), 117–133.

- [17] Schönbucher, P. (1999). A market model for stochastic implied volatility, *Philosophical Transactions of the Royal Society of London. Series A: Mathematical and Physical Sciences* **357**(1758), 2071–2092.
- [18] Schweizer, M. & Wissel, J. (2008). Arbitrage-free market models for option prices: the multi-strike case, *Finance and Stochastics* **12**(4), 469–505.
- [19] Schweizer, M. & Wissel, J. (2008). Term structures of implied volatilities: absence of arbitrage and existence results, *Mathematical Finance* **18**, 77–114.
- [20] Zhu, Y. & Avellaneda, M. (1998). A risk-neutral stochastic volatility model, *International Journal of Theoretical and Applied Finance* **1**(2), 289–310.

Further Reading

Gatheral, J. (2006). *The Volatility Surface: A Practitioner's Guide*, Wiley Finance.

Heath, D., Jarrow, R. & Morton, A. Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation, *Econometrica* **60**(1), 77–105.

Related Articles

Black–Scholes Formula; Dividend Modeling; Exponential Lévy Models; Heath–Jarrow–Morton Approach; Implied Volatility: Long Maturity Behavior; Implied Volatility: Large Strike Asymptotics; Implied Volatility: Volvol Expansion; Implied Volatility Surface; Implied Volatility in Stochastic Volatility Models; Local Volatility Model; Moment Explosions; SABR Model.

VALDO DURRLEMAN

Rating Transition Matrices

Rating transition matrices play an important role in credit risk management both as a method for summarizing the empirical behavior of a rating system and as a tool for computing probabilities of rating migrations in, for example, a portfolio of risky loans. Analysis of statistical properties of rating transition matrices is intimately linked with Markov chains. Even if rating processes in general are not Markovian, statistical analysis of rating systems often focuses on assessing a particular deviation from Markovian behavior. Furthermore, the tractability of the Markovian setting can be preserved in some simple extensions.

Discrete-time Markov Chains

Let the rating process $\eta = (\eta_0, \eta_1, \dots)$ be a discrete-time stochastic process taking values in a finite state space $\{1, \dots, K\}$. If the rating process is a Markov chain, the probability of making a particular transition between time t and time $t + 1$ does not depend on the history before time t , and one-step transition probabilities of the form

$$p_{ij}(t; t + 1) = \Pr(\eta_{t+1} = j \mid \eta_t = i) \quad (1)$$

describes the evolution of the chain. If the one-step transition probabilities are independent of time, we call the chain time homogeneous and write

$$p_{ij} = \Pr(\eta_{t+1} = j \mid \eta_t = i) \quad (2)$$

The one-period transition matrix of the chain is then given as

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1K} \\ \vdots & & \vdots \\ p_{K1} & \cdots & p_{KK} \end{pmatrix} \quad (3)$$

where $\sum_{j=1}^K p_{ij} = 1$ for all i .

Consider a sample of N firms whose transitions between different states are observed at discrete dates $t = 0, \dots, T$. Now introduce the following notation:

- $n_{ij}(t)$ = number of firms that went from i at date $t - 1$ to j at date t .
- $N_i(t) = \sum_{j=1}^K n_{ij}(t)$ = number of firm exposures recorded at the beginning of transition periods.
- $N_{ij}(T) = \sum_{t=1}^T n_{ij}(t)$ = total number of transitions observed from i to j over the entire period.

If we do not assume time homogeneity, we can estimate each element of the one-step transition probability matrix using the maximum-likelihood estimator

$$\widehat{p}_{ij}(t - 1; t) = \frac{n_{ij}(t)}{n_i(t - 1)} \quad (4)$$

which simply is the fraction of firms that made the transition divided by the number of firms which could have made the transition.

Assuming time homogeneity, the maximum-likelihood estimator of the transition probabilities matrix is

$$\widehat{p}_{ij} = \frac{N_{ij}(T)}{N_i(T)} \quad (5)$$

for all $i, j \in K$. This estimator is different from the estimator obtained by estimating a sequence of 1-year transition matrices and then computing the average of each element at a time. The latter method will weigh years with few observations as heavily as years with many observations. If the viewpoint is that there is variation in 1-year transition probabilities over time due to, for example, business cycle fluctuations, the averaging can be justified as a way of obtaining an unconditional 1-year default probability over the cycle.

Rating agencies often form a cohort of firms at a particular date, say January 1, 1980, and record transition frequencies over a fixed time horizon, say 5 years. This can be done in a straightforward way using only information on the initial rating and final rating after 5 years, assuming that all companies that are in the cohort, to begin with, stay in the sample. In practice, rating withdrawals occur, that is, firms or debt issues cease to have a rating. According to [4], the vast majority of withdrawals are due to debt maturing, being redeemed or called. It is traditional in the rating literature to view these events as “noninformative” censoring. One way to deal with withdrawals is to eliminate the firms from the sample and in essence use only those firms that do not have their rating withdrawn in the 5-year period. Another way is to estimate a sequence of

- $n_i(t)$ = number of firms in state i at date t .

1-year transition probability matrices using the 1-year estimator and then estimate the 5-year matrix as the product of 1-year matrices. In this case, information of a firm whose rating is withdrawn is used for the years where it is still present in the sample. Both methods rely on the assumption of withdrawals being noninformative.

Continuous-time Markov Chains

When one has access to full rating histories and therefore knows the exact dates of transitions, the continuous-time formulation offers significant advantages in terms of tractability. Recall that the family of transition matrices for a time-homogeneous Markov chain in continuous time on a finite state space can be described by an associated generator matrix, that is, a $K \times K$ matrix Λ , whose elements satisfy

$$\begin{aligned}\lambda_{ij} &\geq 0 \text{ for } i \neq j \\ \lambda_{ii} &= -\sum_{j \neq i} \lambda_{ij}\end{aligned}\quad (6)$$

Let $P(t)$ denote the $K \times K$ matrix of transition probabilities, that is, $p_{ij}(t) = P(\eta_t = j | \eta_0 = i)$. Then

$$P(t) = \exp(\Lambda t) \quad (7)$$

where the right hand side is the matrix exponential of the matrix Λt obtained by multiplying all entries of Λ by t .

In case a row consists of all zeros, the chain is absorbed in that state when it hits it. It is convenient to work with the default states as absorbing states even if firms in practice may recover and leave the default state. If we ask what the probability is that a firm will default before time T then this can be read from the transition matrix $P(T)$ when we have defined default to be an absorbing state. If the state is not absorbing, but P allows the chain to jump back into the nondefault rating categories, then the transition probability matrix for time T will only give the probability of being in default at time T and this (smaller) probability is typically not the one we are interested in for risk management purposes.

Assume that we have observed a collection of firms between time 0 and time T . The maximum-likelihood estimator for the off-diagonal elements of

the generator matrix is given by

$$\hat{\lambda}_{ij} = \frac{N_{ij}(T)}{\int_0^T Y_i(s) ds} \quad (8)$$

where $Y_i(s)$ is the number of firms in rating class i at time s and $N_{ij}(T)$ is the total number of direct transitions over the period from i to j , where $i \neq j$. The denominator counts the number of “firm-years” spent in state i .

Any period a firm spends in a state will be picked up through the denominator. In this sense all information is being used. Note also how (noninformative) censoring is handled automatically: When a firm leaves the sample, it simply stops contributing to the denominator. Also, this method will produce estimates of transition probabilities for “rare transitions”, even if the rare transitions have not been observed in the sample. For more on this, see [9].

Nonhomogeneous Chains

For statistical specifications and applications to pricing, the concept of a nonhomogeneous chain is useful. In complete analogy with the discrete-time case, the definition of the Markov property does not change when we drop the assumption of time homogeneity, but the description of the family of transition matrices requires that we keep track of calendar dates instead of just time lengths.

For each pair of states i, j with $i \neq j$, let A_{ij} be a nondecreasing right-continuous (and with left limits) function, which is zero at time zero. Let

$$A_{ii}(t) = -\sum_{j \neq i} A_{ij}(t) \quad (9)$$

and assume that

$$\Delta A_{ii}(t) \geq -1 \quad (10)$$

Then there exists a Markov process with state space $1, \dots, K$ whose transition matrix is given by

$$\begin{aligned}P(s, t) &= \Pi_{[s, t]}(I + dA) \\ &\equiv \lim_{\max |t_i - t_{i-1}| \rightarrow 0} \Pi_i(I + A(t_i) - A(t_{i-1}))\end{aligned}\quad (11)$$

where $s \leq t_1 \leq t_n \leq t$. One can think of the probabilistic behavior as follows: Given that the chain is in state i at time s the probability that it remains in that state at least until t (assuming that $\Delta A_{ii}(u) > -1$ for $u \leq t$) is given by

$$P(\Delta\eta_u = 0 \text{ for } s < u \leq t | \eta_s = i) \\ = \exp(-(A_{ii}(t) - A_{ii}(s))) \quad (12)$$

We are interested in testing assumptions on the intensity measure when it can be represented through integrated intensities, that is, we assume that there exists integrable functions (or transition intensities) $\lambda_{ij}(\cdot)$ such that

$$A_{ij}(t) = \int_0^t \lambda_{ij}(s) ds \quad (13)$$

for every pair of states i, j with $i \neq j$.

In this case, given that the chain jumps away from i at date t , the probability that it jumps to state j is given by $\frac{\lambda_{ij}(t)}{\sum_{k \neq i} \lambda_{ik}(t)}$.

A homogeneous Markov chain with intensity matrix Λ has $A_{ij}(t) = \lambda_{ij}t$ and in this special case we can write $P(s, t) = \exp(\Lambda(t - s))$.

For a method for estimating the continuous-time transition probabilities nonparametrically using the so-called Aalen–Johansen estimator, see, for example, [2]. The specification of individual transition intensities allows us to use hazard regressions on specific rating transitions. For an example of nonparametric techniques, see [5]. A Cox regression approach can be found in [9].

Empirical Observations

There is a large literature on the statistical properties of the observed rating transitions, mainly for firms rated by Moody's and Standard and Poors. It has been acknowledged for a long time that the observed processes are not time homogeneous and not Markov. This is consistent with stated objectives of rating agencies of trying to avoid rating reversals and seeking to change ratings only when the change in credit quality is seen as enduring—a property sometimes referred to as “rating through the cycle”. This is in contrast to “point-in-time” rating. The distinction between the two approaches is not rigorous, but a rough indication of the difference is that

a primary concern of through-the-cycle rating is the correct ranking of the firm's default probabilities (or expected loss) over a longer time horizon, whereas a point-in-time is more concerned with following actual, shorter-term default probabilities seeking to maintain a constant meaning of riskiness associated with each rating category.

The degree to which transition probabilities depend on the previous rating history, business cycle variables, and the sector or country to which the rated companies belong has been investigated, for example, in papers [1, 9, 10]. A good entry into the literature is in the special journal issue introduced by Cantor [3].

Rating agencies have a system of modifiers that effectively enlarge the state space. For example, Moody's operates with a watchlist and long-term outlooks. Being on a watchlist signals a high likelihood of rating action in a particular direction in the near future, and outlooks signal longer term likely rating directions. Hamilton and Cantor [7] investigate the performance of ratings when the state space is enlarged with these modifiers and conclude that they go a long way in reducing dependence on rating history.

Correlated Transitions

In risk management, the risk of loan portfolios and exposures to different counterparties in derivatives contracts depends critically on the extent to which the credit ratings of different loans and counterparties are correlated.

We finish by briefly outlining two ways of incorporating dependence into rating migrations. For the first approach, see, for example, [6]; we map rating probabilities into thresholds. The idea is easily illustrated through an example. If firm 1 is currently rated i and we know the (say) 1-year transition probabilities $p_{i1}, \dots, p_{iK}$, then we can model the transition to the various categories using a standard Gaussian random variable ϵ_1 and defining thresholds $a_1 > a_2 > \dots > a_{K-1}$ such that

$$p_{iK} = P(\epsilon_1 \leq a_{K-1}) = \Phi(a_{K-1}) \quad (14)$$

$$p_{i,K-1} = P(a_{K-1} \leq \epsilon_1 \leq a_{K-2}) \\ = \Phi(a_{K-2}) - \Phi(a_{K-1}) \quad (15) \\ \vdots$$

$$p_{i1} = P(a_1 \leq \epsilon_1) = 1 - \Phi(a_1) \quad (16)$$

Similarly, for firm 2, we can define thresholds $b_1, \dots, b_{K-1}$ and a standard random normal variable ϵ_2 so that the transition probabilities are matched as earlier. Letting ϵ_1 and ϵ_2 be correlated with correlation coefficient ρ induces correlation into the migration patterns of the two firms. This can be extended to a large collection of firms using a full correlation matrix obtained, for example, by looking at equity return correlations.

A second approach, which makes it possible to link up rating dynamics with continuous-time pricing models, is proposed in [8]. The idea here is to model the “conditional generator” of a Markov process as the product of a constant generator Λ and a strictly positive affine process μ , that is, conditionally on a realization of the process μ , the Markov chain is time non-homogeneous with the transition intensity $\lambda_{ij}(s) = \mu(s)\lambda_{ij}$. This framework allows for closed form computation of transition probabilities in a setting where rating migrations are correlated through dependence on state variables.

References

- [1] Altman, E. & Kao, D.L. (1992). The implications of corporate bond rating drift, *Financial Analysts Journal* **48**(3), 64–75.
- [2] Andersen, P.K., Borgan, O., Gill, R. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer, New York.
- [3] Cantor, R. (2004). An introduction to recent research on credit ratings, *Journal of Banking and Finance* **28**, 2565–2573.
- [4] Cantor, R. (2008). *Moody's Guidelines for the Withdrawal of Ratings*, Rating Methodology, Moody's Investors Service, New York.
- [5] Fledelius, P., Lando, D. & Nielsen, J. (2004). Non-parametric analysis of rating transition and default data, *Journal of Investment Management* **2**(2), 71–85.
- [6] Gupton, G., Finger, C. & Bhatia, M. (1997). *Credit Metrics—Technical Document*, Morgan Guaranty Trust Company.
- [7] Hamilton, D. & Cantor, R. (2004). *Rating Transitions and Defaults Conditional on Watchlists, Outlook and Rating History*, Special comment, Moody's Investors Service, New York.
- [8] Lando, D. (1998). On Cox processes and credit risky securities, *Review of Derivatives Research* **2**, 99–120.
- [9] Lando, D. & Skødeberg, T. (2002). Analyzing rating transitions and rating drift with continuous observations, *The Journal of Banking and Finance* **26**, 423–444.
- [10] Nickell, P., Perraudin, W. & Varotto, S. (2000). Stability of ratings transitions, *Journal of Banking and Finance* **24**, 203–227.

DAVID LANDO

Credit Migration Models

It is nowadays widely recognized that portfolio models are an essential tool for a proper and effective management of credit portfolios, be it from the perspective of a corporate bank, a mortgage bank, a consumer finance provider, or a fixed-income asset manager. Traditional credit management was, to a large extent, focused on the stand-alone analysis and monitoring of the credit quality of obligors or counterparties. Frequently, the credit process did also include *ad hoc* exposure-based limit-setting policies that were devised in order to prevent excessive risk concentrations. This approach was scrutinized in the 1990s, when the financial industry started to realize that univariate models for obligor default had to be extended to a portfolio context. It was recognized that credit rating and loss recovery models, although a crucial element in the assessment of credit risk, fail to explain some of the important stylized facts of credit loss distributions, if the stochastic dependence of obligor defaults is neglected. From a statistical point of view, not only the skewness and the relatively heavy upper tails of credit portfolio loss distributions, but also the historically observed variation of default rates and the clustering of bankruptcies in single sectors are clearly inconsistent with stochastic independence of defaults. From an economics point of view, it is plausible that default rates are connected to the intrinsic fluctuations of business cycles; relationships between default rates and the economic environment have indeed been established in numerous empirical studies [5]. All these insights supported the quest for tractable credit portfolio models that reflect these stylized facts.

Apart from an accurate statistical description of credit losses, a portfolio model can serve many more purposes. In contrast to a univariate approach, a credit portfolio framework allows to quantify the diversification effects between credit instruments. This makes it, for example, possible to evaluate the impact on the total risk when securities are added or removed from a portfolio. In the same vein, the risk numbers produced by a portfolio model help to identify possible hedges. Ultimately, the use of a portfolio model facilitates the active management of credit portfolios and the efficient allocation of capital. Less of a pure risk management matter is the use of portfolio models for risk-adjusted pricing

(see **Loan Valuation**) or performance measurement. The total portfolio risk is commonly considered as a capital which the lender should hold in order to buffer large losses. For nontraded assets, such as loans or mortgages, the costs for holding this risk capital are typically transferred to the borrowers by means of a surcharge on interest rates. Calculating these surcharges necessitates that the total portfolio risk capital is broken down to borrower (or instrument) level risk contributions. Only in a portfolio model framework, where the dependence between obligors and the resulting diversification benefits are correctly captured, this risk contribution can be determined in an economically rational and fair fashion. We mention that risk contributions can also be applied in order to determine the *ex post* (historical) risk-adjusted performance of instruments or subportfolios. Credit portfolio models also play an important role in the pricing of credit derivatives or structured products, such as credit default swaps or CDSs. For the correct pricing of many of these credit instruments, it is crucial that the dependence between obligor default times are well modeled.

Overview of Credit Migration-based Models

This article gives a survey on migration-based portfolio models, that is, models that describe the joint evolution of credit ratings. The ancestor of all such models is CreditMetrics^{®a}, which was introduced by the US investment bank J.P. Morgan. In 1997, J.P. Morgan and cosponsors from the financial industry published a comprehensive technical document [13] on CreditMetrics, in an effort to set industry standards and to create more transparency in credit risk management. This publication attracted a lot of attention and proved to stimulate research in credit risk. To this date, the CreditMetrics or derivations thereof have been implemented by a large number of financial institutions. Before we turn to a detailed description of CreditMetrics, it might be worth to mention two related models. CreditPortfolioView by McKinsey & Co is credit-migration-based as well. However, in contrast to CreditMetrics, which assumes temporally constant transition matrices, it is endowed with an estimator of credit migration probabilities based on macroeconomic observables. is dedicated to its discussion.

The second link concerns the longer standing KMV model.^b An outline of the KMV methodology can be extracted from an article by Kealhofer and Bohn [16]. In both CreditMetrics and KMV, the obligor correlation is generated in a similar fashion, that is, with a dependence structure following a Gaussian copula. The main differences concern the number of credit states and the source of probabilities of default (PDs). The KMV model operates on a continuum of states, namely, the so-called expected default frequencies (Moody's KMV EDF^{TMc}), basically estimated PDs, whereas CreditMetrics is restricted to a finite number of credit rating states. For this reason, KMV is strictly spoken not a credit-migration-based model and therefore only touched in this article. As remarked by McNeil *et al.* [19], a discretization of EDF would translate KMV to a model which, apart from parametrization, is structurally equivalent to CreditMetrics. Secondly, while for CreditMetrics rating transition matrices are the required exogenous inputs, the KMV counterparts, EDF of listed companies, are estimated through a proprietary method, which is basically an extension of the celebrated Merton model [20] for firm default. Inputs to the EDF model are historical time-series of equity prices together with company debt information, with which the unobserved asset value processes are reconstructed and a quantity called *distance to default* (DD) is calculated for every firm. This DD is used as a predictor of EDF; the relationship is determined by a nonlinear regression of historical default data against historical DD values. It is beyond the scope of this article to provide more details and so we refer to [2] or [17] for an account of the EDF methodology.

The CreditMetrics Model

CreditMetrics models the distribution of the credit portfolio value at a future time, from which risk measures can be derived. The changes of portfolio value are caused by credit migrations of the underlying instruments. In the following, we describe the rationale of the main building blocks of CreditMetrics.

Timescale

CreditMetrics was conceived as a discrete time model. It has a user-specified time horizon T that is

reached in one step from the analysis time 0; typically the time horizon is 1 year. It is assumed that the portfolio is static, that is, its composition is not altered during the time period $(0, T)$.

Risk Factors and Valuation

In case of CreditMetrics, the basic assumption is that each instrument is tied to one or several obligors. The user furnishes obligors with a rating from a rating system with a finite number of classes and an absorbing default state. The obligor ratings are the main risk drivers. We index the obligors by $i = 1, \dots, n$ and assume a rating system with rating classes $\{1, \dots, K\}$ that are ordered with respect to the credit quality, and a default class 0. At time 0, the obligor i has the (known) initial rating S_i^{init} , which then becomes S_i^{new} at time T . The change from S_i^{init} to S_i^{new} happens in a random fashion, according to the so-called credit migration probabilities. These probabilities are assumed to be identical for obligors in the same rating class and can therefore be represented by a so-called credit migration (or rating transition) matrix $\mathbf{M} = (m_{jk})_{j,k \in \{0, \dots, K\}}$. Clearly,

$$\mathbb{P}(S_i^{\text{new}} = k | S_i^{\text{init}} = j) = m_{jk} \quad (1)$$

The credit migration matrix is an important input to CreditMetrics. In practice, one often uses rating systems supplied by agencies such as Moody's or Standard&Poor's. The model also allows to work in parallel with several rating systems, depending on the obligor. If public ratings are not available, financial institutions can resort to internal ratings; *see Credit Rating; Internal-ratings-based Approach; Credit Scoring*.

To treat specific positions, CreditMetrics must estimate values for the position contingent on the position's obligor being in each possible future rating state. This is equivalent to estimating the loss (or gain) on the position contingent on each possible rating transition. In the case of default, the recovery rate δ_i determines the proportion of the position's principal that is paid back by the obligor.^d

For the nondefault states, the standard implementation of the model is to value positions based on market factors: the risk-free interest rate curve and a spread curve corresponding to the rating state. For this reason, CreditMetrics is commonly referred to as a *mark-to-market model*. Importantly, the mark-to-market approach incorporates a maturity effect into

the model: other things being equal, a downward credit migration will have a greater impact on a long maturity bond than a short one, given the long bond's higher sensitivity (duration) to the spread widening that is assumed to accompany the migration. However, this approach does require relevant spread curves for positions of all possible rating states.

For positions where there is little market information, or where the mark-to-market approach is inconsistent with an institution's accounting scheme, it is possible to utilize policy-driven rather than market-driven valuation. For example, if an institution has a reserves policy whereby loss reserves are determined by credit rating and maturity, then the change in required reserves can serve as a proxy for the loss on a position, contingent on a particular rating move. In this way, the model can still incorporate a maturity effect, even where a mark-to-market approach is not practical.

Risk Factor Dynamics and Obligor Dependence Structure

In the original formulation of CreditMetrics, foreign exchange (FX) rates and interest rate and spread curves are assumed to be deterministic since one focuses on the rating as the main risk driver. In principle, this assumption could be relaxed.

The migration matrix in the CreditMetrics model specifies the rating dynamics of a *single* obligor, but it does not provide any information about the *joint* obligor credit migrations. In order to capture the obligor dependence structure, CreditMetrics borrows ideas from the Merton structural model for firm default, which links default to the obligor asset value falling short of its liabilities. The assumption of CreditMetrics is that the obligor rating transition is caused by changes of the obligor's asset value, or equivalently, the asset value return. The lower this random return, the lower the new rating; if the asset value return drops below a certain threshold, default occurs. Mathematically, this amounts to defining return buckets for each obligor; the thresholds bounding these buckets depend on the initial obligor rating, the transition probabilities, and the return distribution. The rating of an obligor is determined by the bucket its return falls into. Obviously, the bucket probabilities must coincide with the transition probabilities. Models of this type are also called *threshold models*.

More formally, if R_i denotes the asset return of obligor i over $(0, T]$, then the rating at time T is determined by

$$S_i^{\text{new}} = j \iff d_j^{(i)} < R_i \leq d_{j+1}^{(i)} \quad (2)$$

The increasing thresholds $d_j^{(i)}$ are picked such that the resulting migration probabilities coincide with the ones prescribed by the credit migration matrix. Consequently,

$$d_0^{(i)} = -\infty \quad \text{and} \quad d_{K+1}^{(i)} = +\infty$$

$$G_i(d_{j+1}^{(i)}) - G_i(d_j^{(i)}) = \mathbb{P}(S_i^{\text{new}} = j \mid S_i^{\text{init}}) \quad (3)$$

where G_i is the cumulative distribution function of R_i . We illustrate the rating transition mechanism in Figure 1, which shows the return distribution and the thresholds for an obligor with an initial rating 2 in a hypothetical rating system with four nondefault classes.

The dependence between obligor ratings stems from the dependence of the asset returns. CreditMetrics assumes that these returns follow a linear factor model with multivariate normal factors and independent Gaussian innovations. This means that

$$R_i = \alpha_i + \sum_{\ell=1}^p \beta_{i\ell} F_\ell + \sigma_i \epsilon_i \quad (4)$$

where the common factors $\mathbf{F} = (F_1, \dots, F_p)' \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are multivariate Gaussian and the ϵ_i 's are independent and identically distributed (i.i.d.) standard normal variables independent of the factors. The numbers $\beta_{i\ell}$ are also called *factor exposures* or *loadings*, $\sum_{\ell=1}^p \beta_{i\ell} F_\ell$ is the systematic return, σ_i is the volatility of idiosyncratic (or specific) return $\sigma_i \epsilon_i$ of obligor i , and the real parameter α_i is referred to as *alpha*. The dependence between the returns, and consequently the dependence between future ratings, is caused by the exposure of the obligors to the common factors. Usually one normalizes the returns R_i to unit variance; this does not alter the joint distribution of $\mathbf{S}^{\text{new}} = (S_1^{\text{new}}, \dots, S_n^{\text{new}})'$ and leads to adjusted thresholds that are simpler. Not explicitly distinguishing between returns and normalized returns in our notation, equation (4) then reads as

$$R_i = \psi_i(\mathbf{F}) + \sqrt{1 - \text{Var}(\psi_i(\mathbf{F}))} \epsilon_i \sim \mathcal{N}(0, 1) \quad (5)$$

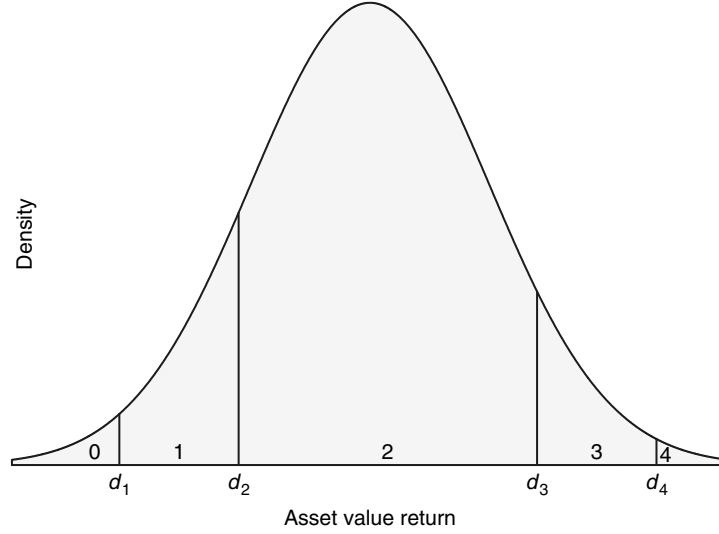


Figure 1 Asset return distribution, thresholds and rating classes

for appropriate affine linear functions ψ_i . The adjusted thresholds are given by

$$d_j^{(i)} = d_j(S_i^{\text{init}}) \quad \text{with} \\ d_j(s) = \Phi^{-1} \left(\sum_{k=0}^{j-1} m_{sk} \right), \quad 0 < j \leq K \quad (6)$$

where Φ is the standard normal distribution function.

As regards the recovery rates δ_i , they are assumed to be independent of each other and of the obligor returns. One stipulates that they are beta-distributed with parameters (a_i, b_i) .

Obligor Dependence Structure from Copulas

As was first recognized by Frey *et al.* [9] (see also [8]), copulas provide an elegant means for describing the obligor dependence structure in credit portfolio models. By virtue of Sklar's theorem, the joint distribution of the random vector $\mathbf{R} = (R_1, \dots, R_n)'$ can be factorized as

$$G(r_1, \dots, r_n) = \mathbb{P}(R_1 \leq r_1, \dots, R_n \leq r_n) \\ = C(G_1(r_1), \dots, G_n(r_n)) \quad (7)$$

where C , the copula associated with $\mathbf{R}$, is the distribution function of a random vector with standard uniform marginal distributions. From standard arguments (*see e.g.*, **Copulas: Estimation; Copulas in Econometrics; or Copulas in Insurance** and references therein),

$$\begin{aligned} \mathbb{P}(S_1^{\text{new}} = s_1, \dots, S_n^{\text{new}} = s_n) \\ &= \mathbb{P}(d_{s_1}^{(1)} < R_1 \leq d_{s_1+1}^{(1)}, \dots, d_{s_n}^{(n)} < R_n \leq d_{s_n+1}^{(n)}) \\ &= \int_{d_{s_1}^{(1)}}^{d_{s_1+1}^{(1)}} \cdots \int_{d_{s_n}^{(n)}}^{d_{s_n+1}^{(n)}} dG(r_1, \dots, r_n) \\ &= \int_{G_1(d_{s_1}^{(1)})}^{G_1(d_{s_1+1}^{(1)})} \cdots \int_{G_n(d_{s_n}^{(n)})}^{G_n(d_{s_n+1}^{(n)})} dC(u_1, \dots, u_n) \end{aligned} \quad (8)$$

Note that the integration limits $G_i(d_{s_i}^{(i)}) = \sum_{k=0}^{s_i-1} m_{s_i^{\text{init}},k}$ do not depend on G_i . This implies that the joint distribution of the ratings vector $\mathbf{S}^{\text{new}}$ is determined by the initial obligor ratings, the credit migration matrix $\mathbf{M}$ and the copula C ; the marginal distributions G_i do not matter.

This result helps to categorize threshold credit portfolio models; models using the same families of copulas can be considered as structurally equivalent. The copula associated with a Gaussian random vector

is called *Gaussian copula* and depends on the correlation matrix only. Since the returns $\mathbf{R}$ are multivariate Gaussian, the original CreditMetrics model [13] is referred to as *having a Gaussian copula*. Replacing the Gaussian copula family by other families gives different models. Frey *et al.* [9] study the CreditMetrics model with Student- t copulas and find that the tail of the loss distribution is considerably fatter as compared with the Gaussian copula with identical correlation parameters.

CreditMetrics as a Mixture Model

We now interpret the CreditMetrics model in a conditional fashion in order to better understand the meaning of the factors $\mathbf{F}$. To this end, we look at the vector of default indicators $\mathbf{D} = (D_1, \dots, D_n)'$, where $D_i = \mathbf{I}_{\{S_i^{\text{new}} = 0\}}$. We denote the default probabilities by $\bar{p}_i = \mathbb{P}(D_i = 1)$. Then conditional on the factors $\mathbf{F}$, the vector $\mathbf{D}$ consists of *independent* Bernoulli random variables with success probabilities

$$\begin{aligned} p_i(\mathbf{F}) &= \mathbb{P}(D_i = 1 \mid \mathbf{F}) \\ &= \Phi \left(\frac{\Phi^{-1}(\bar{p}_i) - \psi_i(\mathbf{F})}{\sqrt{1 - \text{Var}(\psi_i(\mathbf{F}))}} \right) \end{aligned} \quad (9)$$

We deduce that one can simulate $\mathbf{D}$ by first drawing $\mathbf{F}$ from a multivariate Gaussian distribution and then generating independent Bernoulli random variables with success probabilities $p_i(\mathbf{F})$. From this angle, $\mathbf{F}$ represents the state of the economy that determines the obligor default probabilities. The distribution of $\mathbf{D}$, which is obtained from “mixing” the conditional distributions of $\mathbf{D}$ by $\mathbf{F}$, is a so-called Bernoulli mixture:

$$\begin{aligned} &\mathbb{P}(D_1 = d_1, \dots, D_n = d_n) \\ &= \mathbb{E}(\mathbb{P}(D_1 = d_1, \dots, D_n = d_n \mid \mathbf{F})) \\ &= \mathbb{E} \left(\prod_{i=1}^n p_i(\mathbf{F})^{d_i} (1 - p_i(\mathbf{F}))^{1-d_i} \right) \end{aligned} \quad (10)$$

The conditional view offers computational advantages. Finger [6] exploits it for the determination of credit portfolio distributions. Using the Poisson approximation for sums of independent Bernoulli variables, various authors have shown that CreditMetrics is approximated by a Poisson mixture model;

this allows one to make a link to CreditRisk+ (see [3, 8, 11] for details).

Asymptotic Behavior of CreditMetrics

In CreditMetrics, risk measures such as VaR (Value-at-Risk) or ES (expected shortfall) cannot be expressed in terms of simple closed formulas. For their estimation, one has to resort to Monte-Carlo (MC) simulation or other numerical methods (see **Credit Portfolio Simulation**). Although the topic of approximations in credit portfolio models is covered in **Large Pool Approximations; Saddlepoint Approximation**, we provide a brief discussion because the asymptotic results provide important qualitative insights.

Concerning approximation, research has dealt with strong limits of the relative portfolio loss and the tail behavior of the loss distribution when the number of obligors tends to infinity. While the derivation of strong limits consists of straightforward applications of the strong law of large numbers, the analysis of the tail behavior of the loss is more involved. For the tail behavior, we refer to [18] and a recent article by Glasserman *et al.* [10].

We next present the idea of the so-called large pool approximation. To this end, we work in a simplified framework that is adapted to loan portfolios. We assume that recovery rates, spreads, and interest rates are all equal to zero. Every obligor i has an outstanding loan of size e_i . Then the loss in the period $(0, T]$ is given by

$$L_n = \sum_{i=1}^n e_i D_i \quad (11)$$

We define total exposure by $e = \sum_{i=1}^n e_i$ and set $\bar{L}_n = L_n/e$ for the relative loss of the portfolio. We want to verify to which extent the specific risk caused by the obligor-specific returns ϵ_i is diversified away when the number of obligors grows. To this end, we decompose the relative loss into a systematic and an obligor-specific component:

$$\bar{L}_n = \mathbb{E}(\bar{L}_n \mid \mathbf{F}) + \Delta_n \quad (12)$$

It is straightforward to show that obligor-specific variance tends to zero as $n \rightarrow \infty$, provided the

Herfindahl index, which measures exposure concentration, converges to zero:

$$\text{Var}(\Delta_n) \rightarrow 0 \quad \text{if} \quad H_n = \frac{1}{e} \sum_{i=1}^n e_i^2 \rightarrow 0 \quad (13)$$

The stronger property that Δ_n converges almost surely to zero holds true (see e.g. [8] for a proof). This justifies the following approximation for large portfolios with no severe exposure concentration:

$$\bar{L}_n \approx \mathbb{E}(\bar{L}_n | \mathbf{F}) = \frac{1}{e} \sum_{i=1}^n p_i(\mathbf{F}) \quad (14)$$

This means that for n large enough, the relative loss virtually coincides with its systematic component, or in other words, the idiosyncratic risks are diversified away and one is left with systematic risk (caused by $\mathbf{F}$) only. The systematic risk cannot be diversified away other than by hedges that impact $\mathbb{E}(\bar{L}_n | \mathbf{F})$. Despite this positive result, one should be aware that the portfolio R-squared is typically significantly lower than the R-squared in long-only equity portfolios with a similar number of securities. This indicates that in credit portfolios the idiosyncratic risks are comparatively more important than in equity portfolios, or in other words, more securities are necessary to diversify a credit portfolio. This feature is due to the fact that the default correlations $\text{Corr}(D_i, D_{i'})$ are very low (less than 5%) for typical values of asset correlations (10–30%) and default probabilities.^e

The large pool approximation (14) lies at the core of the capital requirement in the Basel II internal ratings-based (IRB) approach. The IRB approach imposes a constant correlation Gaussian one-factor model, that is, $\psi_i(\mathbf{F}) = \sqrt{\rho}\mathbf{F}$ in (5), and $\mathbf{F} \sim \mathcal{N}(0, 1)$ is a latent (unobserved) factor. IRB assumes that the portfolio is infinitely fine-grained, that is, it neglects Δ_n and calculates risk directly from $\mathbb{E}(\bar{L}_n | \mathbf{F})$. Since the latter is the sum of comonotonic^f (or perfectly positive dependent) obligor contributions, portfolio VaR is just the sum of the obligor VaR contributions. For an explanation of the additional multiplicative adjustments for maturities and rules for the choice of ρ_i applied by the IRB approach, we refer to **Internal-ratings-based Approach**, [7, 12].

If the default probabilities in the constant correlation Gaussian one-factor model are constant and

equal to $\bar{p}$, the limit $\mathbb{E}(\bar{L}_n | \mathbf{F})$ is a so-called probit-normal random variable:

$$\bar{L}_n \approx p(\mathbf{F}) = \Phi\left(\frac{\Phi^{-1}(\bar{p}) - \sqrt{\rho}\mathbf{F}}{\sqrt{1-\rho}}\right) \quad (15)$$

This result goes back to [23, 24]. Studying the properties of the probit-normal distribution reveals that the tail behavior of $\bar{L}_n$ is essentially governed by the correlation; the larger the ρ , the thicker the tail of the distribution of $\bar{L}_n$. Along the same lines one can analyze the tails of credit-migration-based models with non-Gaussian copulas, as was done by Lucas *et al.* [18].

Model Estimation

Statistical inference poses a big challenge in CreditMetrics and related models. No estimation method that we know of is truly flawless because assumptions that are difficult to verify are ever involved. As we have seen, the model (2) is fully determined by the credit migration matrix and the matrix of asset correlations. In what follows, we take the existence of a credit migration matrix for granted (*see Rating Transition Matrices*) and restrict our discussion to the estimation of asset correlations. *See also Portfolio Credit Risk: Statistical Methods* for a general discussion of statistical methods for the calibration of credit portfolio models. We distinguish between *direct* and *indirect* approaches. The direct approaches start by estimating the exposures and variances in the factor model (4), typically by regressions against certain predefined factors; the asset correlations are then easily derived. In the indirect approaches, asset correlations are inferred from historical default data.

CreditMetrics uses a direct approach. A major difficulty is that firm asset values cannot be directly observed, even for companies that have publicly traded equity. CreditMetrics circumvents this problem by assuming that asset return correlations are proxied by equity return correlations, or in other words, it regards the R_i 's in equation (4) as equity returns. The CreditMetrics technical document [13] suggests to use MSCI industry and country index returns as explanatory factors. In principle, one could work with any other set of factors. A nice benefit of letting equity returns drive the rating transitions is the

fact that CreditMetrics can be naturally embedded into market risk models of the RiskMetrics^{®g} type because the CreditMetrics risk factors are already part of the RiskMetrics factor universe. This allows the aggregation of credit and market risks in a straightforward fashion.

So far we have explained the case of obligors with publicly traded equity. For private firms, the betas are mostly set by resorting to economic arguments. Often one uses the obligor's country and industry or sector affiliations. Say, a firm has two equally large lines of businesses, which belong to the US Information Technology and US Consumer Discretionary sectors, respectively. Then the betas with respect to the MSCI US Information Technology and MSCI US Consumer Discretionary sector index returns would both be set to 0.5, and all other betas to zero. Another piece to specify is the variance of the idiosyncratic term in equation (4), or equivalently, the R-squared. Here CreditMetrics stipulates that R-squareds obey $R^2 = 1/(1 + A^\gamma \exp(\lambda))$, where A is the book value of the firm's total assets and γ and λ are fixed parameters estimated from a cross-section of traded stocks; we refer to [14] for a critical appraisal of this method. Alternatively, R-squared can be set by the experienced user.

We mention that KMV uses a direct approach as well, but in contrast to CreditMetrics it reconstructs the asset values from the equity price history using the Merton model framework; see [17] for details about this reconstruction and [16] for a description of the structure of the KMV factor model.

The indirect approaches apply statistical inference to time-series of count data of defaults. These methods are constricted by the sparsity of clean data. Since defaults rarely happen and the time-series are short, one creates groups of obligors and assumes that asset correlations in these groups are constant. There exist several studies following an indirect approach. Bluhm *et al.* [1] back out asset correlations in rating classes from standard deviations of historical default rates. De Servigny and Renault [22] infer default correlation from observed joint defaults. Demey *et al.* [4] use maximum-likelihood estimation and reduce the number of parameters by assuming that both asset correlations between and within rating classes are constants. Hamerle and Rösch [15] also apply maximum-likelihood estimation, but advocate the use of lagged macroeconomic variables as additional predictors.

End Notes

^aRiskMetrics and CreditMetrics are registered trademarks of RiskMetrics Group, Inc. and its affiliates.

^bSince the original development of the model, the KMV company was acquired by Moody's, Inc. and is now part of Moody's KMV.

^cMoody's KMV EDF, which we refer to as simply EDF, is a trademark of Moody's KMV. See www.moodyskmv.com.

^dEquivalently, the loss is determined by the loss given default (LGD) specified for the position.

^eThe picture is similar when R-squared is replaced by squared ratios of other risk measures.

^fSee e.g., McNeil *et al.* [19] for formal definition of comonotonicity.

^gSee Mina and Xiaao [21] for an introduction to RiskMetrics.

References

- [1] Bluhm, C., Overbeck, L. & Wagner, C. (2003). *An Introduction to Credit Risk Modeling*, John Wiley & Sons.
- [2] Crosbie, P. & Bohn, J.R. (2003). *Modeling Default Risk*. Available at www.moodyskmv.com.
- [3] Crouhy, M., Galai, D. & Mark, R. (2000). A comparative analysis of current credit risk models, *Journal of Banking and Finance* **24**, 59–117.
- [4] Demey, P., Jouanin, J.-F., Roget, C. & Roncalli, T. (2004). Maximum likelihood estimate of default correlations, *Risk*, November, 104–108.
- [5] Duffie, D. & Singleton, K.J. (2003). *Credit Risk: Pricing, Measurement, and Management*, Princeton Series in Finance, Princeton University Press.
- [6] Finger, C.C. (1999). Conditional approaches for CreditMetrics portfolio distributions, *CreditMetrics Monitor* April, 14–33.
- [7] Finger, C.C. (2001). The one-factor CreditMetrics model in the new Basel Capital Accord, *RiskMetrics Journal* 9–18.
- [8] Frey, R. & McNeil, A.J. (2003). Dependent defaults in models of portfolio credit risk, *Journal of Risk* **6**(1), 59–92.
- [9] Frey, R., McNeil, A.J. & Nyfeler, M. (2002). Copulas and credit models, *Risk*, October, 111–114.
- [10] Glasserman, P., Kang, W. & Shahabuddin, P. (2007). Large deviations in multifactor portfolio credit risk, *Mathematical Finance* **17**(3), 345–379.
- [11] Gordy, M.B. (2000). A comparative anatomy of credit risk models, *Journal of Banking and Finance* **24**, 119–149.
- [12] Gordy, M.B. (2003). A risk-factor model foundation for ratings-based bank capital rules, *Journal of Financial Intermediation* **12**, 199–232.

- [13] Gupton, G.M., Finger, C.C. & Bhatia, M. (1997). *CreditMetrics—Technical Document*, J.P. Morgan & Co. Incorporated.
- [14] Hahnenstein, L. (2004). Calibrating the CreditMetrics correlation concept—empirical evidence from Germany, *Financial Markets and Portfoliomangement* **18**, 358–381.
- [15] Hamerle, A. & Rösch, D. (2006). Parameterizing credit risk models, *Journal of Credit Risk* **2**(4), 101–122.
- [16] Kealhofer, S. & Bohn, J.R. (2001). *Portfolio Management of Default Risk*. Available at www.moodyskmv.com
- [17] Lando, D. (2004). *Credit Risk Modeling*, Princeton Series in Finance, Princeton University Press.
- [18] Lucas, A., Klaassen, P., Spreij, P. & Straetmans, S. (2001). An analytical approach to credit risk in large corporate bond and loan portfolios, *Journal of Banking and Finance* **2**, 1635–1664.
- [19] McNeil, A.J., Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management*, Princeton Series in Finance, Princeton University Press.
- [20] Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.
- [21] Mina, J. & Xiaao, J.Y. (2001). *Return to RiskMetrics: The Evolution of a Standard*, RiskMetrics Group.
- [22] de Servigny, A. & Renault, O. (2003). Correlation evidence, *Risk*, July, 90–94.
- [23] Vasicek, O.A. (1987). *Probability of Loss on Loan Portfolio*. Available at www.moodyskmv.com
- [24] Vasicek, O.A. (2002). Loan portfolio value, *Risk*, December, 160–162.

Related Articles

Exposure to Default and Loss Given Default; Gaussian Copula Model; Large Pool Approximations; Structural Default Risk Models; Rating Transition Matrices.

DANIEL STRAUMANN & CHRISTOPHER C.
FINGER

Structural Default Risk Models

Structural models of default risk for individual firms originate from the seminal work of Merton [25]. Default is linked to the economic fundamentals of the considered firm *via* the assumption that default occurs if the value of the firm's assets, modeled as a geometric Brownian motion, falls below some default threshold (the firm's liabilities) at some future point in time (the maturity of a zero-coupon bond). A significant extension of this methodology was proposed by Black and Cox [7], who continuously test for default. Hence, in their model, the time of default is a first-passage time. Further generalizations address stochastic interest rates, more general assumptions on the default threshold, the definition of the default event, and discontinuous processes as model for the firm's assets [10, 12, 20–22, 34]. On a high level, these innovations aim at making the model-induced term structure of default probabilities flexible enough to allow for a precise fit of the model to observed bond prices and credit default swap (CDS) spreads.

The growing popularity of derivatives on credit portfolios, for example, collateralized debt obligations (CDOs) and n th to default baskets, and advanced demands on risk-management solutions produced a need for portfolio models that simultaneously explain the credit quality of multiple firms. Since corporate defaults in a globalized economy are not independent, a multivariate default model has to explain univariate default probabilities and the dependence among the default events. A natural assumption for a multivariate structural-default model is to introduce dependence by assuming correlated asset values, leading to dependent default events. Zhou [33] motivates this approach by the observation: "The fortunes of individual companies are linked together via industry-specific and/or general economic conditions." The first portfolio model of this class was formulated by Vasicek [31] and can be classified as a multivariate generalization of the work by Merton [25]. This model is discussed in some detail in the section "*The Model of Vasicek*", as it constitutes the basis for most of today's generalizations and is used to assess the regulatory capital for loan portfolios within the Basel II framework.

A major advantage of (multivariate) structural-default models is the appealing economic interpretation of the definition of default. Additionally, comovements of the individual firm-value processes might also be interpreted as being the result of common risk factors. Moreover, modeling the evolution of the firm's values as some multivariate stochastic process naturally implies a dynamic model, which is highly desirable in risk-management and pricing applications. The downside of this class of models is the mathematical challenge of computing the portfolio-loss distribution or even bivariate default correlations. Hence, most of the proposed models rely on simplifying assumptions or can be solved only *via* a Monte Carlo simulation.

Merton-type Models

The Model of Vasicek

In his short memo [31], Vasicek considers a portfolio of n loans with unit nominal and maturity T . Each individual firm-value process is modeled as a geometric Brownian motion defined by the stochastic differential equation:

$$\begin{aligned} dV_t^i &= V_t^i(\gamma^i dt + \sigma^i dW_t^i), \\ V_0^i &> 0, \quad i \in \{1, \dots, n\} \end{aligned} \quad (1)$$

The first simplification, often referred to as *homogeneous portfolio assumption*, is to assume identical default probabilities for all firms. In the current setup, this assumption corresponds to identical parameters $V_0 \equiv V_0^i$, $\sigma \equiv \sigma^i$, and $\gamma \equiv \gamma^i$. Moreover, an identical correlation across all bivariate pairs of Brownian motions W^i and W^j is assumed. Using Itô's formula and replacing the growth rate γ by the risk-free interest rate r , we find

$$V_T^i \stackrel{d}{=} V_0^i \exp((r - 0.5\sigma^2)T + \sigma\sqrt{T}X^i) \quad (2)$$

where $X^i := W_T^i/\sqrt{T}$ follows a standard normal distribution. Given some default threshold $d_T \equiv d_T^i$, one can immediately compute the probability of default at time T , since the distribution of the firm value at time T is known explicitly. Moreover, since default can only happen at maturity, only the distribution of V_T^i is of importance and not the dynamic model leading to it. By scaling the

original default threshold, default can alternatively be expressed in terms of the standard normally distributed variable X^i . More precisely, assuming the default probability of firm i at time T is given by p^i , the default threshold with respect to X^i is $K^i = \Phi^{-1}(p^i)$, where Φ^{-1} is the quantile function of the standard normal distribution.

To incorporate correlation among the companies, one explains X^i by a common market factor M and an idiosyncratic risk factor ϵ^i , that is,

$$X^i \stackrel{d}{=} \sqrt{\rho}M + \sqrt{1-\rho}\epsilon^i, \quad \rho \in (0, 1) \quad (3)$$

where $M, \{\epsilon^i\}_{i=1}^n$ are independent standard normally distributed random variables. Consequently, $\text{Cor}(X^i, X^j) = \rho$ for $i \neq j$, and each X^i is again distributed according to the standard normal law. By conditioning on the common market factor M , the firm's values and default events are independent. The result is the so-called *conditionally independent model*. We denote this conditional default probability by $p^i(M)$ and obtain

$$p^i(m) = \mathbb{P}(X^i < K^i | M = m) = \Phi\left(\frac{K^i - \sqrt{\rho}m}{\sqrt{1-\rho}}\right) \quad (4)$$

Furthermore, all companies are assumed to have identical default probabilities $p \equiv p^i$. Now Ω is defined as the random variable that describes the fraction of defaults in the portfolio up to time t . The distribution of Ω depends on two parameters: the individual default probability p and the correlation ρ . In what follows, this distribution is denoted by $F_{p,\rho}^{(n)}(x) = \mathbb{P}(\Omega \leq x)$. It is crucial that firms are independent given M , since the probability that exactly k firms default can be derived by integrating out the market factor:

$$\mathbb{P}(n\Omega = k) = \int_{-\infty}^{\infty} \mathbb{P}(n\Omega = k | M = m) \times \phi(m) dm \quad (5)$$

$$= \int_{-\infty}^{\infty} \binom{n}{k} p(m)^k (1-p(m))^{n-k} \times \phi(m) dm \quad (6)$$

where ϕ is the density of a standard normal distribution and $k \in \{0, \dots, n\}$. For large portfolios, evaluating the binomial coefficient is numerically critical and

can be avoided by applying the law of large numbers; this approach is called *large portfolio approximation*. The key observation is that $\mathbb{P}(\Omega \leq x | M = m) \rightarrow \mathbf{1}_{\{p(m) \leq x\}}$ for $(n \rightarrow \infty)$. A straightforward calculation, see, for example, [29] for details, establishes

$$F_{p,\rho}^{(n)}(x) \rightarrow F_{p,\rho}^{\infty}(x) := \Phi\left(\frac{\sqrt{1-\rho}\Phi^{-1}(x) - \Phi^{-1}(p)}{\sqrt{\rho}}\right), \quad (n \rightarrow \infty) \quad (7)$$

This approximation is continuous and strictly increasing in x . As it further maps the unit interval onto itself, it is a distribution function, too. It is also worth mentioning that the quality of this approximation is typically good; see [29] for a discussion.

The IRB Approach in Basel II

Vasicek's asymptotic loss distribution or, more precisely, its quantile function

$$K_{p,\rho}(y) := \Phi\left(\frac{\Phi^{-1}(y)\sqrt{\rho} + \Phi^{-1}(p)}{\sqrt{1-\rho}}\right) \quad (8)$$

plays a major role in today's regulatory world. The core of the first pillar of the Basel II accord [4] is the internal rating-based (IRB) approach for calculating capital requirements for loan portfolios. Within this framework, banks classify their loans by asset class and credit quality into homogeneous buckets and use their own internal rating systems to estimate risk characteristics such as loss-given default (LGD), the expected exposure at default (EAD), and the one-year default probability (PD), that is, $PD = p_1$. It is worth mentioning that estimating LGD and PD independently contradicts the empirical observation that recovery rates and default rates are inversely related; see, for example, [2] and [11]. Still, banks are free to choose a certain internal rating system, as long as they can demonstrate its accuracy and meet certain data requirements. In the second step, these credit characteristics are used in the IRB formula to assess the minimum capital requirements for the unexpected loss *via* the factor

$$K_{\text{IRB}} = \text{LGD} \cdot (K_{PD,\rho}(99.9\%) - PD) \cdot MA \quad (9)$$

The risk-weighted assets (RWA) are then obtained by

$$RWA = \frac{K_{IRB} \cdot EAD}{0.08} = 12.5 \cdot K_{IRB} \cdot EAD \quad (10)$$

where 0.08 corresponds to the 8% minimum capital ratio. The very conservative one-year 99.9%-quantile in equation (9) is part of the Basel II accord and might be interpreted as some cushion regarding the underlying simplifications in Vasicek's model. The factor MA is the *maturity adjustment* and calculated *via* (some exceptions apply)

$$MA = \frac{1 + (M - 2.5) \cdot b(PD)}{1 - 1.5 \cdot b(PD)},$$

$$M = \min \left\{ \frac{\sum_t t \cdot CF_t}{\sum_t CF_t}, 5 \right\} \quad (11)$$

and $b(PD) = (0.11852 - 0.05478 \cdot \log PD)^2$, where CF_t denotes the expected cash-flow at time t . M accounts for the fact that loans with longer (shorter) maturity than one year require a higher (lower) capital charge. Finally, the crucial correlation parameter needs to be specified. Basel II uses a convex combination between some lower ρ^l and upper ρ^u correlation whose weights depend on the default probability of the respective loan, that is,

$$\rho = \rho^l a(PD) + \rho^u (1 - a(PD)),$$

$$a(x) = \frac{1 - e^{-50x}}{1 - e^{-50}} \quad (12)$$

For corporate credits, the correlation-adjustment factor

$$SM_{ad}(S) = -0.04 \cdot \left(1 - \frac{\max\{5, S\} - 5}{45} \right) \times \mathbf{1}_{\{S \leq 50\}} \quad (13)$$

is added to ρ for borrowers with reported annual sales $S \leq 50$, measured in millions of Euros. The specific form of $a(x)$ and the adjustment factor $SM_{ad}(S)$ being negative stem from the empirical observation [23] that large firms that bear more systemic risk are more correlated compared to small firms that

are more likely to default due to idiosyncratic reasons. (ρ^l, ρ^u) depend on the type of loan and are specified as (0.12, 0.24) for sovereign, corporate, and bank loans; (0.12, 0.30) for highly volatile commercial real estate loans; $\rho^l = \rho^u = 0.15$ for residential mortgages; $\rho^l = \rho^u = 0.04$ for revolving retail loans such as credit cards; and finally (0.03, 0.16) for other retail exposures, where in this case the weight function $a(x)$ is computed with exponents -35 instead of -50 .

The IRB approach is sometimes criticized for the strong assumptions that are required to derive Vasicek's distribution. However, one should recognize the IRB approach as a compromise that provides a common language for regulators, banks, and investors to communicate and establishes comparable risk estimates across banks. The IRB formula is discussed in depth in [5, 30].

Generalizations Using Other Distributions

It is well known that the model [31] does not yield a satisfactory fit to market quotes of tranches of CDOs. More precisely, an implied correlation smile is present when the model is inverted for the correlation parameter tranche by tranche. Especially tail events with multiple defaults are underrepresented in a Gaussian world, making a precise fit to senior tranches of a CDO impossible. To overcome this shortcoming, a natural assumption is to give up normality in equation (3) and consider other heavier tailed distributions. For the derivation leading to $F_{p,\rho}^\infty$ in equation (7), the stability of the normal distribution under convolutions is essential in equation (3). Hence, natural choices for generalizations are other infinitely divisible distributions, which are connected to Lévy processes; see, for example, [8]. These generalizations add flexibility to the model and can additionally imply a dependence structure with tail dependence, making multiple defaults more likely. Specific models in this spirit include, for example, the NIG model of Kalemánova *et al.* [17], the VG model of Moosbrucker [27], and the BVG model of Baxter [6]. Following [1], we now derive a *large homogeneous portfolio approximation* in a general Lévy framework.

Let $X = \{X_t\}_{t \in [0,1]}$ be a Lévy process (see **Lévy Processes**) with $X_1 \sim H_1$ for some infinitely divisible distribution H_1 . Assume X_1 to be standardized to zero mean and unit variance. Given a correlation

$\rho \in (0, 1)$, define in analogy to equation (3) for independent copies $\{X^i\}_{i=1}^n$ of X the random variables V^i by

$$V^i := X_\rho + X_{1-\rho}^i, \quad i \in \{1, \dots, n\} \quad (14)$$

Here, the common market factor is represented by X_ρ , and the idiosyncratic risk of firm i is captured in $X_{1-\rho}^i$. Using the Lévy properties of X , each V^i is again distributed according to H_1 and $\text{Cor}(V^i, V^j) = \rho$ for $i \neq j$. In what follows, we denote by H_t^{-1} the inverse of the distribution function of X_t . The homogeneous portfolio assumption in the present setup translates to identical univariate default probabilities up to time T , abbreviated as $p \equiv p^i$, identical threshold levels $K_T = H_1^{-1}(p) \equiv K_T^i$, and unit notional of each firm. The probability of exactly k defaults in the portfolio is then again obtained as

$$\mathbb{P}(n\Omega = k) = \int_{-\infty}^{\infty} \mathbb{P}(n\Omega = k | X_\rho = m) dH_\rho(m), \quad k \in \{0, \dots, n\} \quad (15)$$

Similar to Vasicek's model, the conditional distribution of the number of defaults given $X_\rho = m$ is a binomial distribution with n trials and success probability $p(m) = \mathbb{P}(V^i \leq K_T | X_\rho = m) = H_{1-\rho}(K_T - m)$. The large portfolio assumption, that is, letting the number of firms n tend to infinity, then gives

$$F_{p,\rho}^\infty(x) = 1 - H_\rho(H_1^{-1}(p) - H_{1-\rho}^{-1}(x)) \quad (16)$$

as distribution function of the fractional loss in an infinite granular portfolio; see [1] for a complete proof. Let us finally remark that evaluating H_t and H_t^{-1} requires numerical routines for most choices of $X_1 \sim H_1$.

The Model of Willemann

The starting point for Willemann [32] is the univariate jump-diffusion model of Zhou [34]. This model assumes a discontinuous firm-value process of the form

$$\begin{aligned} dV_t &= V_t((\gamma - \lambda v)dt + \sigma dW_t + (\Pi - 1)dN_t), \\ V_0 &> 0 \end{aligned} \quad (17)$$

where N_t is a Poisson process with intensity $\lambda > 0$ and the jumps Π are log-normally distributed with

expected jump size $v = \mathbb{E}[\Pi - 1]$. The advantage of supporting negative jumps on a univariate level is that default events are no longer predictable, which translates to positive short-term credit spreads. Willemann [32] incorporates dependence to the individual firm-value processes by the classical decomposition of each Brownian motion into a market factor and an idiosyncratic component. Moreover, it is assumed that all firm-value processes jump together, that is, all processes are driven by the same Poisson process N_t . Consequently, this construction allows for two layers of correlation: diffusion and jump correlation; the latter being the main innovation of this setup.

The default threshold of firm i is set to $K_t^i = e^{-\phi^i t} K_0^i$ for some positive constants ϕ^i and K_0^i . This declining form is chosen to increase short-term spreads, but might also imply that the fit to individual CDS gets worse with increase in time. To achieve semianalytical results for the portfolio-loss distribution, default is tested on a grid. The advantage of this simplification is that only the distribution of each firm-value process at the grid points is required, instead of functionals as $\inf_{s \in [0, t]} V_s$. Individual default probabilities up to time t can then be computed conditional on the number of jumps up to time t , which is a Poisson-distributed random variable. Since the specific choice of jump-size distribution is compatible to the Brownian motion of the model, this leads to an infinite sum of normally distributed random variables. Moreover, all default events are independent conditional on the market factor and the number of jumps. Hence, the portfolio-loss distribution can be found by integrating out these common factors and using a recursion technique similar to [3, 16]. Willemann [32] demonstrates quite successfully how the model is simultaneously fitted (in seconds) to individual CDS spreads and the tranches of a CDO.

A Remark on Asset and Default Correlation

Modeling asset values as correlated stochastic processes introduces dependence to the resulting default times. Still, this relation is not trivial and deserves some caution, especially when it comes to estimating the model's asset-correlation parameter. We follow [24] in defining the default correlation of two firms (up to time t) as

$$\begin{aligned}\rho_t^D &:= \text{Cor}(\mathbb{1}_{\{\tau^1 \leq t\}}, \mathbb{1}_{\{\tau^2 \leq t\}}) \\ &= \frac{\mathbb{P}(P_t^1, P_t^2) - \mathbb{P}(P_t^1)\mathbb{P}(P_t^2)}{\sqrt{\mathbb{P}(P_t^1)(1 - \mathbb{P}(P_t^1))}\sqrt{\mathbb{P}(P_t^2)(1 - \mathbb{P}(P_t^2))}}\end{aligned}\quad (18)$$

where $P_t^i := \{\tau^i \leq t\}$, $i \in \{1, 2\}$. Most structural-default models share the commonality that evaluating $\mathbb{P}(P_t^1, P_t^2)$, the probability of a joint default of both firms up to time t , is quite difficult; an exception being the case of two companies with Gaussian factors coupled as described in equation (3). This example is, therefore, used to illustrate the nonlinear relation of asset and default correlation. A joint default in this setup corresponds to a simultaneous drop of both factors X^1 and X^2 below their respective default threshold $K^i = \Phi^{-1}(p_t^i)$, $i \in \{1, 2\}$. Since the vector (X^1, X^2) follows a two-dimensional normal distribution with mean vector $(0, 0)$ and the asset-correlation ρ as correlation parameter, we obtain

$$\mathbb{P}(P_t^1, P_t^2) = \Phi_2(K^1, K^2; \rho) \quad (19)$$

which is used to produce Figure 1. This example illustrates that small asset correlations induce only a negligible default correlation.

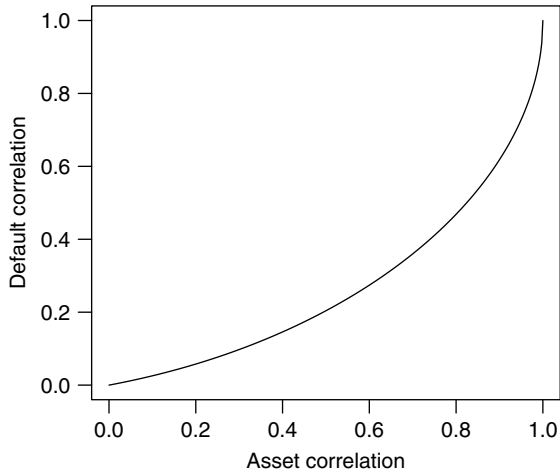


Figure 1 Default correlation ρ_t^D as a function of asset correlation $\rho \in [0, 1]$, with $\mathbb{P}(P_t^1) = \mathbb{P}(P_t^2) = 0.05$ and $t = 1$

Being able to convert default to asset correlations (and *vice versa*) opens the possibility of estimating the model's asset correlation using historical default correlations (and *vice versa*); see, for example, [14]. This approach is relevant since asset values are not directly observable, making an estimation of asset correlations delicate. It is an ongoing debate whether indirectly observed changes in asset values, computed from changes in the respective firm's equity, or observed defaults are the better source of data for the estimation of the model's correlation parameter. In both cases, pointing out the respective limitations is much simpler than providing theoretical evidence for the methodology. Empirically estimating default correlations (based on groups of firms with similar characteristics) requires a large set of observations, since corporate defaults are rare events. This makes the approach vulnerable to structural changes such as new bankruptcy rules. On the other hand, daily equity prices are readily available for most firms. When this latter source of data is used, the difficulty lies in transforming equity to asset returns, see, for example, [9], from which the correlation might be estimated. In addition, one should be aware that equity prices might change for reasons that are not related to credit risk.

First-passage Time Models

The starting point for most multivariate first passage-time models is equation (1). Compared to models in the spirit of the work by Merton [25], the time of default is now defined as suggested in [7], that is,

$$\tau^i := \inf\{t \geq 0 : V_t^i \leq d_t^i\}, \quad i \in \{1, \dots, n\} \quad (20)$$

where d_t^i is the default threshold of firm i at time t . From a modeling perspective, this definition overcomes the unrealistic assumption of default being restricted to maturity. This observation is even more crucial in a portfolio environment when bonds with different maturities are monitored simultaneously. More precisely, a first-passage model naturally induces a dynamic model for the default correlation (since the firm-value processes evolve dynamically over time) and allows the computation of consistent default correlations over any time horizon.

However, the main drawback of this model class is its computational intractability. This stems from

the fact that the joint distribution of the minimum of several firm-value processes is required, which is already a challenging problem for univariate marginals. The following section collects models where analytical results or numerical routines are available to overcome this problem.

The Model of Zhou

Zhou studies [33] a portfolio of two firms whose asset-value processes are modeled as in equation (1) with correlated Brownian motions. The default thresholds are assumed to be exponential, that is, $d_t^i = e^{\lambda^i t} K^i$ for $i \in \{1, 2\}$. The degree of dependence of both firms is measured in terms of their *default correlation* up to time t , that is, as $\text{Cor}(\mathbf{1}_{\{\tau^1 \leq t\}}, \mathbf{1}_{\{\tau^2 \leq t\}})$. The key observation is that results of Rebholz [28] can be applied to give an analytical representation of the default correlation in terms of an infinite sum of indefinite integrals over modified Bessel functions. Sensitivity analysis of the model parameters indicates that the model-induced default correlations for short maturities are close to zero. This observation needs to be considered when portfolio derivatives with short maturities are priced within such a framework.

The Model of Giesecke

Giesecke [13] considers a portfolio of n firms whose value processes evolve according to some vector-valued stochastic process $(V^1, \dots, V^n)$, where default is again defined as in equation (20). The key innovation is to replace the vector of default thresholds by an initially unobservable random vector $(d^1, \dots, d^n)$ whose dependence structure is represented by some copula. It is shown that the model-induced copula of default times is a function of the copula of default thresholds and the copula of the vector of historical lows of the firm-value processes.

On a univariate level, the assumption of an unobservable random threshold overcomes the predictability of individual defaults, which is responsible for vanishing credit spreads for short maturities; see [10] for a related model. Short-term spreads [13] are positive as long as the respective firm-value process is close to its historical low. The consequence of this construction on a portfolio level is also remarkable. Observing a corporate default τ^i reveals the respective default threshold d^i to all investors. This

piece of information allows to update the knowledge on all other default thresholds, leading to *contagious* jumps in credit spreads of the remaining firms. Giesecke [13] also presents an explicit example of two firms with independent value processes modeled as geometric Brownian motions and default thresholds coupled *via* a Clayton copula. While this simplified example illustrates the desired contagion effect of the model, it also highlights the challenge of finding analytic results in a realistic framework.

Models Relying on Monte Carlo Simulations

This section briefly presents two first-passage time models that rely on Monte Carlo simulations for the pricing of CDOs.

The n firm-value processes [15] are defined as in equation (1); the model can therefore be considered as a generalization of Zhou's [33] bivariate model to larger portfolios. The default thresholds are rewritten in terms of the driving Brownian motions. Asset correlation is introduced by n_F risk factors, that is, the Brownian motion of firm i is replaced by

$$dW_t^i := \sum_{j=1}^{n_F} \alpha_{i,j} dF_t^j + \left(1 - \sum_{j=1}^{n_F} \alpha_{i,j}^2\right)^{\frac{1}{2}} dU_t^i, \quad i \in \{1, \dots, n\} \quad (21)$$

where $\alpha_{i,j}$ is the sensitivity of firm i to changes of the risk factor F^j and U^i is the idiosyncratic risk of this firm. All processes F^j and U^i are independent Brownian motions. Hull *et al.* [15] also consider extensions to stochastic correlations, stochastic recovery rates, and stochastic volatilities and compare these in terms of their fitting capability to CDO tranches. An interesting conclusion that also applies to similar first-passage time models is drawn when the model is compared to a copula model. It is argued that the default environment in a copula model is static for the whole life of the model, while the dynamic nature of equation (21) allows to have bad default environments in one year, followed by good environments later. Hence, the use of one or more common risk factors implies a sound economic model for cyclical correlation.

Kiesel and Scherer [18] present another multivariate extension of the work by Zhou [34]. They model the firm-value process of the company i as

the exponential of a jump-diffusion process with two-sided exponentially distributed jumps Y_{ij} , that is,

$$V_t^i = V_0^i \exp(X_t^i),$$

$$X_t^i = \gamma^i t + \sigma^i W_t^i + \sum_{j=1}^{N_t(b^i)} Y_{ij}, \quad V_0^i > 0 \quad (22)$$

where the Brownian motions of different firms are again correlated *via* a factor decomposition. The novelty in their approach is the use of a Poisson process N_t as ticker for jumps in the market that is thinned-out with probability $(1 - b^i)$ to induce jumps in V^i . Consequently, some but not necessarily all firms jump (and possibly default) together. As a result of common jumps, the model allows for default clusters that extend the cyclical correlation induced by common continuous factors. For this choice of jump distribution, the marginals of the model can be calibrated to CDS quotes using the Laplace transform of first-passage times of X^i , which is derived in [19]. The multivariate model is solved *via* a Brownian-bridge Monte Carlo simulation in the spirit of the work by Metwally and Atiya [26].

Conclusion

Structural-default models allow for an appealing interpretation of corporate default: companies operate as long as they have sufficient assets. A clear economic interpretation also holds for the way dependence is introduced to a portfolio of companies: comovements of the firm-value processes might be seen as the result of common risk factors, to which economic interpretations might also apply. This rationale can also be used to empirically estimate the correlation structure of the model from market data. Summarizing, the dependence structure and univariate marginals are simultaneously explained. Moreover, since each company is modeled explicitly, called *bottom-up approach*, it is also possible to price portfolio derivatives and individual risk consistently; a major advantage over *top-down models* that purely focus on the portfolio-loss process. In addition, the current asset level might be mapped to some credit rating, implying a dynamic model of rating changes including default. Finally, the dynamic nature of the modeled firm-value processes translates to a dynamic model for the default correlation and

the portfolio-loss process; a desired property in risk-management solutions and for the pricing of (exotic) credit-portfolio derivatives.

The downside of multivariate structural-default models lies in the difficulty of translating the model to analytical formulas for default correlations and the portfolio-loss distribution. This becomes especially apparent when the simplifying assumption in [31] and its generalizations are reconsidered; the bottom-up nature of structural-default models is entirely given up in order to compute the portfolio-loss distribution in closed form. The price to pay for a more realistic framework typically is a Monte Carlo simulation. However, if such a simulation is efficiently implemented, a realistic dynamic model for a portfolio of credit-risky assets is available.

Acknowledgments

Research support by Daniela Neykova, Technische Universität München, is gratefully acknowledged.

References

- [1] Albrecher, H., Ladoucette, S. & Schoutens, W. (2007). A generic one-factor Lévy model for pricing synthetic CDOs, in *Advances in Mathematical Finance*, M.C. Fu, R.A. Jarrow, J.J. Yen & R.J. Elliott, eds, Birkhaeuser.
- [2] Altman, E., Resti, A. & Sironi, A. (2004). Default recovery rates in credit risk modeling: a review of the literature and empirical evidence, *Economic Notes* **33**(2), 183–208.
- [3] Andersen, L. & Sidenius, J. (2004). Extensions of the Gaussian copula: random recovery and random factor loadings, *Journal of Credit Risk* **1**(1), 29–70.
- [4] Basel Committee on Banking Supervision (2004). *International Convergence of Capital Measurement and Capital Standards—A Revised Framework*, retrieved from <http://www.bis.org/publ/bcbs107.pdf>.
- [5] Basel Committee on Banking Supervision (2005). *An Explanatory Note on the Basel II IRB Risk Weight Functions*, retrieved from <http://www.bis.org/bcbs/irbrisk-weight.pdf>.
- [6] Baxter, M. (2006). *Dynamic Modelling of Single-name Credits and CDO Tranches*. Working paper, Nomura Fixed Income Quant Group.
- [7] Black, F. & Cox, J. (1976). Valuing corporate securities: some effects of bond indenture provisions, *Journal of Finance* **31**(2), 351–367.
- [8] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, Financial Mathematics Series, Chapman and Hall/CRC.

- [9] Crosbie, P. & Bohn, J. *Modeling Default Risk*, KMV Corporation, retrieved from <http://www.moodyskmv.com/research/files/wp/ModelingDefaultRisk.pdf>.
- [10] Duffie, D. & Lando, D. (2001). The term structure of credit spreads with incomplete accounting information, *Econometrica* **69**, 633–664.
- [11] Frye, J. (2000). Depressing recoveries, *Risk* **13**(11), 106–111.
- [12] Geske, R. (1977). The valuation of corporate liabilities as compound options, *Journal of Financial and Quantitative Analysis* **12**(4), 541–552.
- [13] Giesecke, K. (2004). Correlated default with incomplete information, *Journal of Banking and Finance* **28**(7), 1521–1545.
- [14] Gordy, M. (2000). A comparative anatomy of credit risk models, *Journal of Banking and Finance* **24**(1), 119–149.
- [15] Hull, J., Predescu, M. & White, A. (2005). *The Valuation of Correlation-dependent Credit Derivatives using a Structural Model*. Working paper, retrieved from <http://www.rotman.utoronto.ca/hull/DownloadablePublications/StructuralModel.pdf>
- [16] Hull, J. & White, A. (2004). Valuation of a CDO and an n-th to default CDS without a Monte Carlo simulation, *Journal of Derivatives* **12**(2), 8–23.
- [17] Kalemanna, A., Schmid, B. & Werner, R. (2007). The normal inverse Gaussian distribution for synthetic CDO pricing, *Journal of Derivatives* **14**(3), 80–93.
- [18] Kiesel, R. & Scherer, M. (2007). *Dynamic Credit Portfolio Modelling in Structural Models with Jumps*. working paper, retrieved from http://www.uni-ulm.de/fileadmin/website_uni_ulm/mawi.inst.050/people/kiesel/publications/Kiesel_Scherer_Dec07.pdf.
- [19] Kou, S. & Wang, H. (2003). First passage times of a jump diffusion process, *Advances in Applied Probability* **35**, 504–531.
- [20] Leland, H. (1994). Corporate debt value, bond covenants, and optimal capital structure, *Journal of Finance* **49**(4), 1213–1252.
- [21] Leland, H. & Toft, K. (1996). Optimal capital structure, endogenous bankruptcy, and the term structure of credit spreads, *Journal of Finance* **51**(3), 987–1019.
- [22] Longstaff, F. & Schwartz, E. (1995). A simple approach to valuing risky fixed and floating rate debt, *Journal of Finance* **50**(3), 789–819.
- [23] Lopez, J. (2004). The empirical relationship between average asset correlation, firm probability of default, and asset size, *Journal of Financial Intermediation* **13**(2), 265–283.
- [24] Lucas, D. (1995). Default correlation and credit analysis, *Journal of Fixed Income* **4**(4), 76–87.
- [25] Merton, R. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470. Reprinted as Chapter 12 in Merton, R. (1990) *Continuous-time Finance*, Blackwell.
- [26] Metwally, S. & Atiya, A. (2002). Using Brownian bridge for fast simulation of jump-diffusion processes and barrier options, *The Journal of Derivatives* **10**(1), 43–54.
- [27] Moosbrucker, T. (2006). *Pricing CDOs with Correlated Variance Gamma Distributions*. Research report, Department of Banking, University of Cologne.
- [28] Rebholz, J. (1994). *Planar Diffusions with Applications to Mathematical Finance*, PhD thesis, University of California, Berkeley.
- [29] Schönbucher, P. (2003). *Credit Derivatives Pricing Models: Models, Pricing, Implementation*, Wiley Finance.
- [30] Thomas, H. & Wang, Z. (2005). Interpreting the internal ratings-based capital requirements in Basel II, *Journal of Banking Regulation* **6**, 274–289.
- [31] Vasicek, O. (1987). *Probability of Loss on Loan Portfolio*, KMV Corporation, retrieved from http://www.moodyskmv.com/research/whitepaper/Probability_of_Loss_on_Loan_Portfolio.pdf
- [32] Willemann, S. (2007). Fitting the CDO correlation skew: a tractable structural jump-diffusion model, *The Journal of Credit Risk* **3**(1), 63–90.
- [33] Zhou, C. (2001). An analysis of default correlations and multiple defaults, *Review of Financial Studies* **14**, 555–576.
- [34] Zhou, C. (2001). The term structure of credit spreads with jump risk, *Journal of Banking and Finance* **25**, 2015–2040.

Further Reading

- Lipton, A. (2002). Assets with jumps, *Risk* **15**(9), 149–153.
- Lipton, A. & Sepp, A. (2009). Credit value adjustment for credit default swaps via the structural default model, *The Journal of Credit Risk* **5**(2), 125.

Related Articles

Default Barrier Models; Modeling Correlation of Structured Instruments in a Portfolio Setting; Gaussian Copula Model; Internal-ratings-based Approach; Reduced Form Credit Risk Models.

RÜDIGER KIESEL & MATTHIAS A. SCHERER

CreditRisk+

CreditRisk+ is a portfolio credit risk model developed by the bank Credit Suisse, who published the methodology in 1997 [2].

A portfolio credit risk model is a means of estimating the statistical distribution of the aggregate loss from defaults in a portfolio of loans or other credit-risky instruments over a period of time. More generally, changes in credit quality other than default can be considered, but CreditRisk+ in its original form is focused only on default. The most widely used portfolio credit risk models are undoubtedly the so-called structural models, including models based on the Gaussian copula framework (*see Structural Default Risk Models*). CreditRisk+ performs its calculation in a different way to these models, but it is recognized that CreditRisk+ and Gaussian copula models have a similar conceptual basis. A detailed discussion can be found in [4, 7].

Financial institutions use portfolio credit risk models to estimate aggregate credit losses at high percentiles, corresponding to very bad outcomes (often known as the *tail* of the loss distribution). These estimates are then used in setting and allocating economic capital (*see Economic Capital*) and determining portfolio performance measures such as risk-adjusted return on capital (*see Risk-adjusted Return on Capital (RAROC)*).

Portfolio credit risk models have two elements. The first is a set of statistical assumptions about the effect of economic influences on the likelihood of individual borrowers defaulting, and about how much the individual losses might be when they default. The second element is an algorithm for calculating the resulting loss distribution under these assumptions for a specific portfolio. Unlike most portfolio credit risk models, CreditRisk+ calculates the loss distribution using a numerical technique that avoids Monte Carlo simulation. The other distinction of CreditRisk+ is that it was presented as a methodology rather than as a software implementation. Practitioners and institutions have developed their own implementations, leading to a number of significant variants and improvements of the original model. The model has also been used by regulators and central banks: CreditRisk+ played a role in the

early formulation of the Basel accord (see [5]) and has been used by central banks to analyze country-wide panel data on defaults (an example is reported in [1]).

For these reasons, since its introduction in 1997, CreditRisk+ has consistently attracted the interest of practitioners, financial regulators, and academics, who have generated a significant body of literature on the model. An account of CreditRisk+ and its subsequent developments can be found in [6].

The CreditRisk+ Algorithm

The function of CreditRisk+ is to transform data about the creditworthiness of individual borrowers into a portfolio-level assessment of risk. In most portfolio credit risk models, this step requires Monte Carlo simulation (*see Credit Portfolio Simulation*). However, CreditRisk+ avoids simulation by using an efficient numerical algorithm, as outlined below.

The approach confers advantages in terms of speed of computation and enhanced understanding of the drivers of the resulting distribution: many useful statistics, such as the moments of the loss distribution, are given by simple formulae in CreditRisk+, whose relationship to the risk management features of the situation is transparent. On the other hand, owing to its analytic nature, CreditRisk+ is a relatively inflexible portfolio model, and as such has tended to find application where transparency and ease of calculation are more important than flexible parameterization.

To understand the CreditRisk+ calculation, we consider a portfolio containing N loans, where we wish to assess the loss distribution over a one-year time horizon. (The model can be applied to bonds or derivatives counterparties, but the main features of the calculation are the same.) To run CreditRisk+, a number R of economic factors must be chosen. This can be the number of distinct economic influences on the portfolio that are considered to exist (say, the number of geographical regions or industries significantly represented in the portfolio), but it is often assumed in practice that $R = 1$, in which case the model is said to be in “one-factor” mode. CreditRisk+ with one factor gives an assessment of risk that ignores subtle industry or geographic diversification, but can capture the correct overall amount of economic and concentration risk present

in the portfolio, and is sufficient for many purposes. In any event, typically R is much less than N , the number of loans, reflecting the fact that all the significant influences on the portfolio affect many borrowers at once.

For each loan i , where $1 \leq i \leq N$, the model needs the following input data:

1. Long-term average probability of default p_i : This is the probability that the obligor will default over the year, typically estimated from the credit rating (*see Credit Rating*).
2. Loss on default E_i : This is typically estimated as the loan notional less an estimated recovery amount (*see Recovery Rate*).
3. Economic factor loadings: These are given by $\theta_{i,j}$, for $1 \leq j \leq R$, where R is the number of factors introduced above. $\theta_{i,j}$ must be nonnegative numbers satisfying $\sum_{j=1}^R \theta_{i,j} = 1$ for each i .

The factor loadings $\theta_{i,j}$ require some further explanation: they represent the sensitivity of the obligor i to each of the R economic factors assumed to influence the portfolio. In general, determining suitable values for $\theta_{i,j}$ is one of the main difficulties of using CreditRisk+, and analogous difficulties exist for all portfolio models. Note, however, that if R is chosen to be 1 (“one-factor mode” as described above), then we must have $\theta_{i,1} = 1$ for all i , and there is no information requirement. This reflects the fact that one-factor mode ignores the subtle industry or geographic diversification effects in the portfolio, but is, nevertheless, a popular mode of use of the model due to the simpler parameter requirements.

To understand how CreditRisk+ processes this data, let $X_1, \dots, X_R$ be random variables, each with mean $E(X_j) = 1$. The variable X_j represents the economic influence of sector j over the year. In common with most portfolio credit risk models, CreditRisk+ does not incorporate economic prediction. Instead, uncertainty about the economy is reflected by representing economic factors as random variables in this way. CreditRisk+ then assumes that the realized probability of default P_i for loan i is given by the following critical relationship:

$$P_i = p_i(\theta_{i,1}X_1 + \dots + \theta_{i,R}X_R) \quad (1)$$

The realized default probability P_i depends not only on the long-term average probability of default

p_i but also on the random variables $X_1, \dots, X_R$. Note that because $E(X_j) = 1$ and $\sum_{j=1}^R \theta_{i,j} = 1$, we have

$$E(P_i) = p_i(\theta_{i,1} + \dots + \theta_{i,R}) = p_i \quad (2)$$

so that the long-term average default probability (or equivalently, the average of the default probabilities across all states of the economy) is p_i as required. In a particular year, however, P_i will differ from its long-term average. If the borrower i is sensitive to a factor j , (i.e., $\theta_{i,j} > 0$), and if a large value is drawn for X_j , then this represents a poor economy with a negative impact on the obligor i , and we will tend to have $P_i > p_i$, meaning that the obligor i is more likely to default in this particular year than on average. Because the same will be true of other obligors i' with $\theta_{i',j} > 0$, the economic influence represented by factor j can affect a large number of obligors at once. This mechanism incorporates systematic risk, which affects many obligors at once and so cannot be diversified away. The same mechanism in various forms is present in all commonly used portfolio credit risk models.

Two technical assumptions are now made in CreditRisk+:

1. The random variables X_j , $1 \leq j \leq R$, are independent, and each has a Gamma distribution with mean 1 and variance β_j .
2. For each loan i , $1 \leq i \leq N$, the loss given default E_i is a positive integer.

The first assumption is made to facilitate the CreditRisk+ numerical algorithm. In other credit risk models, notably the Gaussian copula models, the variables that play the role of the X_j are assumed to be normally distributed. Although these assumptions seem very different, in fact for many applications they have little effect on the final risk estimate. Assumption (1) can, however, lead to difficulties in parameterizing CreditRisk+.

The second assumption, known as *bucketing* of exposures, also requires some further explanation. Without this assumption, E_i could be any positive amounts, all expressed in units of a common reference currency. An insight of CreditRisk+ is that the precise values of E_i are not critical: E_i can be rounded to whole numbers without significantly affecting the aggregate risk assessment (a simple way of estimating the resulting error is given in Section A4.2 of [2]). The amount of rounding depends on

how E_i are expressed before rounding; for example, it is common to express E_i in millions, so that a loss on default of say 24.35, meaning 24.35 million units of the reference currency, would be rounded to 25.

After bucketing of exposures, the aggregate loss from the portfolio must itself be a whole number (in the example above, this would mean a whole number of millions of the reference currency). The loss distribution can therefore be summarized in terms of its *probability generating function*

$$G(z) = \sum_{n=0}^{\infty} A_n z^n \quad (3)$$

where A_n denotes the probability that the aggregate loss is exactly n . To obtain the loss distribution, we need the numerical value of A_n , for $n = 0$ (corresponding to no loss), $1, 2, \dots$ up to a desired point. For CreditRisk+, with the inputs described above, it can be shown that the probability generating function (3) is given explicitly as

$$G(z) = \prod_{j=1}^R \left(1 - \beta_j \left(\sum_{i=1}^N \theta_{i,j} p_i (z^{E_i} - 1) \right) \right)^{-1/\beta_j} \quad (4)$$

For the derivation of this equation, see, for example [2], Section A9 or [6], Chapter 2. The derivation involves a further approximation, known as the *Poisson approximation*, which can roughly be described as assuming that the default probabilities p_i are small enough that their squares can be neglected. CreditRisk+ then uses an approach related to the so-called Panjer algorithm, which was developed originally for use in actuarial aggregate claim estimation. This relies on the fact that there exist polynomials $P(z)$ and $Q(z)$, whose coefficients can be computed explicitly from the input data *via* equation (4), and which satisfy

$$P(z) \frac{dG(z)}{dz} = Q(z)G(z) \quad (5)$$

Equating the coefficients of z^n on each side of this identity, for each $n \geq 0$, leads finally to a simple *recurrence relationship* between A_n in equation (3). The recurrence relationship expresses the value of A_n for each n , in terms of the earlier coefficients $A_0, \dots, A_{n-1}$. The calculation is started

by calculating A_0 , which is the probability of no loss, by setting $z = 0$ in equation (4) to give the explicit formula

$$A_0 = G(0) = \prod_{j=1}^R \left(1 + \beta_j \left(\sum_{i=1}^N \theta_{i,j} p_i \right) \right)^{-1/\beta_j} \quad (6)$$

and the recurrence relation then allows efficient calculation of A_n up to any desired level. For a complete treatment of this algorithm, see, for example, [6], Chapter 2.

Later Developments of CreditRisk+

Many enhancements to CreditRisk+ have been proposed by various authors (see the introduction to [6] for a discussion of some of the drawbacks of the original model). Developments have fallen into the following broad themes:

1. alternative calculation algorithms, such as saddlepoint approximation, Fourier inversion, and the method of Giese [3];
2. improved capital allocation methods, notably the method of Haaf and Tasche;
3. inclusion of additional risks, such as migration risk and uncertain recovery rates;
4. improved methods for determining inputs, particularly the economic factor loadings $\theta_{i,j}$;
5. application to novel situations such as default probability estimation [8]; and
6. asymptotic formulae, notably the application of the “granularity adjustment” [5].

The reader is also referred to [6] for details on many of these developments.

References

- [1] Balzarotti, V., Castro, C. & Powell, A. (2004). *Reforming Capital Requirements in Emerging Countries: Calibrating Basel II using Historical Argentine Credit Bureau Data and CreditRisk+*. Working Paper, Universidad Torcuato Di Tella, Centro de Investigación en Finanzas.
- [2] Credit Suisse Financial Products (1997). *CreditRisk+, a Credit Risk Management Framework*, Credit Suisse Financial Products, London.
- [3] Giese, G. (2003). Enhancing CreditRisk+, *Risk* **16**(4), 73–77.
- [4] Gordy, M. (2000). A comparative anatomy of Credit Risk Models, *Journal of Banking and Finance* **24**, 119–149.

- [5] Gordy, M. (2004). Granularity adjustment in portfolio Credit Risk Measurement, in *Risk Measures for the 21st Century*, G. Szego, ed., John Wiley & Sons, Heidelberg.
- [6] Grundlach M. & Lehrbass, F. (eds) (2004). *CreditRisk+ in the Banking Industry*, Springer Finance.
- [7] Koyluoglu, H.U. & Hickman, A. (1998). Reconcilable differences, *Risk* **11**(10), 56–62.
- [8] Wilde, T. & Jackson, L. (2006). Low default portfolios without simulation, *Risk* **19**(8), 60–63.

Related Articles

Credit Risk; Gaussian Copula Model; Structural Default Risk Models.

TOM WILDE

Large Pool Approximations

The loss distribution of a large credit portfolio can be valued by Monte Carlo methods. This is perhaps the most common approach used by practitioners today. The problem is that Monte Carlo methods are computationally intensive usually taking a significant amount of time to achieve the required accuracy. Therefore, although such methods may lend themselves to pricing and structuring of credit derivatives, they are not appropriate for risk management where simulation and stress testing are required. In fact, nesting a second level of simulation, for pricing, within the risk management simulation represents a performance challenge.

Analytical approximations of losses of large portfolios represent an efficient alternative to Monte Carlo simulation. The following methods can be applied for approximation of a large portfolio's loss distribution: the law of large numbers (LLN), the central limit theorem (CLT), and large deviation theory.

The analytical methods for approximation of credit portfolio losses are usually applied in an additive scheme: the portfolio losses due to default, $\mathcal{L}$, over some fixed time horizon (single step) are represented as

$$\mathcal{L} = \sum_{k=1}^K L_k \quad (1)$$

where L_k is the loss of the k th name in the portfolio and K is the number of names. Application of limit theorems for stochastic process becomes quite natural as K increases. The main technical difficulties are related to dependency of default events and losses of the counterparties.

The analytical methods for portfolio losses are applied in the conditional independence framework pioneered in [14] (see also [7, 9]), based on the assumption that there is a random vector, X , such that conditional on the values of X , the default events are independent. Usually, X is interpreted as a vector of credit drivers describing the state of the economy or a sector of the economy, at the end of the time horizon [3, 5, 8, 12]. In multistep models, X can be a random process describing the dynamics of the credit drivers [7]. In this case, computation of conditional default and migration probabilities requires efficient

numerical quadratures for multidimensional integrals [7]. The multistep portfolio modeling is applied when it is necessary to incorporate the effect of stochastic portfolio exposure, as in the integrated market and credit risk framework in [7].

The notation $\mathbb{P}_x$ denotes the regular conditional probability measure, conditional on $X = x$; $\mathbb{E}_x$ is the corresponding conditional expectation operator.

A general approach to approximate the distribution of the random variable $\mathcal{L}$ can be described as follows:

1. Choose a sufficiently rich family of distributions, F_θ , such that $\theta \mapsto F_\theta$ is a Borel measurable mapping of a vector of parameters θ .
2. Fix a value of the variable, $X = x$ and compute parameters, $\theta(x)$, of the approximating family of distributions, $F_{\theta(x)}(\ell)$, such that the conditional distribution, $\mathbb{P}_x(\mathcal{L} \leq \ell) = \mathbb{P}(\mathcal{L} \leq \ell | X = x)$, is approximated by $F_{\theta(x)}(\ell)$. (It is assumed that $x \mapsto \theta(x)$ is Borel measurable, so that $x \mapsto F_{\theta(x)}(\ell)$ is also measurable, for each ℓ .)
3. Find the unconditional approximating distribution by integration over the distribution, G_X , of the variable X :

$$F^*(\ell) = \int F_{\theta(x)}(\ell) dG_X(x) \quad (2)$$

Law of Large Numbers: Vasicek Approximation

The first key result was obtained in [14, 13] for homogeneous portfolios. The K random variables, L_k , can be expressed as $L_k = N \cdot I_k$, where I_k is the indicator of default of the k th name and N is the constant loss given default. The random variables I_k are identically distributed and their sum, $\nu = \sum_{k=1}^K I_k$, is the number of names in default. The portfolio losses $\mathcal{L} = N\nu$.

The variable, X , in the Vasicek model is latent and has a standard normal distribution, $\Phi(x)$. Conditional on $X = x$, the default events are independent and ν has a binomial distribution with parameter $p(x) = \mathbb{P}(I_k = 1 | X = x)$ so that

$$\int_{-\infty}^{\infty} p(x) d\Phi(x) = \pi_* \quad (3)$$

where π_* is the common unconditional probability of default. The unconditional distribution of ν is then a

generalized binomial distribution and

$$\begin{aligned}\mathbb{P}(\mathcal{L} = mN) &= \mathbb{P}(v = m) \\ &= \binom{K}{m} \int_{-\infty}^{\infty} p^m(x) q^{K-m}(x) d\Phi(x), \\ m &= 0, 1, \dots, K\end{aligned}\quad (4)$$

where $q(x) = 1 - p(x)$.

The following specification^a of the conditional default probability is widely used in the literature [6, 14], and so on.

$$p(x) = \Phi\left(\frac{H - \beta x}{\sigma}\right) \quad \beta^2 + \sigma^2 = 1 \quad (5)$$

where $H = \Phi^{-1}(\pi_*)$, and β is a parameter that determines the correlation between default events.

Consider the ratio $v_K = v/K$ determining the portfolio losses. If $\beta = 0$, then $p(x) \equiv \pi_*$ and

$$\lim_{K \rightarrow \infty} v_K = \pi_* \quad \text{almost surely} \quad (6)$$

in accordance with the strong law of large numbers. If $\beta \neq 0$ the limit in equation (6) is in distribution, to a random variable with the same distribution as $\xi = p(X)$. Thus, one obtains

$$\lim_{K \rightarrow \infty} \mathbb{P}(v_K \leq \ell) = \Phi\left(\frac{\sigma \Phi^{-1}(\ell) - H}{\beta}\right), \quad 0 \leq \ell \leq 1 \quad (7)$$

It follows from equation (7) that the quantile approximation, ℓ_q^* , corresponding to the probability q , is

$$\ell_q^* = N \Phi\left(\frac{\beta \Phi^{-1}(q) + H}{\sigma}\right) \quad (8)$$

In terms of the general approach, one has $\theta = \mu$ with $F_\theta(\ell) = \mathbb{1}_{[\mu, \infty)}(\ell)$ and $\theta(x) \equiv \mu(x) = KNp(x)$.

Central Limit Theorem

The heterogeneous case is treated at the outset, as it is no more difficult than the homogeneous case, which is described as a special case at the end. Once again, X is univariate and latent. Denote by N_k , the loss

given default of the k th name in the portfolio, and by $p_k(x)$, the conditional default probability of the k th name. Then the conditional mean, $\mu(x)$, and the conditional variance, $\sigma^2(x)$, of the portfolio losses are

$$\begin{aligned}\mu(x) &= \sum_{k=1}^K N_k p_k(x), \\ \sigma^2(x) &= \sum_{k=1}^K N_k^2 p_k(x) \cdot (1 - p_k(x))\end{aligned} \quad (9)$$

Under mild conditions on the notionals, N_k (which are vacuous in the homogeneous case), the conditional distribution of the portfolio losses satisfies

$$\mathbb{P}_x\left(\frac{\mathcal{L} - \mu(x)}{\sigma(x)} \leq \ell\right) \rightarrow \Phi(\ell) \quad \text{as } K \rightarrow \infty \quad (10)$$

Let a probability, q , $0 < q < 1$, be fixed and consider the equation

$$q = \mathbb{P}(\mathcal{L} \leq \ell_q) \quad (11)$$

for the quantile of the distribution of the random variable $\mathcal{L}$. One has

$$\begin{aligned}\mathbb{P}(\mathcal{L} \leq \ell_q) &= \int_{-\infty}^{\infty} \mathbb{P}_x(\mathcal{L} \leq \ell_q) d\Phi(x) \\ &\doteq \int_{-\infty}^{\infty} \Phi\left(\frac{\ell_q - \mu(x)}{\sigma(x)}\right) d\Phi(x)\end{aligned} \quad (12)$$

Therefore the quantile approximation, ℓ_q^* , is the solution of the equation

$$q = \int_{-\infty}^{\infty} \Phi\left(\frac{\ell_q^* - \mu(x)}{\sigma(x)}\right) d\Phi(x) \quad (13)$$

In terms of the general approach, one has $\theta = (\mu, \sigma)$ with $F_\theta(\ell) = \Phi\left(\frac{\ell - \mu}{\sigma}\right)$ and $\theta(x) = (\mu(x), \sigma(x))$.

In the case of a homogeneous portfolio, considered in [14, 13] one has the simplifications

$$\mu(x) = KNp(x) \quad \sigma^2(x) = KN^2(p(x) - p^2(x)) \quad (14)$$

The normal approximation is just the classical central limit theorem (CLT). The equation for the quantile approximation simplifies to

$$q = \int_{-\infty}^{\infty} \Phi \left(\frac{\ell_q^*/N - Kp(x)}{\sqrt{Kp(x)(1-p(x))}} \right) d\Phi(x) \quad (15)$$

Generalized Poisson Approximation

Consider a homogeneous portfolio, for which the number, K , of obligors is moderately large but not very large. If also the conditional mean number of default events in the portfolio, $K \cdot p(x)$, takes moderate values, the conditional distribution of v might be better approximated by a Poisson distribution

$$\mathbb{P}_x(v = m) = \exp(-\lambda(x)) \frac{\lambda^m(x)}{m!}, \quad m = 0, 1, 2, \dots \quad (16)$$

than by a normal distribution, where $\lambda(x) = Kp(x)$. In this case, the (unconditional) portfolio losses can be approximated by the generalized Poisson distribution: for moderately large K ,

$$\mathbb{P}(\mathcal{L} = mN) = \int e^{-\lambda(x)} \frac{\lambda^m(x)}{m!} dG_X(x), \quad m = 0, 1, 2, \dots \quad (17)$$

In terms of the general approach, one has F_θ being the Poisson distribution function with mean θ and $\theta(x) = \lambda(x)$.

In particular, for the quantile approximation, one obtains

$$q = \int \sum_{m=0}^{\ell_q^*/N} e^{-\lambda(x)} \frac{\lambda^m(x)}{m!} dG_X(x) \quad (18)$$

Compound Poisson Approximation

In order to extend the result of the previous section to heterogeneous portfolios, one needs to consider compound Poisson distributed random variables. The compound Poisson distribution is a well known approximation in insurance models [11]. In risk management of credit derivatives, the approach

was used in [2] and [6] for synthetic collateralized debt obligation (CDO) pricing. The same approach is applicable for approximation of portfolio losses.

In the case of a heterogeneous portfolio, it is not sufficient to approximate the distribution of the number of losses suffered. One must keep track of who defaults or at least the sizes of the individual potential losses because, given only the number of defaults, one cannot infer the losses incurred. To see how this added complexity is handled and how the compound Poisson distribution arises quite naturally, the simplest heterogeneous case is analyzed first; namely, when there are only two distinct recovery-adjusted notional values among the obligors in the portfolio.

Denote by $N_{(1)}$ and $N_{(2)}$, the two distinct values of the recovery-adjusted notionals in the pool. The portfolio then divides into two groups: one with obligors having the common recovery-adjusted notional equaling $N_{(1)}$; the other having common recovery-adjusted notional equaling $N_{(2)}$. Denote the number of defaults in each of the two groups, by v_1 and v_2 , respectively. Conditionally, their distributions are independent and can be approximated by a Poisson distribution with conditional mean $\lambda_i(x) = \sum_{k: N_k = N_{(i)}} p_k(x)$, $i = 1, 2$, provided both group sizes are moderately large. (This assumption on the group sizes, is only being made in the context of this example.) The total number of defaults in the portfolio, $v = v_1 + v_2$, is conditionally Poisson with conditional mean $\lambda(x) = \lambda_1(x) + \lambda_2(x)$. The total portfolio loss is the sum of the losses of the first and second groups:

$$\mathcal{L} = v_1 N_{(1)} + v_2 N_{(2)} \quad (19)$$

As a positive linear combination of conditionally independent Poisson random variables, $\mathcal{L}$ is conditionally a compound Poisson random variable with the same distribution as that of

$$\tilde{\mathcal{L}} := \sum_{j=1}^v \mathcal{N}^{(j)} \quad (20)$$

where $\mathcal{N}^{(j)}$ is a conditionally independent and identically distributed (i.i.d.) sequence of random variables, each taking two values, $N_{(1)}$ and $N_{(2)}$, with corresponding conditional probabilities $\frac{\lambda_1(x)}{\lambda(x)}$ and $\frac{\lambda_2(x)}{\lambda(x)}$ and conditionally independent of v . (This is an elementary

calculation using the conditional characteristic functions of $\mathcal{L}$ and $\tilde{\mathcal{L}}$.) More formally, the conditional distribution of $\mathcal{N}^{(j)}$ is

$$f(N; x) \equiv \mathbb{P}_x(\mathcal{N}^{(j)} = N) = \begin{cases} \frac{\lambda_1(x)}{\lambda(x)}, & N = N_{(1)} \\ \frac{\lambda_2(x)}{\lambda(x)}, & N = N_{(2)} \end{cases} \quad (21)$$

In the general case where the recovery-adjusted notionals take more than two values, the conditional distribution of the random variable, $\mathcal{N}^{(j)}$, is

$$f(N; x) = \sum_{k: N_k = N} p_k(x) / \lambda(x) \quad (22)$$

where $\lambda(x) = \sum_{k=1}^K p_k(x)$ and N represents a possible individual loss.

In the special case where p_k does not depend on k , f is simply the relative frequency of the notional values and does not depend on x :

$$f(N) = [\#k \in \{1, 2, \dots, K\} : N_k = N] / K \quad (23)$$

In general, the function $f(N; x)$ is a probability mass function with respect to N , which approximates the conditional probability that the portfolio loss is of size N , given that there has been only one default.

More generally, it can be shown that

$$\mathbb{P}_x(\mathcal{L} = N | v = m) \approx f^{*m}(N; x) \quad (24)$$

where f^{*m} denotes the m -fold convolution of f with itself, as a probability mass function (for notational convenience, $f^{*1} \equiv f$ and $f^{*0}(N; x) = 1$ if and only if $N = 0$). Given that there have been exactly m defaults, the pool loss amounts to a sum of m notional amounts but, as one does not know who defaulted, in the heterogeneous case there is still some randomness left; that randomness is captured (approximately) by f^{*m} .

Assuming that a monetary unit has been chosen and that all recovery-adjusted notionals are expressed as integers—that is, integer multiples of the monetary unit—one has the following result [6]:

Theorem 1 *In the limiting case of a large portfolio (K large), the following approximate equality holds in distribution under $\mathbb{P}_x$ (i.e., conditional on $X = x$):*

for fixed $i = 1, 2, \dots, n$,

$$\mathcal{L} \stackrel{\mathcal{D}}{\approx} \sum_{m=1}^v \mathcal{N}^{(m)} \quad (25)$$

where $(\mathcal{N}^{(m)})_{m=1}^K$ is an i.i.d. sequence of random variables with common probability mass function f and independent of v , the number of defaults in the pool, which is approximately Poisson distributed under $\mathbb{P}_x$

$$v \stackrel{\mathcal{D}}{\approx} \text{Pois}(\lambda(x)) \quad (26)$$

More precisely,

$$\begin{aligned} \max_N \left| \mathbb{P}_x(\mathcal{L} = N) - \mathbb{P}_x\left(\sum_{m=1}^v \mathcal{N}^{(m)} = N\right) \right| \\ = \mathcal{O}\left(\sum_{k=1}^K (p_k(x))^2\right) \end{aligned} \quad (27)$$

For the unconditional loss distribution,

$$\mathbb{P}(\mathcal{L} \leq \ell) = \int \sum_{N \leq \ell} \sum_{m=0}^{\infty} f^{*m}(N; x) \frac{e^{-\lambda(x)} \lambda^m(x)}{m!} dG_X(x) \quad (28)$$

In terms of the general approach, one has F_θ being the compound Poisson distribution function with parameter $\theta = (\theta_1, \theta_2, \dots, \theta_K) \in [0, 1]^K$ and $\theta(x) = (p_1(x), p_2(x), \dots, p_K(x))$; F_θ is defined as

$$F_\theta = \sum_{N \leq \ell} \sum_{m=0}^{\infty} f^{*m}(N) \frac{e^{-\lambda} \lambda^m}{m!} \quad (29)$$

where $\lambda := \sum_{k=1}^K \theta_k$, $f(N) := \lambda^{-1} \sum_{k: N_k = N} \theta_k$. In practice, the convolutions, would be calculated recursively using the fast Fourier transform.

Large Deviations

Approximations based on large deviation theory usually lead to exponential approximations of the tail of the conditional portfolio loss distribution. These approximations are derived using the saddlepoint

method for the characteristic function of the portfolio losses,

$$\begin{aligned}\Psi_{\mathcal{L}}(s) &= \mathbb{E}[\exp(is\mathcal{L})] \\ &= \int \prod_{k=1}^K (1 - p_k(x) + p_k(x)e^{isN_k}) dG_X(x)\end{aligned}\quad (30)$$

The technical details can be found in [1] (*see also Saddlepoint Approximation*).

Other Methods

There are some methods of approximation that deal only with quantiles of the loss distribution directly, focusing on quantiles with high quantile probability, which is the case of interest for credit risk. The large deviation approximations are examples of such methods.

Another one of these methods is due to Pykhtin [12] who, building on the work of Martin and Wilde [10], adapted the tools of an earlier investigation [4] in market-risk sensitivity to position sizes, to the credit risk setting. Note that this method is a direct, analytical approximation to the quantile of the *unconditional* loss distribution using an approximate model, unlike the other semianalytic methods described so far, which calculate the quantile by making analytical approximations to the conditional loss distribution (conditional on a systemic credit scenario). It is also worth noting that the result is in closed form, a qualitative description of which is given here.

Pykhtin's approach can be described at a high level as follows. It consists of a three-stage series of approximations:

1. A single-factor model, which is an approximation based on an LLN type of loss function; that is, it is a Vasicek type of model.
 - a) The single factor is built as a weighted sum of the portfolio's counterparties' credit drivers.
 - b) The weights are chosen to maximize the single factor's correlation with the drivers.
 - c) The weights use the counterparties' loss characteristics such as default probabilities and losses given default.
2. An analytic adjustment (approximation) to a full multifactor model that is still based on an LLN type of loss function. This adjustment is called a *multifactor* adjustment.
3. An analytic adjustment, bridging the LLN-type loss function of the second stage to the usual Merton-type one with full specific risk. This adjustment is called a *granularity* adjustment.

The reason behind the terminology for the two adjustments, is that for a single-factor model, the multifactor adjustment vanishes, whereas for an infinitely granular portfolio (i.e., a very large, homogeneous one), the granularity adjustment vanishes.

The approximations, in both the second and third stages, are based on a single formula for quantile approximation, due originally to Gouriéroux *et al.* [4]. The formula is a second-order Taylor expansion, for the quantile, in a small parameter that is used to express the full loss model as a perturbation of the single-factor model. The first-order Taylor coefficient is the difference between the single-factor (conditional) loss and the conditional expected loss of the full model, conditional on the single factor. The single factor is constructed so that the first-order Taylor term vanishes.

The second-order Taylor coefficient is related to the conditional variance of the full loss, conditional on the single factor. The well-known conditional variance decomposition from statistics is used to split the Taylor coefficient into two terms, which are the approximations in the second and third stages.

The end result for the entire adjustment to the single-factor quantile, is expressed as a sum of four quadratic forms in the recovery-adjusted exposures, with coefficients involving the bivariate and univariate normal cumulative distribution functions, evaluated in terms of the input statistical parameters of the model. The result is thus in closed form. The reader is referred to [12] for the quantitative details of the construction, the formulae for the terms in the quantile approximation, and a study of the scope of applicability of the method.

End Notes

^aThis specification is a partial case of the famous Gaussian copula model [9].

References

- [1] Dembo, A., Deuschel, J.-D. & Duffie, D. (2004). Large portfolio losses, *Finance and Stochastics* **8**(1), 3–16.
- [2] De Prisco, B., Iscoe, I. & Kreinin, A. (2005). Loss in translation, *Risk* **18**(6), 77–82.
- [3] Gordy, M. (2003). A risk-factor model foundation for ratings-based bank capital rules, *Journal of Financial Intermediation* **12**(3), 199–232.
- [4] Gouriéroux, C., Laurent, J.-P. & Scaillet, O. (2000). Sensitivity analysis of values at risk, *Journal of Empirical Finance* **7**, 225–245.
- [5] Huang, X., Oosterlee, C. & Mesters, M. (2007). Computation of VaR and VaR contribution in the Vasicek portfolio credit loss model: a comparative study, *The Journal of Credit Risk* **3**(3), 75–96.
- [6] Iscoe, I. & Kreinin, A. (2007). Valuation of synthetic CDOs, *Journal of Banking and Finance* **31**, 3357–3376.
- [7] Iscoe, I., Kreinin, A. & Rosen, D. (1999). Integrated market and credit risk portfolio model, *Algorithmics Research Quarterly* **2**(3), 21–38.
- [8] Koyluoglu, H.U. & Hickman, A. (1998). *A Generalized Framework for Credit Risk Portfolio Models*, Working paper, CSFP Capital.
- [9] Li, D. (1999). *On Default Correlation: A Copula Function Approach*, The RiskMetrics group, Working paper, 99-07.
- [10] Martin, R. & Wilde, T. (2002). Unsystematic credit risk, *Risk* **15**(11), 123–128.
- [11] Panjer, H. & Willmot, G. (1992). *Insurance Risk Models*, Society of Actuaries, Schaumburg.
- [12] Pykhtin, M. (2004). Multi-factor adjustment, *Risk* March, 85–90.
- [13] Vasicek, O. (1987). *Probability of Loss on Loan Portfolio*, KMV, available at www.kmv.com
- [14] Vasicek, O. (2002). Loan portfolio value, *Risk*, December.

Further Reading

- Emmer, S. & Tasche, D. (2003). *Calculating Credit Risk Capital Charges with the One-Factor Model*, Working Paper, September 2003.
- Gordy, M. (2002). Saddlepoint approximations of credit risk, *Journal of Banking and Finance* **26**, 1335–1353.
- Gordy, M. & Jones, D. (2003). Random tranches, *Risk* March, 78–83.
- Gregory, J. & Laurent, J.-P. (2003). I will survive, *Journal of Risk* **16**(6), 103–108.
- Hull, J. & White, A. (2003). Valuation of a CDO and an n^{th} to default CDS without Monte Carlo simulation, *Journal of Derivatives* **12**(2), 8–23.
- Laurent, J.-P. & Gregory, J. (2003). Basket default swaps, CDO's and factor copulas, *Presentation at the Conference Quant'03*, London, September 2003, p. 21, www.defaultrisk.com
- Schönbucher, P. (2003). *Credit Derivatives Pricing Models*, John Wiley & Sons.

IAN ISCOE & ALEX KREININ

Saddlepoint Approximation

The classical method known variously as the *saddlepoint approximation*, the *method of steepest descents*, the *method of stationary phase*, or the *Laplace method*, applies to contour integrals that can be written in the form

$$I(s) = \int_{\mathcal{C}} e^{sf(\zeta)} d\zeta \quad (1)$$

where f , an analytic function, has a real part that goes to minus infinity at both ends of the contour $\mathcal{C}$. The fundamental idea is that the value of the integral when $s > 0$ is large should be dominated by contributions from the neighborhoods of points where the real part of f has a saddlepoint. Early use was made of the method by Debye to produce asymptotics of Bessel functions, as reviewed in, for example, [8]. Daniels [3] wrote a definitive work on the saddlepoint approximation in statistics. Later, these ideas evolved into the theory of large deviations, initiated by Varadhan in [7], which seeks to determine rigorous asymptotics for the probability of rare events.

If we write $\zeta = x + iy$, elementary complex analysis implies that the surface over the (x, y) plane with graph $\Re f$ has zero mean curvature, so any critical point ζ^* (a point where $f' = 0$) will be a saddlepoint of the modulus $|e^{sf(\zeta)}|$. The level curves of $\Re f$ and $\Im f$ form families of orthogonal trajectories: the curves of steepest descent of $\Re f$ are the level curves of $\Im f$, and *vice versa*. Thus the curve of steepest descent of the function $\Re f$ through ζ^* is also a curve on which $\Im f$ is constant. In other words, it is a curve of “stationary phase”. On such a curve, the modulus of $e^{sf(\zeta)}$ will have a sharp maximum at ζ^* . If the contour $\mathcal{C}$ can be deformed to follow the curve of steepest descent through a unique critical point ζ^* , and the modulus of $e^{sf(\zeta)}$ is negligible elsewhere, the dominant contribution to the integral for large s can be computed by a local computation in the neighborhood of ζ^* . In more complex applications, several critical points may need to be accounted for.

The tangent line to the steepest descent curve at ζ^* can be parameterized by $w \in \mathbb{R}$ by the equation

$$(sf^{(2)}(\zeta^*))^{1/2}(\zeta - \zeta^*) = iw \quad (2)$$

(care is needed here to select the correct sign of the complex square root), and on this line, the Taylor expansion of f about ζ^* implies

$$f(z) = f(\zeta^*) + \sum_{n \geq 2} \frac{1}{n!} f^{(n)}(\zeta^*) \left(\frac{iw}{s(f^{(2)}(\zeta^*))^{1/2}} \right)^n \quad (3)$$

One can write the integrand in the form

$$\begin{aligned} e^{sf(z)} &\sim e^{sf(\zeta^*) - w^2/2} \left[1 - is^{-1/2} \frac{f^{(3)}(\zeta^*)}{3!(f^{(2)}(\zeta^*))^{3/2}} w^3 \right. \\ &\quad \left. + s^{-1} \left(\frac{f^{(4)}(\zeta^*)}{4!(f^{(2)}(\zeta^*))^2} w^4 \right. \right. \\ &\quad \left. \left. - \frac{(f^{(3)}(\zeta^*))^2}{2!(3!)^2(f^{(2)}(\zeta^*))^3} w^6 \right) + \dots \right] \quad (4) \end{aligned}$$

Now approximating the integral over $\mathcal{C}$ by the integral over the tangent line parameterized by w leads to a series of Gaussian integrals, each of which can be computed explicitly. The terms with an odd power of w all vanish, leading to the result

$$\begin{aligned} I(s) &\sim i \left(\frac{2\pi}{sf^{(2)}(\zeta^*)} \right)^{1/2} e^{sf(\zeta^*)} \\ &\quad \times \left[1 + s^{-1} \left(\frac{3f^{(4)}(\zeta^*)}{4!(f^{(2)}(\zeta^*))^2} \right. \right. \\ &\quad \left. \left. - \frac{5 \cdot 3 \cdot (f^{(3)}(\zeta^*))^2}{2!(3!)^2(f^{(2)}(\zeta^*))^3} \right) + \dots \right] \quad (5) \end{aligned}$$

Daniels’ Application to Statistics

Daniels [3] presented an asymptotic expansion for the probability density function (pdf) $f_n(x)$ of the mean $\bar{X}_n$ of n i.i.d. copies of a continuous random variable X with cumulative probability function $F(x)$ and pdf $f(x) = F'(x)$. Assuming that the moment generating function

$$M(\tau) = e^{\Psi(\tau)} = \int_{-\infty}^{\infty} e^{\tau x} f(x) dx \quad (6)$$

is finite for τ in an open interval $(-c_1, c_2)$ containing the origin, the Fourier inversion theorem implies that

$$f_n(x) = \frac{n}{2\pi i} \int_{\alpha - i\infty}^{\alpha + i\infty} e^{n(\Psi(\tau) - \tau x)} d\tau \quad (7)$$

for any real $\alpha \in (-c_1, c_2)$. This integral is now amenable to a saddlepoint treatment as follows.

For each x in the support of f , one can show that the saddlepoint condition

$$\Psi'(\tau) - x = 0 \quad (8)$$

has a unique real solution $\tau^* = \tau^*(x)$. One now evaluates the integral given by equation (7) with $\alpha = \tau^*$, and uses Taylor expansion and the substitution $w = -i\sqrt{n\Psi''(\tau^*)}(\tau - \tau^*)$ to write

$$\begin{aligned} f_n(x) &\sim \frac{\sqrt{n}}{2\pi\sqrt{\Psi''(\tau^*)}} \int_{-\infty}^{\infty} e^{n(\Psi(\tau^*) - \tau^*x) - w^2/2} \\ &\times \left[1 + in^{-1/2}(\Psi''(\tau^*))^{-3/2}\Psi^{(3)}(\tau^*)w^3/3! \right. \\ &\left. + n^{-1}(\Psi''(\tau^*))^{-2}\Psi^{(4)}(\tau^*)w^4/4! + \dots \right] dw \end{aligned} \quad (9)$$

Each term in this expansion is a Gaussian integral that can be evaluated in closed form. The odd terms all vanish, leaving an expansion in powers of n^{-1} :

$$\begin{aligned} f_n(x) &\sim g_n(x) \left[1 + n^{-1} \left(\frac{\Psi^{(4)}(\tau^*)}{8(\Psi''(\tau^*))^2} \right. \right. \\ &\left. \left. - \frac{5(\Psi^{(3)}(\tau^*))^2}{24(\Psi''(\tau^*))^3} \right) + O(n^{-2}) \right] \end{aligned} \quad (10)$$

where the leading term (called the *saddlepoint approximation*) is given by

$$g_n(x) = \left(\frac{n}{2\pi\Psi''(\tau^*)} \right)^{1/2} e^{n(\Psi(\tau^*) - \tau^*x)} \quad (11)$$

The function $I(x) = \sup_{\tau} \tau x - \Psi(\tau) = \tau^*x - \Psi(\tau^*)$ that appears in this expression is the Legendre transform of the cumulant generating function Ψ , and is known as the *rate function* or *Cramér function* of the random variable X . The *large deviation principle*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P(\bar{X}_n > x) = -I(x) \quad \text{for } x > E[X] \quad (12)$$

holds for very general X . Another observation is that the Edgeworth expansion of statistics comes out in a similar way, but takes 0 instead of τ^* as the center of the Taylor expansion.

One can show, using a lemma due to Watson [8], that equation (10) is an *asymptotic expansion*,

which means roughly that when truncated at any order of n^{-1} , the remainder is of the same magnitude as the first omitted term. A more precise statement of the magnitude of the remainder is difficult to establish: the lack of a general error analysis is an acknowledged deficiency of the saddlepoint method.

Applications to Portfolio Credit Risk

The problem of portfolio credit risk measures and the problem of evaluating arbitrage-free pricing of *collateralized debt obligations* (CDOs) both boil down to computation of the probability distribution of the portfolio loss at a set of times, and can be amenable to a saddlepoint treatment. To illustrate this fact, we consider a simple portfolio of credit risky instruments (e.g., corporate loans or credit default swaps), and investigate the properties of the losses caused by default of the obligors. Let $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ be a filtered probability space that contains all of the random elements: P may be either the physical or the risk-neutral probability measure. The portfolio is defined by the following basic quantities:

- M reference obligors with notional amounts N_j , $j = 1, 2, \dots, M$;
- the default time τ_j of the j th credit, an $\mathcal{F}_t$ stopping time;
- the fractional recovery R_j after default of the j th obligor;
- the loss $l_j = (1 - R_j)N_j/N$ caused by default of the j th obligor as a fraction of the total notional $N = \sum_j N_j$;
- the cumulative portfolio loss $L(t) = \sum_j l_j I(\tau_j \leq t)$ up to time t as a fraction of the total notional.

For simplicity, we make the following assumptions:

1. The discount factor is $v(t) = e^{-rt}$ for a constant interest rate $r \geq 0$.
2. The fractional recovery values R_j and hence l_j are deterministic constants.
3. There is a sub σ -algebra $\mathcal{H} \subset \mathcal{F}$ generated by a d -dimensional random variable Y , the “condition”, such that the default times τ_j are mutually conditionally independent under $\mathcal{H}$. The marginal distribution of Y is denoted by P_Y and has pdf $\rho_Y(y)$, $y \in \mathbb{R}^d$.

The most important consequence of these assumptions is that, conditioned on $\mathcal{H}$, the fractional loss $L(t)$ is a sum of independent (but not identical) Bernoulli random variables. For fixed values of the time t and conditioning random variable Y , we note that $\hat{L} := L(t)|_Y \sim \sum_j l_j X_j$ where $X_j \sim \text{Bern}(p_j(t, y))$, $p_j = \text{Prob}(\tau_j \leq t | Y = y)$. The following functions are associated with the random variable $\hat{L}$:

1. the pdf $\rho(x) := F^{(-1)}(x)$ (in our simple example, it is a sum of delta functions supported on the interval $[0, 1]$);
2. the cumulative distribution function (CDF) $F^{(0)}(x) = E[I(\hat{L} \leq x)]$;
3. the higher conditional moment functions $F^{(m)}(x) = (m!)^{-1} E[(x - \hat{L})^+{}^m]$, $m = 1, 2, \dots$;
4. the cumulant generating function (CGF) $\Psi(u) = \log(E[e^{u\hat{L}}])$.

When we need to make explicit the dependence on t, y we write $F^{(m)}(x|t, y)$. The unconditional versions of these functions are given by

$$F^{(m)}(x|t) = E[F^{(m)}(x|t, Y)] = \int_{\mathbb{R}} F^{(m)}(x|t, y) \times \rho_Y(dy), \quad m = -1, 0, \dots \quad (13)$$

According to these definitions, for all $m = 0, 1, \dots$ we have the integration formula

$$F^{(m)}(x) = \int_0^x F^{(m-1)}(z) dz \quad (14)$$

Credit Risk Measures

In risk management, the key quantities that determine the economic capital requirement for such a credit risky portfolio are the *Value at Risk* (VaR) and *Conditional Value at Risk* (CVaR) for a fixed time horizon T and a fixed confidence level $\alpha < 1$. These are defined as follows:

$$\text{VaR}_\alpha(L_T) = \inf\{x | F^{(0)}(x|T) > \alpha\} \quad (15)$$

$$\begin{aligned} \text{CVaR}_\alpha(L_T) &= \frac{E[(L_T - x)^+]}{1 - \alpha} \\ &= \frac{F^{(1)}(x|T) + E[L_T] - x}{1 - \alpha} \end{aligned} \quad (16)$$

Here, we need to take P to be the physical measure.

CDO Pricing

CDOs are portfolio credit swaps that can be schematically decomposed into two types of basic contingent claims whose cash flows depend on the portfolio loss L_t . These cash flows are analogous to insurance and premium payments paid periodically (typically, quarterly) on dates t_k , $k = 1, \dots, K$, to cover default losses within a “tranche” that occurred during that period.

The writer (the *insurer*) of one unit of a default leg for a tranche with attachment levels $0 \leq a < b \leq 1$ pays the holder (the *buyer of insurance*) at each date t_k all default losses within the interval $[a, b]$ that occurred over $[t_{k-1}, t_k]$. The time 0 arbitrage price of such a contract is

$$W_{a,b} = \sum_k e^{-rt_k} E[(b - L_{t_k})^+ - (b - L_{t_{k-1}})^+ - (a - L_{t_k})^+ + (a - L_{t_{k-1}})^+] \quad (17)$$

where E is now the expectation with respect to some risk-neutral measure. The writer of one unit of a premium leg for a tranche with attachment levels $a < b$ (the insured) pays the holder (the insurer) on each date t_k an amount jointly proportional to the year fraction $t_k - t_{k-1}$ and the amount remaining in the tranche. We ignore a possible “accrual term” that account for defaults between payment dates. The time 0 arbitrage price of such a contract is

$$V_{a,b} = \sum_k e^{-rt_k} (t_k - t_{k-1}) E[(b - L_{t_k})^+ - (a - L_{t_k})^+] \quad (18)$$

The *CDO rate* $s_{a,b}$ for this contract at time 0 is the number of units of the premium leg that has the same value as one unit of the default leg, that is, $s_{a,b} = W_{a,b}/V_{a,b}$.

Saddlepoint Approximations for $F^{(m)}$

We see that the credit risk management problem and the CDO pricing problem both boil down to finding an efficient method to compute $E[F^{(m)}(x|t, y)]$ for $m = 0, 1$ and a large but finite set of values (x, t, y) . For the conditional loss $\hat{L} = L_t|Y = y$, the CGF is

explicit

$$\Psi(u) = \sum_{j=1}^M \log[1 - p_j + p_j e^{u l_j}] \quad (19)$$

We suppose that the conditional default probabilities $p_j = p_j(t, y)$ are known. A number of different strategies can be used to compute this distribution accurately:

1. In the fully homogeneous case when $p_j = p, l_j = l$, the distribution is binomial.
2. When $l_j = l$, but p_j are variable (the homogeneous notional case), these probabilities can be computed highly efficiently by a recursive algorithm in [1, 5].
3. When both l_j, p_j are variable, it has been noted in [2, 4, 6, 9] that a saddlepoint treatment of these problems offer superior performance over a naive Edgeworth expansion.

We now consider the fully nonhomogeneous case and begin by using the Laplace inversion theorem to write

$$\rho(x) = F^{(-1)}(x) = \frac{1}{2\pi} \int_{\alpha-i\infty}^{\alpha+i\infty} e^{\Psi(\tau)-\tau x} d\tau \quad (20)$$

Since ρ is a sum of delta functions, this formula must be understood in the distributional sense, and holds for any real α . When $\alpha < 0$,

$$\begin{aligned} F^{(0)}(x) &= \frac{1}{2\pi} \int_{\alpha-i\infty}^{\alpha+i\infty} e^{\Psi(\tau)} \frac{1 - e^{-\tau x}}{\tau} d\tau \\ &= -\frac{1}{2\pi} \int_{\alpha-i\infty}^{\alpha+i\infty} \tau^{-1} e^{\Psi(\tau)-\tau x} d\tau \end{aligned} \quad (21)$$

In the last step in this argument, one term is zero because $e^{\Psi(\tau)}$ is analytic and decays rapidly as $\Re \tau \rightarrow -\infty$. Similarly, for $m = 1, 2, \dots$ one can show that

$$F^{(m)}(x) = (-1)^{m+1} \frac{1}{2\pi} \int_{\alpha-i\infty}^{\alpha+i\infty} \tau^{-m-1} e^{\Psi(\tau)-\tau x} d\tau \quad (22)$$

provided $\alpha < 0$. It is also useful to consider the functions

$$G^{(m)}(x) := (-1)^{m+1} \frac{1}{2\pi} \int_{\alpha-i\infty}^{\alpha+i\infty} \tau^{-m-1} e^{\Psi(\tau)-\tau x} d\tau \quad (23)$$

defined when $\alpha > 0$. One can show by evaluating the residue at $\tau = 0$ that

$$F^{(0)}(x) = G^{(m)}(x) - 1 \quad (24)$$

$$F^{(1)}(x) = G^{(1)}(x) - E[L] + x \quad (25)$$

with similar formulas relating $F^{(m)}$ and $G^{(m)}$ for $m = 2, 3, \dots$

Since the conditional portfolio loss is a sum of similar, but not identical, independent random variables, we can follow the argument of Daniels to produce an expansion for the functions $F^{(m)}$. Some extra features are involved: the cumulant generating function is not N times something, but rather a sum of N (easily computed) terms; we must deal with the factor τ^{-m-1} ; we must deal with the fact that critical points of the exponent in these integrals may be on the positive or negative real axis and there is a pole at $\tau = 0$. To treat the most general case, we move the factor τ^{-m-1} into the exponent and consider the saddlepoint condition

$$\Psi'(\tau) - (m+1)/\tau - x = 0 \quad (26)$$

Proposition 5.1 from [9] shows that a choice of two real saddlepoints solving this equation is typically available:

Proposition 1 *Suppose that $p_j, l_j > 0$ for all j . Then*

1. *There is a solution τ^* , unique if it exists, of $\Psi'(\tau) - x = 0$ if and only if $0 < x < \sum_j l_j$. If $E[\hat{L}] > x > 0$, then $\tau^* > 0$ and if $E[\hat{L}] < x < \sum_j l_j$, then $\tau^* < 0$.*
2. *For each $m \geq 0$, there is exactly one solution τ_m^- of equation (26) on $(-\infty, 0)$, if $x < \sum_j l_j$ and no solution on $(-\infty, 0)$, if $x \geq \sum_j l_j$. Moreover, when $x < \sum_j l_j$, the sequence $\{\tau_m^-\}_{m \geq 0}$ is monotonically decreasing in m .*
3. *For each $m \geq 0$, there is exactly one solution τ_m^+ of equation (26) on $(0, \infty)$, if $x > 0$ and no solution on $(0, \infty)$, if $x \leq 0$. Moreover, when $x > 0$ the sequence $\{\tau_m^+\}_{m \geq 0}$ is monotonically increasing in m .*

At this point, the methods in [2] and [9] differ. We consider first the method in [9] for computing $F^{(m)}, m = 0, 1$. The argument of Daniels directly is applied, but with the following strategy for choosing the saddlepoint. Whenever $x < E[\hat{L}]$, τ_m^- is chosen as the center of the Taylor expansion for the integral in equation (22). Whenever $x > E[\hat{L}]$, instead, τ_m^+ is chosen as the center of the Taylor expansion for the integral in equation (23), and either of equations (24)

or (25) is used. Thus for example, when $x > E[L]$, the approximation for $m = 1$ is

$$F^{(1)}(x) \sim x - E[L] + \frac{e^{\tau_1^+ x + \Psi(\tau_1^+)}}{\sqrt{2\pi \Psi^{(2)}(\tau_1^+)}} \times \left[1 + \frac{\Psi^{(4)}(\tau_1^+)}{8(\Psi^{(2)}(\tau_1^+))^2} - \frac{5(\Psi^{(3)}(\tau_1^+))^2}{24(\Psi^{(2)}(\tau_1^+))^3} + \dots \right] \quad (27)$$

In [2], the $m = -1$ solution τ^* , suggested by large deviation theory, is chosen as the center of the Taylor expansion, even for $m \neq -1$. The factor τ^{-m-1} is then included with the other nonexponentiated terms, leading to an asymptotic expansion with terms of the form

$$\int_{-\infty}^{\infty} e^{-w^2/2} (w + w_0)^{-m-1} w^k dw w_0 = \tau^* / \sqrt{\Psi^{(2)}(\tau^*)} \quad (28)$$

These integrals can be evaluated in closed form, but are somewhat complicated, and more terms are needed for a given order of accuracy.

Numerical implementation of the saddlepoint method for portfolio credit problems thus boils down to efficient computation of the appropriate solutions of the saddlepoint condition given by equation (26). This is a relatively straightforward application of one-dimensional Newton–Raphson iteration, but must be done for a large number of values of (x, t, y) . For typical parameter values and up to 2^{10} obligors, [9] report that saddlepoints were usually found in under 10 iterations, which suggests that a saddlepoint expansion will run no more than about 10 times

slower than the Edgeworth expansion with the same number of terms. However, both [2] and [9] observe that the accuracy of the saddlepoint expansion is often far greater.

Acknowledgments

Research underlying this article was supported by the Natural Sciences and Engineering Research Council of Canada and MITACS, Canada.

References

- [1] Andersen, L., Sidenius, J. & Basu, S. (2003). All your hedges in one basket, *Risk* **16**, 67–72.
- [2] Antonov, A., Mechkov, S. & Misirpashaev, T. (2005). *Analytical Techniques for Synthetic CDOs and Credit Default Risk Measures*, Numerix Preprint http://www.defaultrisk.com/pp_crdrv_77.htm.
- [3] Daniels, H.E. (1954). Saddlepoint approximations in statistics, *Annals of Mathematical Statistics* **25**, 631–650.
- [4] Gordy, M. (2002). Saddlepoint approximation of credit risk, *Journal of Banking Finance* **26**(2), 1335–1353.
- [5] Hull, J. & White, A. (2004). Valuation of a CDO and an n th to default CDS without Monte Carlo simulation, *Journal of Derivatives* **2**, 8–23.
- [6] Martin, R., Thompson, K. & Browne, C. (2003). Taking to the saddle, in *Credit Risk Modelling: The Cutting-edge Collection*, M. Gordy, ed, Riskbooks, London.
- [7] Varadhan, S.R.S. (1966). Asymptotic probabilities and differential equations, *Communications on Pure and Applied Mathematics* **19**, 261–286.
- [8] Watson, G.N. (1995). *A Treatise on the Theory of Bessel Functions*, 2nd Edition, Cambridge University Press, Cambridge, reprint of the second (1944) edition.
- [9] Yang, J.P., Hurd, T.R. & Zhang, X.P. (2006). Saddlepoint approximation method for pricing CDOs, *Journal of Computational Finance* **10**, 1–20.

THOMAS R. HURD

Credit Scoring

Credit scoring models play a fundamental role in the risk management practice at most banks. Commercial banks' primary business activity is related to extending credit to borrowers and generating loans and credit assets. A significant component of a bank's risk, therefore, lies in the quality of its assets that needs to be in line with the bank's risk appetite.^a To manage risk efficiently, quantifying it with the most appropriate and advanced tools is an extremely important factor in determining the bank's success.

Credit risk models are used to quantify credit risk at counterparty or transaction level and they differ significantly by the nature of the counterparty (e.g., corporate, small business, private individual). Rating models have a long-term view (through the cycle) and have been always associated with corporate clients, financial institutions, and public sector (*see Credit Rating; Counterparty Credit Risk*). Scoring models, instead, focus more on the short term (point in time) and have been mainly applied to private individuals and, more recently, extended to small- and medium-sized enterprises (SMEs).^b In this article, we focus on credit scoring models, giving an overview of their assessment, implementation, and usage.

Since 1960s, larger organizations have been utilizing credit scoring to quickly and accurately assess the risk level of their prospects, applicants, and existing customers mainly in the consumer-lending business. Increasingly, midsize and smaller organizations are appreciating the benefits of credit scoring as well. The credit score is reflected in a number or letter(s) that summarizes the overall risk utilizing available information on the customer. Credit scoring models predict the probability that an applicant or existing borrower will default or become delinquent over a fixed time horizon.^c The credit score empowers users to make quick decisions or even to automate decisions, and this is extremely desirable when banks are dealing with large volumes of clients and relatively small margin of profits at individual transaction level.

Credit scoring models can be classified into three main categories: application, behavioral, and collection models, depending on the stage of the consumer credit cycle in which they are used. The main difference between them lies in the set of variables that are available to estimate the client's creditworthiness, that is, the earlier the stage in the credit cycle,

the lower the number of specific client information available to the bank. This generally means that application models have a lower prediction power than behavioral and collection models.

Over the last 50 years, several statistical methodologies have been used to build credit scoring models. The very simplistic univariate analysis applied at the beginning (late 1950s) was replaced as soon as academic research started to focus on credit scoring modeling techniques (late 1960s). The seminal works, in this field, of Beaver [10] and Altman [1] introduced the multivariate discriminant analysis (MDA) that became the most popular statistical methodology used to estimate credit scoring models until Ohlson [26], for the first time, applied the conditional logit model to the default prediction's study. Since Ohlson's research (early 1980s), several other statistical techniques have been utilized to improve the prediction power of credit scoring models (e.g., linear regression, probit analysis, Bayesian methods, neural network, etc.), but the logistic regression still remains the most popular method.

Lately, credit scoring has gained new importance with the new Basel Capital Accord. The so-called Basel II replaces the current 1988 capital accord and focuses on techniques that allow banks and supervisors to properly evaluate the various risks that banks face (*see Internal-ratings-based Approach; Regulatory Capital*). Since credit scoring contributes broadly to the internal risk assessment process of an institution, regulators have enforced more strict rules about model development, implementation, and validation to be followed by banks that wish to use their internal models in order to estimate capital requirements.

The remainder of the article is structured as follows. In the second section, we review some of the most relevant research related to credit scoring modeling methodologies. In the third section, following the model lifecycle structure, we analyze the main steps related to the model assessment, implementation, and validation process.

The statistical techniques used for credit scoring are based on the idea of discrimination between several groups in a data sample. These procedures originated in the 1930s and 1940s of the previous century [18]. At that time, some of the finance houses and mail order firms were having difficulties with their credit management. Decision whether to give loans or send merchandise to the applicants was

made judgmentally by credit analysts. The decision procedure was nonuniform, subjective, and opaque; it depended on the rules of each financial house and on the personal and empirical knowledge of each single clerk. With the rising number of people applying for a credit card, it was impossible to rely only on credit analysts; an automated system was necessary. The first consultancy was formed in San Francisco by Bill Fair and Earl Isaac in the late 1950s.

After the first empirical solutions, academic interest on the topic rose and, given the lack of consumer-lending figures, researchers focused their attention on small business clients. The seminal works in this field were Beaver [10] and Altman [1], who developed univariate and multivariate models, applying an MDA technique to predict business failures using a set of financial ratios.^d

For many years thereafter, MDA was the prevalent statistical technique applied to the default prediction models and it was used by many authors [2, 3, 13, 15, 16, 24, 29]. However, in most of these studies, authors pointed out that two basic assumptions of MDA are often violated when applied to the default prediction problems.^e Moreover, in MDA models, the standardized coefficients cannot be interpreted such as the slopes of a regression equation and, hence, do not indicate the relative importance of the different variables. Considering these MDA's problems, Ohlson [26], for the first time, applied the conditional logit model to the default prediction's study.^f The practical benefits of the logit methodology are that it does not require the restrictive assumptions of MDA and allows working with disproportional samples. The performance of his models, in terms of classification accuracy, was lower than the one reported in the previous studies based on MDA, but he pointed out some reasons to prefer the logistic analysis.

From a statistical point of view, logit regression seems to fit well the characteristics of the default prediction problem, where the dependent variable is binary (default/nondefault) and with the groups being discrete, nonoverlapping, and identifiable. The logit model yields a score between 0 and 1, which conveniently can be transformed in the probability of default (PD) of the client. Lastly, the estimated coefficients can be interpreted separately as the importance or significance of each of the independent variables in the explanation of the estimated PD. After the work of Ohlson [26], most of the academic

literature [5, 11, 19, 27, 30] used logit models to predict default.

Several other statistical techniques have been tested to improve the prediction accuracy of credit scoring models (e.g., linear regression, probit analysis, Bayesian methods, neural network, etc.), but the empirical results have never shown really significant benefits.

Credit Scoring Models Lifecycle

As already mentioned, banks that want to implement the most advanced approach to calculate their minimum capital requirements (i.e., advanced internal rating based approach, A-IRB) are subject to more strict and common rules regarding how their internal models should be developed, implemented, and validated.^g A standard model lifecycle has been designed to be followed by the financial institutions that will want to implement the A-IRB approach. The lifecycle of every model is divided into several phases (assessment, implementation, validation) and regulators have published specific requirements for each one of them. In this section, we describe the key aspects of each model's lifecycle phase.

Model Assessment

Credit scoring models are used to risk rank new or existing clients on the basis of the assumption that the future will be similar to the past. If an applicant or an existing client had a certain behavior in the past (e.g., paid back his debt or not), it is likely that a new applicant or client, with similar characteristics, will show the same behavior. As such, to develop a credit scoring model, we need a sample of past applicants or clients' data related to the same product as the one we want to use our scoring model for. If historical data from the bank are available, an empirical model can be developed. When banks do not have data or do not have a sufficient amount of data to develop an empirical model, an expert or a generic model is the most popular solution.^h

When a data sample covering the time horizon necessary for the statistical analysis (usually at least 1 year) is available, the performance of the clients inside the sample can be observed. We define performance as the default or nondefault event associated with each client.ⁱ This binary variable is the dependent variable used to run the regression analysis. The

characteristics of the client at the beginning of the selected period are the predictors.

Following the literature discussed in the second section, a conditional probability model, logit model, is commonly used by most banks to estimate the 1-year score through a range of variables by maximizing the log-likelihood function. This procedure is used to obtain the estimates of the parameters of the following logit model [20, 21]:

$$P_1(X_i) = \frac{1}{[1 + e^{-(B_0 + B_1 X_{i1} + B_2 X_{i2} + \dots + B_n X_{in})}]} \\ = \frac{1}{[1 + e^{-(D_i)}]} \quad (1)$$

where $P_1(X_i)$ is the score given the vector of attributes X_i ; B_j is the coefficient of attribute j (with $j = 1, \dots, n$); B_0 is the intercept; X_{ij} is the value of the attribute j (with $j = 1, \dots, n$) for customer i ; and D_i is the logit for customer i .

The logistic function implies that the logit score P_1 has a value in $[0,1]$ interval and is increasing in D_i . If D_i approaches minus infinity, P_1 will be zero and if D_i approaches plus infinity, P_1 will be one.

The set of attributes that are used in the regression depends on the type of model that is going to be developed. Application models, employed to decide whether to accept or reject an applicant, typically rely only on personal information about the applicant, given the fact that this is usually the only information available to the bank at that stage.^j Behavioral and collection models include variables describing the status of the relationship between the client and the bank that may add significant prediction power to the model.^k

Once the model is developed, it needs to be tested on a test sample to confirm the soundness of its results. When enough data are available, part of the development sample (hold-out sample) is usually kept for the final test of the model. However, an optimal test of the model would require investigating its performance also on an out-of-time and out-of-universe sample.

Model Implementation

The main advantage of scoring models is to allow banks to implement automated decision systems to manage their retail clients (private individuals and

SMEs). When a large amount of applicants or clients is manually referred to credit analysts to check their information and apply policy rules, most of the benefits associated with the use of scoring models are lost. On the other hand, any scoring model has a “gray” area where it is not able to separate with an acceptable level of confidence between expected “good” clients and expected “bad” ones.^l The main challenge for credit risk managers is to define the most appropriate and efficient thresholds (cutoff) for each scoring model.

In order to maximize the benefits of a scoring model, the optimal cutoff should be set taking into account the misclassification costs related to the type I and type II error rates as Altman *et al.* [2], Taffler [29], and Koh [23] point out. Moreover, we believe that the optimum cutoff value cannot be found without a careful consideration of each particular bank peculiarities (e.g., tolerance for risk, profit-loss objectives, recovery process costs and efficiency, possible marketing strategies). Today, the most advanced banks set cutoffs using profitability analyses at account level.

The availability of sophisticated IT systems has significantly broadened the number of strategies that can be implemented using credit scoring models. The most efficient banks are able to follow the lifecycle of any client, from the application to the end of the relationship, with monthly updated scores calculated by different scorecards related to the phase of the credit cycle where the client is located (e.g., origination, account maintenance, collection, write off). Marketing campaigns (e.g., cross-selling, up-selling), automated limit changes, early collection strategies, and shadow limit management are some of the activities that are fully driven by the output of scoring models in most banks.

Model Validation

Banks that have adopted or are willing to adopt the Basel II IRB-advanced approach are required to put in place a regular cycle of model validation that should include at least monitoring of the model performance and stability, reviewing of the model relationships, and testing of model outputs against outcomes (i.e., backtesting).^m

Considering the relatively short lifecycle of credit scoring models due to the high volatility of retail markets, their validation has always been completed

by banks. Basel II has only given to it a more official shape, prescribing that the validation should be undertaken by a team independent from the one that has developed the models.

Stability and performance (i.e., prediction accuracy) are extremely important information about the quality of the scoring models. As such, they should be tracked and analyzed at least monthly by banks, regardless of the validation exercise. As we have discussed above, often scoring models are used to generate a considerable amount of automated decisions that may have a significant impact on the banking business. Even small changes in the population's characteristics can substantially affect the quality of the models, creating undesired selection bias.

In the literature, we have found several indexes that have been used to assess the performance of the models. The simple type I and type II error rates that quantify the accuracy of each model in correctly classifying defaulted and nondefaulted observations have been the first measures to be applied to scoring models. More recently, the accuracy ratio (AR) and the Gini index have become the most popular measures (see [17] for further details).

Backtesting and benchmarking are an essential part of the scoring models' validation. With the backtesting, we evaluate the calibration and discrimination of a scoring model. Calibration refers to the mapping of a score to a quantitative risk measure (e.g., PD). A scoring model is considered well calibrated if the (*ex ante*) estimated risk measures (PD) deviate only marginally from what has been observed *ex post* (actual default rate per score band). Discrimination measures how well the scoring model provides an ordinal ranking of the risk profile of the observations in the sample; for example, in the credit risk context, discrimination measures to what extent defaulters were assigned low scores and nondefaulters high scores.

Benchmarking is another quantitative validation method that aims at assessing the consistency of the estimated scoring models with those obtained using other estimation techniques and potentially using other data sources. This analysis may be quite difficult to perform for retail portfolios, given the lack of generic benchmarks in the market.¹¹

Lastly, we would like to point out that Basel II specifically requires senior management to be fully involved and aware of the quality and performance

of all the scoring models utilized in the daily business (see [9], par. 438, 439, 660, 718 (LXXVI), 728).

End Notes

^aRisk appetite is defined as the maximum risk the bank is willing to accept in executing its chosen business strategy, to protect itself against events that may have an adverse impact on its profitability, the capital base, or share price (see **Economic Capital Allocation; Economic Capital**).

^bRecently, several studies [4, 12] have shown the importance for banks of classifying SMEs as retail clients and applying credit scoring models developed specifically for them.

^cThe default definition may be significantly different by bank and type of client. The new Basel Capital Accord [9] (par.452) has given a common definition of default (i.e., 90 days past due over 1-year horizon) that is consistently used by most banks today.

^dThe original Z-score model Altman [1] used five ratios: working capital/total assets, retained earnings/total assets, EBIT/total assets, market value equity/BV of total debt, and sales/total assets.

^eMDA is based on two restrictive assumptions: (i) the independent variables included in the model are multivariate normally distributed and (ii) the group dispersion matrices (or variance-covariance matrices) are equal across the failing and the nonfailing group. See [6, 22, 25] for further discussions about this topic.

^fZmijewski [31] was the pioneer in applying probit analysis to predict default, but, until now, logit analysis has given better results in this field.

^gThe new Basel Capital Accord offers financial institutions the possibility to choose between the standardized and the advanced approach to calculate their capital requirements. Only the latter requires banks to use their own internal risk assessment tools to quantify the inputs of the capital requirements formulas (i.e., PD and loss given default).

^hExpert scorecards are based on subjective weights assigned by an analyst, whereas generic scorecards are developed on pooled data from other banks operating in the same market. For a more detailed analysis of the possible solutions that banks can consider when not enough historical data is available, see [28].

ⁱSee end note (b).

^jThe most common application variables used are sociodemographic information about the applicants (e.g., marital status, residence type, time at current address, type of work, time at current work, flag phone, number of children, installment on income, etc.). When a credit bureau is available in the market, the information that can be obtained related to the behavior of the applicant with other financial institutions is an extremely powerful variable to be used in application models.

^kVariables used in behavioral and collection scoring models are calculated and updated at least monthly. As such, the

correlation between these variables and the default event is significantly high. Examples of behavioral variables are as follows: the number of missed installments (current, max last 3/6/12 months, or ever), number of days in excess (current, max last 3/6/12 months, or ever), outstanding on limit, and so on. Behavioral score can be calculated at facility and customer level (when several facilities are related to the same client).

^lDepending on the chosen binary-dependent variable, “good” and “bad” will have different meanings. For credit risk models, these terms are usually associated with nondefaulted and defaulted clients, respectively.

^mSee par. 417 and 718 (XCix) of the new Basel Capital Accord [7–9] (see also **Model Validation; Backtesting**).

ⁿRecently, rating agencies (e.g., Standard & Poor’s and Moody’s) and credit bureau providers (e.g., Fair Isaac and Experian) have started to offer services of benchmarking for retail scoring models. For more details about backtesting and benchmarking techniques, see [14].

References

- [1] Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance* **23**(4), 589–611.
- [2] Altman, E.I., Haldeman, R.G. & Narayanan, P. (1977). Zeta-analysis. A new model to identify bankruptcy risk of corporations, *Journal of Banking and Finance* **1**, 29–54.
- [3] Altman, E.I., Hartzell, J. & Peck, M. (1995). *A Scoring System for Emerging Market Corporate Debt*. Salomon Brothers Emerging Markets Bond Research, May 15.
- [4] Altman, E.I. & Sabato, G. (2005). Effects of the new Basel capital accord on bank capital requirements for SMEs, *Journal of Financial Services Research* **28**(1/3), 15–42.
- [5] Aziz, A., Emanuel, D.C. & Lawson, G.H. (1988). Bankruptcy prediction – an investigation of cash flow based models, *Journal of Management Studies* **25**(5), 419–437.
- [6] Barnes, P. (1982). Methodological implications of non-normality distributed financial ratios, *Journal of Business Finance and Accounting* **9**(1), 51–62.
- [7] Basel Committee on Banking Supervision (2005). *Studies on the Validation of Internal Rating Systems*. Working paper 14, www.bis.org.
- [8] Basel Committee on Banking Supervision (2005). *Update on Work of the Accord Implementation Group Related to Validation Under the Basel II Framework*. Newsletter 4, www.bis.org.
- [9] Basel Committee on Banking Supervision (2006). *International Convergence of Capital Measurement and Capital Standards*. www.bis.org.
- [10] Beaver, W. (1967). Financial ratios predictors of failure, *Journal of Accounting Research* **4**, 71–111.
- [11] Becchetti, L. & Sierra, J. (2003). Bankruptcy risk and productive efficiency in manufacturing firms, *Journal of Banking and Finance* **27**(11), 2099–2120.
- [12] Berger, A.N. & Udell, S.W. (2007). Small business credit scoring and credit availability, *Journal of Small Business Management* **45**(1), 5–22.
- [13] Blum, M. (1974). Failing company discriminant analysis, *Journal of Accounting Research* **12**(1), 1–25.
- [14] Castermans, G., Martens, D., Van Gestel, T., Hamers, B. & Baesens, B. (2007). An overview and framework for PD backtesting and benchmarking, *Proceedings of Credit Scoring and Credit Control X*, Edinburgh, Scotland.
- [15] Deakin, E. (1972). A discriminant analysis of predictors of business failure, *Journal of Accounting Research* **10**(1), 167–179.
- [16] Edmister, R. (1972). An empirical test of financial ratio analysis for small business failure prediction, *Journal of Financial and Quantitative Analysis* **7**(2), 1477–1493.
- [17] Engelmann, B., Hayden, E. & Tasche, D. (2003). Testing rating accuracy, *Risk* **16**(1), 82–86.
- [18] Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**, 179–188.
- [19] Gentry, J.A., Newbold, P. & Whitford, D.T. (1985). Classifying bankrupt firms with funds flow components, *Journal of Accounting Research* **23**(1), 146–160.
- [20] Gujarati, N.D. (2003). *Basic Econometrics*, 4th Edition, McGraw-Hill, London.
- [21] Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd Edition, John Wiley & Sons, New York.
- [22] Karels, G.V. & Prakash, A.J. (1987). Multivariate normality and forecasting of business bankruptcy, *Journal of Business Finance & Accounting* **14**(4), 573–593.
- [23] Koh, H.C. (1992). The sensitivity of optimal cutoff points to misclassification costs of Type I and Type II errors in the going-concern prediction context, *Journal of Business Finance & Accounting* **19**(2), 187–197.
- [24] Lussier, R.N. (1995). A non-financial business success versus failure prediction model for young firms, *Journal of Small Business Management* **33**(1), 8–20.
- [25] Mc Leay, S. & Omar, A. (2000). The sensitivity of prediction models to the non-normality of bounded and unbounded financial ratios, *British Accounting Review* **32**, 213–230.
- [26] Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy, *Journal of Accounting Research* **18**(1), 109–131.
- [27] Platt, H.D. & Platt, M.B. (1990). Development of a class of stable predictive variables: the case of bankruptcy prediction, *Journal of Business Finance & Accounting* **17**(1), 31–51.
- [28] Sabato, G. (2008). Managing credit risk for retail low-default portfolios, in *Credit Risk: Models, Derivatives and Management*, N. Wagner, ed., Financial Mathematics Series, Chapman & Hall/CRC.
- [29] Taffler, R.J. & Tisshaw, H. (1977). Going, going, gone – four factors which predict, *Accountancy* **88**(1083), 50–54.

- [30] Zavgren, C. (1983). The prediction of corporate failure: the state of the art, *Journal of Accounting Literature* **2**, 1–37.
- [31] Zmijewski, M.E. (1984). Methodological issues related to the estimation of financial distress prediction models, *Journal of Accounting Research* **22**, 59–86.

Further Reading

Taffler, R.J. (1982). Forecasting company failure in the UK using discriminant analysis and financial ratio data, *Journal of the Royal Statistical Society* **145**(3), 342–358.

Related Articles

Backtesting; Credit Rating; Credit Risk; Internal-ratings-based Approach; Model Validation.

GABRIELE SABATO

Credit Rating

The cornerstone of credit risk measurement and management for a financial institution is the credit rating, whether supplied by an external credit rating agency (CRA) or generated by an internal credit model. A credit rating represents an overall assessment of the creditworthiness of a borrower, obligor, or counterparty, and is thus meant to reflect only credit or default risk. That obligor may be either a firm or an individual. In this way the rating is a forecast, and like all forecasts it is noisy. For that reason, credit rating agencies make use of discrete ratings. The convention used by the three largest rating agencies, namely, Fitch, Moody's, and Standard & Poor's (S&P) is to have seven credit grades. They are, from best to worst, AAA, AA, A, BBB, BB, B, and CCC using S&P and Fitch's nomenclature, and Aaa, Aa, A, Baa, Ba, B, and Caa using Moody's nomenclature.^b As a firm migrates from a higher to a lower credit rating, that is, it is downgraded, it simply moves closer to default.

The market for credit ratings in the United States is dominated by two players: S&P and Moody's Investor Services; of the smaller rating agencies, only Fitch plays a significant role in the United States (although it has a more substantial presence elsewhere).^c The combined market share of Moody's and S&P is 80%, and once market share of Fitch is added, the total exceeds 95% [20].

To be sure, it is not the obligor but the instrument issued by the obligor that receives a credit rating, though an obligor rating typically corresponds to the credit risk of a senior unsecured debenture issued by that firm. The distinction is not that relevant for corporate bonds, where the obligor rating is commensurate with the rating on a senior unsecured instrument, but is quite relevant for structured credit products such as asset-backed securities (ABS). Nonetheless, as stated in a recent S&P document, "[o]ur ratings represent a uniform measure of credit quality globally and across all types of debt instruments. In other words, an 'AAA' rated corporate bond should exhibit the same degree of credit quality as an 'AAA' rated securitized issue." (44, p.4).

This stated intent implies that an investor can assume that, say, a double-A rated instrument is the same in the United States as in Belgium or Singapore,

regardless of whether that instrument is a standard corporate bond or a structured product such as a tranche on a collateralized debt obligation (CDO) (see **Collateralized Debt Obligations (CDO)**); see also [31]. The actual behavior of rated obligors or instruments may turn out to have more heterogeneity across countries, industries, and product types, and there is substantial supporting evidence. See [37] for evidence of variation across countries of domicile and industries for corporate bond ratings, and [17] for differences between corporate bonds and structured products.

The rating agencies differ about what exactly is assessed. Although Fitch and S&P evaluate an obligor's overall capacity to meet its financial obligation, and hence is best thought of as an estimate of probability of default, Moody's assessment incorporates some judgment of recovery in the event of loss. In the argot of credit risk management, S&P measures PD (probability of default), whereas Moody's measure is somewhat closer to EL (expected loss) [9].^d These differences seem to remain for structured products. In describing their ratings criteria and methodology for structured products, S&P states the following: "[w]e base our ratings framework on the likelihood of default rather than expected loss or loss given default. In other words, our ratings at the rated instrument level don't incorporate any analysis or opinion on post-default recovery prospects." (44, p. 3). Since 2005, Fitch, followed soon after by S&P and Moody's, have started publishing recovery ratings (each on a six-point scale). Market sector coverage has been different, but expanding, across the agencies. Also, application is different between corporate *versus* structured products.^e

Credit ratings issued by the agencies typically represent an unconditional view, sometimes also called *cycle-neutral* or *through the cycle*: the rating agency's own description of their rating methodology broadly supports this view.

(33, p.6,7): .. [O]ne of Moody's goals is to achieve stable *expected* [italics in original] default rates across rating categories and time. ... Moody's believes that giving only a modest weight to cyclical conditions best serves the interests of the bulk of investors.^f

(43, p.41): Standard & Poor's credit ratings are meant to be forward looking; ... Accordingly, the anticipated ups and downs of business cycles—whether industry-specific or related to the

general economy—should be factored into the credit rating all along. . . . The ideal is to rate ‘through the cycle’.

This unconditional or firm-specific view of credit risk stands in contrast to risk measures such as EDFs (expected default frequencies) from Moody’s KMV. An EDF has two principal inputs: firm leverage and asset volatility, where the latter is derived from equity (stock price) volatility; see [24] for a description. As a result, EDFs can change frequently and significantly since they reflect the stock market’s view of risk for that firm at a given point in time, a view which incorporates both systematic and idiosyncratic risk.

Unfortunately, there is substantial evidence that credit rating changes, including changes to default, exhibit procyclical or systematic variation [7, 27, 37], especially for speculative grades [23].

Although this article’s focus is on credit ratings for corporate entities, including special purpose entities such as structured credit, individuals also receive credit ratings or scores. These are important for obtaining either unsecured credit like a credit card, or even a mobile phone, as well as secured credit such as an auto loan or lease or a mortgage. They have received considerable attention of late in the context of subprime mortgages and their securitization. Nonetheless, because the individual credit exposures are typically small, even considering mortgages, banks have tried to automate the credit assessment of retail exposures as much as possible, often with the help of outside firms called *credit bureaus*. See [22] for a discussion with application to credit cards, and [1] for a broader survey on retail credit (see also **Credit Scoring** for a review of credit scoring models).

How to Generate a Credit Rating

One of the earliest studies of predicting firm bankruptcy, perhaps the most obvious form of borrower default, is [2]. Altman constructed a balanced sample of 33 defaulted and 33 nondefaulted firms to build a bankruptcy prediction model using multiple-discriminant analysis. His choice of condition variables—ratios reflecting firm leverage and profitability—has influenced default models to this day, and therefore it is worth showing the final

model here:

$$Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5 \quad (1)$$

where

X_1 = working capital/total assets,

X_2 = retained earnings/total assets,

X_3 = earnings before interest and taxes/total assets,

X_4 = market value of equity/book value of total debt, and

X_5 = sales/total assets.

A large value of Z indicates high credit quality; the firm is far from its default threshold. The simplicity of this model makes it deceptively easy to criticize: have the coefficients remained the same? is it applicable to all industries (financial firms are typically much more highly leveraged— X_4 is small—than nonfinancial firms) or all countries? is the relationship really linear in the conditioning variables? and why are nonfinancial variables such as firm age or some measure of management quality not considered? However, the Altman Z -score endures to this day; it can be found for most publicly traded firms on any Bloomberg terminal.

The next important innovation in credit modeling is arguably Merton’s [32] option-based default model (see **Default Barrier Models**). Merton recognized that a lender is effectively writing a put option on the assets of the borrowing firm; owners and owner-managers (i.e., shareholders) hold the call option. Thus a firm is expected to default when the value of its assets falls below a threshold value determined by its liabilities. To this day all credit models owe an intellectual debt to Merton’s insights. The best-known commercial application of the Merton model is the Moody’s KMV EDF (expected default frequency) model.

Clearly, quantitative information obtained from (public) accounting data, metrics such as leverage, profitability, debt service coverage, and liquidity, is important for arriving at a credit assessment of a firm. In addition, a rating agency, because it has access to private information about the firm, can and does include qualitative information such as the quality of management [3, 28]. Indeed this is partly what makes credit rating agencies unique and important: they aggregate public and private

information into a single measure of creditworthiness (riskiness) and then make that summary statistic—the credit rating—public, essentially providing a public good [47]. By contrast a Moody's KMV EDF makes use only of public information, although it transforms this information using a proprietary methodology.⁸ In fact, rating agencies are in the business of not just information production but, in the words of Boot *et al.* [12], they also act as “information equalizers” [quotes in the original]. In this way, they serve as a coordinating mechanism or focal point to the financial markets.

Model Performance

All credit scoring or rating models map a set of financial and nonfinancial variables into the unit interval: the objective is to generate a probability of default, to separate the defaulters from the nondefaulters. Unsurprisingly, there is a plethora of modeling choices as documented, for instance, by Resti and Sironi [41]. However, in the horse race of default prediction models, the hazard approach as shown in [16, 42] seems to be emerging as the winner. See [11] for a recent overview. While we can say little about the performance on bank internal credit scoring models—they are proprietary—we can examine the empirical default experience of firms with a rating from a credit rating agency.

Highly rated firms default quite rarely. For example, Moody's reports that the 1-year investment grade default rate over the period 1983–2007 was 0.069% or 6.9 bp [35]. This is an average over four letter grade ratings: Aaa through Baa. Thus in a pool of 10 000 investment grade obligors or instruments we would expect seven defaults over the course of 1 year. But what if only four default? What about 11? Higher than expected default could be the result of either bad luck or a bad model, and it is very hard to distinguish between the two, especially for small probabilities (see also [29] and **Backtesting**). Indeed the use of the regulatory color scheme—green, amber, red—which is behind the 1996 Market Risk Amendment to the Basel I, was motivated precisely by this recognition, and in that case the probability to be validated is comparatively large 1% (for 99% VaR) [8] with daily data.

Although rating agencies insist that their ratings scale reflects an ordinal ranking of credit risk,

they also publish default rates for different horizons by rating. Thus we would expect default rates or probabilities to be monotonically increasing as one descends the credit spectrum. Using S&P rating histories, Hanson and Schuermann [23] show formally that monotonicity is violated frequently for most notch-level investment grade 1-year estimated default probabilities. The precision of the PD point estimates is quite low; there have been no defaults over 1 year for triple-A or AA+ (Aa1) rated firms, yet surely we do not believe that the 1-year probability of default is identically equal to zero. The new Basel Capital Accord (*see* **Regulatory Capital**), perhaps with this in mind, has set a lower bound of 3 bp for any PD estimate (10, §285), commensurate with about a single-A rating. Trück and Rachev [46] show the economic impact resulting from such uncertainty using bank internal ratings and a corresponding loan portfolio. Pluto and Tasche [40] propose a conservative approach to generating PD estimates for low-default portfolios.

Despite this lack of statistical precision, Kliger and Sarig [26] show that bond ratings contain price-relevant information by taking advantage of a natural experiment. On April 26, 1982, Moody's introduced overnight modifiers to their rating system, much like the notching used by S&P and Fitch, effectively introducing finer credit rating information about their issuer base without any change in the firm fundamentals. They find that bond prices indeed adjust to the new information, as do stock prices, and that any gains enjoyed by bondholders are offset by losses suffered by stockholders.

Although the 1-year horizon is typical in credit analysis (and is also the horizon used in Basel II), most traded credit instruments have longer maturity. For example, the typical CDS contract (*see* **Credit Default Swaps**) is five years, and over that horizon there are positive empirical default rates for Aaa and Aa, which Moody's reports to be 7.8 bp and 18.3 bp, respectively [35].

The preceding discussion highlights the difficulty of accurately forecasting such small PDs. Empirical estimates of PDs using credit rating histories can be quite noisy, even with over 25 years of data. Under the new Basel Capital Accord (Basel II), US regulators would require banks to have a minimum of seven nondefault rating categories [21].

Internal Ratings

With the roll-out of the New Basel Capital Accord, internal credit ratings will become widespread. To qualify for the internal-ratings-based (IRB) approach (*see Internal-ratings-based Approach*), which allows a bank to use its own internal credit rating, the accord provides the following rather nonspecific guidance (10, §461):

Banks must use information and techniques that take appropriate account of the long-run experience when estimating the average PD for each rating grade. For example, banks may use one or more of the three specific techniques set out below: internal default experience, mapping to external data, and statistical default models.

Since bank internal ratings are proprietary, not much is known (publicly) about their exact construction or about their performance. Carey and Hrycay's [15] study of internal ratings in US banks suggests that such rating systems rely at least to some degree on external ratings themselves, either by using them directly, when available, or calibrating internal credit scores to external ratings. Several books by practitioners and academics with practitioner experience, for example, [18, 30, 38, 41], indicate that the methods used are, perhaps unsurprisingly, much along the lines of the models covered above: statistical approaches—discriminant models in the manner of Altman's Z-score, logistic regression, neural networks, decision trees, and so on—that make use of firm financials, augmented with a judgmental overlay that incorporates qualitative information. This more intangible information is especially important when lending to small and young firms with no footprint in the capital markets, as shown, for instance, by Peterson and Rajan [39]. Ashcraft [5] documents that the failure of even healthy banks is followed by largely permanent declines in real activity, which is attributed to the destruction of private information about informationally opaque borrowers. There are now some papers emerging that attempt to formalize the incorporation of qualitative and subjective information, for example, from a loan officer, into bank internal credit scores or ratings. See, for instance, [25, 45].

Ratings for Structured Credit Products

Corporate bond (obligor) ratings are largely based on firm-specific risk characteristics. Since ABS

structures represent claims on cash flows from a *portfolio* of underlying assets, the rating of a structured credit product must take into account systematic risk. It is correlated losses that matter especially for the more senior (higher rated) tranches, and loss correlation arises through dependence on shared or common (or systematic) risk factors.^h For ABS deals that have a large number of underlying assets, for instance, mortgage-backed securities (MBS), the portfolio is large enough such that all idiosyncratic risk is diversified away leaving only systematic exposure to the risk factors particular to that product class (here, mortgages). By contrast, a substantial amount of idiosyncratic risk may remain in ABS transactions with smaller asset pools, for instance, CDOs [4, 17].

Because these deals are portfolios, the effect of correlation is not the same for all tranches (*see CDO Tranches: Impact on Economic Capital*): equity tranche holders prefer higher correlation, whereas senior tranches prefer lower correlation (tail losses are driven by loss correlation). As correlation increases, so does portfolio loss volatility. The payoff function for the equity tranche is, true to its name, like a call option. Indeed equity itself is a call option on the assets of the underlying firm, and the value of a call option is increasing in volatility. If the equity tranche is long a call option, then the senior tranche is short a call option, so that their payoffs behave in an opposite manner. The impact of increased correlation on the value of mezzanine tranches is ambiguous and depends on the structure of a particular deal [19]. By contrast, correlation with systematic risk factors should not matter for corporate ratings (*see also Base Correlation and Modeling Correlation of Structured Instruments in a Portfolio Setting* for details on default correlation modeling).

As a result of the portfolio nature of the rated products, the ratings migration behavior may also be different than for ordinary obligor ratings. Moody's Investors Services [34] reports that rating changes are much more common for corporate bond than for structured product ratings, but the magnitude of changes (number of notches up- or downgraded) was nearly double for the structured products.ⁱ There are potentially two reasons for this difference: model error or greater sensitivity of performance to systemic factors.

The modeling approach for rating structured credit products is in flux as of this writing, driven by the poor performance during the turmoil in the credit

markets in 2007 and 2008. Moody's, for instance, has recently proposed adding two new risk measures for structured finance transactions [36]. First, an "assumption volatility score" (or "V Score") that would rate the uncertainty of a rating and the potential for future ratings volatility on a scale of 1–5 (low to high). Second, a "loss sensitivity" that would estimate the number of notches a tranche would be downgraded if the expected loss of the collateral pool were increased to the 95th percentile of the original loss distribution. Moody's decided to develop these risk measures in addition to rather than in substitution for the standard credit ratings, which are on the same scale as their corporate ratings, precisely to allow investors to make a baseline comparison to other rated securities.

End Notes

^aAny views expressed represent those of the authors only and not necessarily those of the Federal Reserve Bank of New York or the Federal Reserve System.

^bFor no reason other than convenience and expediency, we make use of the Fitch and S&P nomenclature for the remainder of the article.

^cAs of this writing, there are ten "accredited" rating agencies in the United States (see [6] for a discussion on what it means to be "accredited"): A.M. Best, Dominion Bond Rating Service (DBRS), Egan-Jones Rating Company, Fitch Inc., Japan Credit Rating Agency Ltd., Moody's Investors Services, Inc., LACE Financial Corp., Rating and Investment Information, Inc, Realpoint LLC, Standard & Poor's Ratings Services. There are several other firms that provide credit opinions; one such example is Eagan-Jones Rating. An extensive list of rating agencies across the globe can be found at http://www.defaultrisk.com/rating_agencies.htm. For an exposition of the history of the credit rating industry, see [47]. For a detailed institutional discussion, see [14].

^dSpecifically, $EL = PD \times LGD$, where LGD is loss given default. However, given the paucity of LGD data, little variation in EL that exists at the obligor (as opposed to instrument) level can be attributed to variation in LGD making the distinction between the agencies modest at best.

^eSee http://www.fitchratings.com/corporate/fitchResources.cfm?detail=1&rd_file=intro#rtng_actn.

^fThis view was recently reinforced in [13]; both authors work for Moody's.

^gIn particular, the mapping from distance-to-default to EDF is proprietary.

^hNote that correlation includes more than just economic conditions, as it includes (i) model risk by the agencies, (ii) originator and arranger effects, and (iii) servicer effects.

ⁱThe recent rash of downgrades for structured credit products in the wake of the subprime credit crisis may change this stylized fact.

References

- [1] Allen, L., DeLong, G. & Saunders, A. (2004). Issues in the credit risk modeling of retail markets, *Journal of Banking and Finance* **28**, 727–752.
- [2] Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance* **20**, 589–609.
- [3] Altman, E.I. & Rijken, H.A. (2004). How rating agencies achieve rating stability, *Journal of Banking and Finance* **28**, 2679–2714.
- [4] Amato, J.D. & Remolona, E.M. (2005). *The Pricing of Unexpected Credit Losses*. BIS Working Paper No. 190.
- [5] Ashcraft, A. (2005). Are banks really special? New evidence from the FDIC-induced failure of healthy banks, *American Economic Review* **95**, 1712–1730.
- [6] Ashcraft, A. & Schuermann, T. (2008). *Understanding the Securitization of Subprime Mortgage Credit. Foundations and Trends in Finance* **2** 191–309.
- [7] Bangia, A., Diebold, F.X., Kronimus, A., Schagen, C. & Schuermann, T. (2002). Ratings migration and the business cycle, with applications to credit portfolio stress testing, *Journal of Banking and Finance* **26**(2/3), 445–474.
- [8] Basel Committee on Banking Supervision (1996). *Amendment to the Capital Accord to Incorporate Market Risks*. Basel Committee Publication No. 24. Available: www.bis.org/publ/bcbs24.pdf.
- [9] Basel Committee on Banking Supervision (2000). *Credit Ratings and Complementary Sources of Credit Quality Information*. BCBS Working Paper No. 3, available at <http://www.bis.org/publ/bcbs.wp3.htm>.
- [10] Basel Committee on Banking Supervision (2005). *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Available at <http://www.bis.org/publ/bcbs118.htm>.
- [11] Bharath, S.T. & Shumway, T. (2008). Forecasting default with the Merton distance to default model, *Review of Financial Studies* **21**, 1339–1369.
- [12] Boot, A.W.A., Milbourn, T.T. & Schmeits, A. (2006). Credit ratings as coordination mechanisms, *Review of Financial Studies* **19**, 81–118.
- [13] Cantor, R. & Mann, C. (2007). Analyzing the tradeoff between ratings accuracy and stability, *Journal of Fixed Income* **16**(4), 60–68.
- [14] Cantor, R. & Packer, F. (1995). The credit rating industry, *Journal of Fixed Income* **5**(3), 10–34.
- [15] Carey, M. & Hrycay, M. (2001). Parameterizing credit risk models with rating data, *Journal of Banking and Finance* **25**, 197–270.
- [16] Chava, S. & Jarrow, R.A. (2004). Bankruptcy prediction with industry effects, *Review of Finance* **8**, 537–569.

- [17] Committee on the Global Financial System (2005). *The Role of Ratings in Structured Finance: Issues and Implications*. Available at <http://www.bis.org/publ/cgfs23.htm>, January.
- [18] Crouhy, M., Galai, D. & Mark, R. (2000). *Risk Management*, McGraw-Hill, New York.
- [19] Duffie, Darrell. (2007). *Innovations in Credit Risk Transfer: Implications for Financial Stability*. Stanford University GSB Working Paper, available at <http://www.stanford.edu/~duffie/BIS.pdf>.
- [20] The Economist (2007). *Measuring the Measurers*. May 31.
- [21] Federal Reserve Board (2003). *Supervisory Guidance on Internal Ratings-Based Systems for Corporate Credit*. Attachment 2 in <http://www.federalreserve.gov/boarddocs/meetings/2003/20030711/attachment.pdf>.
- [22] Gross, D. & Souleles, N. (2002). An empirical analysis of personal bankruptcy and delinquency, *Review of Financial Studies* **15**, 319–347.
- [23] Hanson, S.G. & Schuermann, T. (2006). Confidence intervals for probabilities of default, *Journal of Banking and Finance* **30**(8), 2281–2301.
- [24] Kealhofer, S. & Kurbat, M. (2002). Predictive Merton models, *Risk* February, 67–71.
- [25] Kiefer, N. (2007). The probability approach to default estimation, *Risk* July, 146–150.
- [26] Kliger, D. & Sarig, O. (2000). The information value of bond ratings, *Journal of Finance* **55**(6), 2879–2902.
- [27] Lando, D. & Skødeberg, T. (2002). Analyzing ratings transitions and rating drift with continuous observations, *Journal of Banking and Finance* **26**(2/3), 423–444.
- [28] Löffler, G. (2004). Ratings versus market-based measures of default risk in portfolio governance, *Journal of Banking and Finance* **28**, 2715–2746.
- [29] Lopez, J.A. & Saidenberg, M. (2000). Evaluating credit risk models, *Journal of Banking and Finance* **24**, 151–165.
- [30] Marrison, C. (2002). *Fundamentals of Risk Measurement*, McGraw Hill, New York.
- [31] Mason, J.R. & Rosner, J. (2007). *Where Did the Risk Go? How Misapplied Bond Ratings Cause Mortgage Backed Securities and Collateralized Debt Obligation Market Disruptions*. Hudson Institute Working Paper.
- [32] Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.
- [33] Moody's Investors Services (1999). *Rating Methodology: The Evolving Meanings of Moody's Bond Ratings*, Moody's Global Credit Research, New York.
- [34] Moody's Investors Services (2007). *Structured Finance Rating Transitions: 1983–2006*. Special Comment, Moody's Global Credit Research, New York.
- [35] Moody's Investors Services (2008). *Corporate Default and Recovery Rates: 1920–2007*. Special Comment, Moody's Global Credit Research, New York.
- [36] Moody's Investors Services (2008). *Introducing Assumption Volatility Scores and Loss Sensitivities for Structured Finance Securities*, Moody's Global Credit Policy, New York.
- [37] Nickell, P., Perraudin, W. & Varotto, S. (2000). Stability of rating transitions, *Journal of Banking and Finance* **24**, 203–227.
- [38] Ong, M.K. (1999). *Internal Credit Risk Models: Capital Allocation and Performance Measurement*, Risk Books, London.
- [39] Peterson, M.A. & Rajan, R. (2002). Does distance still matter: the information revolution in small business lending, *Journal of Finance* **57**, 2533–2570.
- [40] Pluto, K. & Tasche, D. (2005). Thinking positively, *Risk* August, 76–82.
- [41] Resti, A. & Sironi, A. (2007). *Risk Management and Shareholders' Value in Banking*, John Wiley & Sons, New York.
- [42] Shumway, T. (2001). Forecasting bankruptcy more accurately: a simple hazard model, *Journal of Business* **74**, 101–124.
- [43] Standard and Poor's (2001). *Rating Methodology: Evaluating the Issuer*, Standard & Poor's Credit Ratings, New York.
- [44] Standard and Poor's (2007). *Principles-Based Rating Methodology for Global Structured Finance Securities*, Standard & Poor's RatingsDirect Research, New York.
- [45] Stefanescu, C., Tunaru, R. & Turnbull, S. (2008). *The Credit Rating Process and Estimation of Transition Probabilities: A Bayesian Approach*. London Business School working paper.
- [46] Trück, S. & Rachev, S.T. (2005). Credit portfolio risk and PD confidence sets through the business cycle, *Journal of Credit Risk* **1**(4), 61–88.
- [47] White, L. (2002). The credit rating industry: an industrial organization analysis, in *Ratings, Rating Agencies and the Global Financial System*, R.M. Levich, C. Reinhart & G. Majnoni, eds, Kluwer, Amsterdam, NL. pp. 41–64.

Related Articles

Collateralized Debt Obligations (CDO); Credit Migration Models; Credit Risk; CreditRisk+; Credit Scoring; Internal-ratings-based Approach; Rating Transition Matrices; Structured Finance Rating Methodologies.

ADAM ASHCRAFT & TIL SCHUERMANN

Portfolio Credit Risk: Statistical Methods

This article gives a brief overview over statistical methods for estimating the parameters of credit portfolio models from default data. The focus is on models for default probability and correlations; for recovery rates, (see **Recovery Rate**). First, a rather general model setting is introduced along the lines of the models of McNeil and Wendin [10] and others, who depict portfolio models as generalized linear mixed models (GLMMs). Then, we describe the most common estimation techniques, which are the method of moments and maximum likelihood. An excellent reference for other estimation techniques is [10], in particular for Bayes estimation.

A Single Obligor's Default Risk

Let D_{it} denote an indicator variable for obligor i 's default in time period t such that

$$D_{it} = \begin{cases} 1 & \text{borrower } i \text{ defaults in } t \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$i \in \mathbb{N}_t, t = 1, \dots, T$, where $\mathbb{N}_t$ is the set of firms under consideration at the beginning of time period t and $n_t = |\mathbb{N}_t|$ is their cardinal number. The default indicator variable can be motivated in terms of a threshold approach wherein default is said to occur when a continuous variable falls below a threshold. This approach is based on the asset value model due to Merton [11] where the firm declares bankruptcy when the value of its asset is below the principal value of its debt at maturity. Let V_{it} denote the asset value return of firm i at time t ($i \in \mathbb{N}_t, t = 1, \dots, T$), or more generally a variable representing an obligor's credit quality. Then, the obligor defaults when V_{it} falls below a threshold c_{it} , that is, $D_{it} = 1 \Leftrightarrow V_{it} < c_{it}$. Crucial parts in credit risk management now are the modeling of V_{it} , the parameterization of c_{it} , and finally the estimation of parameter. In most industry credit risk models, such as credit metrics, credit risk plus, and credit portfolio view, the triggering variable V_{it} is driven by two sources of random factors. The first is a systematic random factor F_t following a distribution G , which affects all firms jointly and therefore cannot

be diversified away. The second is an idiosyncratic part ε_{it} with variance σ^2 , which is specific for each firm, independent between the firms and from the systematic factor. The default threshold c_{it} is mostly modeled *via* credit ratings that reflect an aggregated summary of a firm's risk characteristics. In a simple case, both variables may be expressed as linear functions of the respective risk drivers, such that $V_{it} = -\omega F_t + \varepsilon_{it}$ and $c_{it} = \alpha + \beta' \mathbf{x}_{it}$ where α , β , and ω are unknown parameters, and $\mathbf{x}_{it}$ is a design vector consisting of observable covariates for obligor i , which may be time- and obligor-specific (such as balance sheet ratios) or only time-specific (such as macroeconomic variables). Then the firm defaults if $V_{it} < \alpha + \beta' \mathbf{x}_{it}$. As shown in [7], the aforementioned credit risk models mainly differ in the distributional assumptions regarding the common and the idiosyncratic random factors driving the firm value as discussed below.

The probability of the firm's default conditional on the random factor f_t can be expressed as

$$\begin{aligned} CPD_{it}(f_t) &= P(D_{it} = 1 | \mathbf{x}_{it}, f_t) \\ &= P(V_{it} < \alpha + \beta' \mathbf{x}_{it} | \mathbf{x}_{it}, f_t) \\ &= P(\varepsilon_{it} < \alpha + \beta' \mathbf{x}_{it} + \omega f_t) \\ &= f(\alpha + \beta' \mathbf{x}_{it} + \omega f_t) \end{aligned} \quad (2)$$

which in statistical terms is a *GLMM*; see [10] and the references cited therein. f_t is a realization of the systematic factor F_t , which is called the *time-specific random effect*. $f(\cdot) : \mathbb{R} \rightarrow (0, 1)$ denotes a response or link function given by the distribution of the idiosyncratic random error ε_{it} . In the credit metrics model, the idiosyncratic errors are standard normally distributed ($\sigma^2 = 1$), whereas in the credit portfolio view model the idiosyncratic error follows a logistic distribution ($\sigma^2 = \pi^2/3$). This leads to the common link functions of the probit $f(y) = \Phi(y)$ or the logit function $f(y) = 1/(1 + \exp(-y))$. In the credit risk plus approach, the systematic and the unsystematic factors are linked multiplicatively rather than linearly and their distributions are Gamma and Exponential, respectively. For details, we refer to [7].

The (with respect to the random effect unconditional) probability of default (PD) is given by the

expectation

$$\begin{aligned} PD_{ti} &= P(D_{ti} = 1 | \mathbf{x}_{ti}) \\ &= \int f(\alpha + \beta' \mathbf{x}_{ti} + \omega f_t) dG(f_t) \end{aligned} \quad (3)$$

which depends on the distribution of the random factor. For example, in the credit metrics model, the random effect is assumed to follow a standard normal distribution. Then, in the probit model, the simple expression for the unconditional PD

$$PD_{ti} = P(D_{ti} = 1 | \mathbf{x}_{ti}) = \Phi(\tilde{\alpha} + \tilde{\beta}' \mathbf{x}_{ti}) = \Phi(\tilde{c}_{ti}) \quad (4)$$

results, where $\tilde{c}_{ti} = \tilde{\alpha} + \tilde{\beta}' \mathbf{x}_{ti}$, $\tilde{\alpha} = \alpha / \sqrt{1 + \omega^2}$ and $\tilde{\beta} = \beta / \sqrt{1 + \omega^2}$. The correlation between the latent variables of obligor i and j , $i \neq j$ is given by $\rho_{ij} \equiv \rho = \frac{\omega^2}{1 + \omega^2}$, which is sometimes referred to as *asset correlation* since the latent variables are interpreted as asset value return. See [5] for a detailed description of correlations. For the aforementioned distributions in the credit portfolio view and the credit risk plus approach and the empirical estimation, compare [8].

Portfolio Default Risk

The vector of default indicators of the portfolio is denoted by $\mathbf{D}_t = (D_{t1}, \dots, D_{tn_t})$. Conditional on the systematic random factor and given the $\mathbf{x}_{ti}$, the defaults are independent. Then the joint distribution of defaults conditional on the systematic factor is given by

$$\begin{aligned} P(\mathbf{D}_t = \mathbf{d}_t | \mathbf{x}_{ti}, f_t) &= \prod_{i \in \mathbb{N}_t} P(D_{ti} = 1 | \mathbf{x}_{ti}, f_t)^{d_{ti}} \\ &\times (1 - P(D_{ti} = 1 | \mathbf{x}_{ti}, f_t))^{1-d_{ti}} \end{aligned} \quad (5)$$

which is also known as a *Bernoulli mixture model* [6]. The unconditional distribution (where unconditional refers w.r.t. the random effect) is obtained as

$$P(\mathbf{D}_t = \mathbf{d}_t | \mathbf{x}_{ti}) = \int P(\mathbf{D}_t = \mathbf{d}_t | \mathbf{x}_{ti}, f_t) dG(f_t) \quad (6)$$

Extensions of model (5) include more than one random effect and are, therefore, called *multifactor models*.

A special case of this model results if the obligors are homogeneous in $\mathbf{x}_{ti}$, that is, $\mathbf{x}_{ti} = \mathbf{x}_t$ and thus $c_{ti} = c_t$ for all i . Then, all obligors exhibit the same conditional PD and the Bernoulli mixture distribution (5) drops to the binomial mixture distribution. Let $D_t = \sum_{i \in \mathbb{N}_t} D_{ti}$ be the number of defaults. Then,

$$\begin{aligned} P(D_t = d_t | \mathbf{x}_t, f_t) &= \binom{n_t}{d_t} P(D_{ti} = 1 | \mathbf{x}_t, f_t)^{d_t} \\ &\times (1 - P(D_{ti} = 1 | \mathbf{x}_t, f_t))^{n_t - d_t} \end{aligned} \quad (7)$$

with the unconditional distribution analogous to equation (6). Define the default rate $p_t = d_t / n_t$ as the ratio of defaulting obligors divided by the total number of obligors. As shown in [15, 16], the distribution of the default rate converges against the ‘‘Vasicek’’-distribution if the random effect is standard normally distributed and the number of obligors goes to infinity. The density is then given as

$$\begin{aligned} f(p_t) &= \frac{\sqrt{1 - \rho}}{\sqrt{\rho}} \cdot \exp\left(\frac{1}{2}(\Phi^{-1}(p_t))^2\right. \\ &\quad \left. - \frac{1}{2\rho^2}(c_t - \sqrt{1 - \rho^2} \cdot \Phi^{-1}(p_t))^2\right) \end{aligned} \quad (8)$$

For a thorough description of large pool approximations, see [9].

Estimation Techniques

There are basically two ways of estimating the unknown model parameters. As the first way, one can use asset values and asset value returns as in the KMV approach [1]. Given the level of liabilities, the default probabilities can be derived. Correlation estimates are obtained by calculating historical correlations from asset value returns. As the crucial part of these methods is deriving the asset values and the capital structure of the firm rather than the statistical procedures, they are not discussed here in detail. Instead, we refer to [3] and the references cited therein. As the second way, the parameters can be estimated using time series $\mathbf{d}_1, \dots, \mathbf{d}_T$ of observed default events. The simplest methods can

be employed for the case of the homogeneous portfolio with time-constant parameters where closed-form solutions for the estimators exist. In the GLMM model, more advanced numerical techniques have to be used. Here, we briefly describe the method of moments and the maximum-likelihood method. For Bayes estimation, we refer to [10] and the references cited therein.

Method of Moments

If the obligors in the portfolio or the segment are homogeneous and the parameters are constant, only two parameters are to be estimated, namely, the PD and the correlation. Gordy [7] applies the method of moments estimator to the probit model. He shows that expectation and variance of the conditional default probability are

$$E(CPD(F_t)) = PD \quad (9)$$

and

$$\begin{aligned} \text{Var}(CPD(F_t)) \\ = \Phi_2(\Phi^{-1}(PD), \Phi^{-1}(PD), \rho) - PD^2 \end{aligned} \quad (10)$$

where $\Phi_2(\cdot)$ is the bivariate normal cumulative distribution function for two random variates, with expectation zero and variance one each and correlation ρ . An unbiased estimator for the unconditional PD is given by the average default rate:

$$\bar{p} = \frac{1}{T} \sum_{t=1}^T p_t \quad (11)$$

The left-hand side of equation (10) can be estimated by the sample variance of the default rate:

$$s_p^2 = \frac{1}{T-1} \sum_{t=1}^T (p_t - \bar{p})^2 \quad (12)$$

Given the two estimates, the asset correlation ρ can be backed out numerically from equation (10). Gordy [7] also provides a finite sample adjustment for the estimator. However, this modified estimator turns out to perform similar to the simple estimator [3].

Maximum-likelihood Method

In the limiting case (8), asymptotic maximum-likelihood estimators of the (homogeneous) PD and

the (homogeneous) asset correlation can be derived as

$$\hat{\rho} = \frac{m_2/T - m_1^2/T^2}{1 + m_2/T - m_1^2/T^2} \quad (13)$$

$$\widehat{PD} = \Phi(T^{-1} \sqrt{1 - \hat{\rho}} m_1) \quad (14)$$

where $m_1 = \sum_{t=1}^T p_t$ and $m_2 = \sum_{t=1}^T p_t^2$ [4].

In the general case of the GLMM where obligors are heterogeneous, the log-likelihood is given via equation (6) as

$$\begin{aligned} l = \sum_{t=1}^T \ln \int \prod_{i \in \mathbb{N}_t} P(D_{ti} = 1 | \mathbf{x}_{ti}, f_i)^{d_{ti}} \\ (1 - P(D_{ti} = 1 | \mathbf{x}_{ti}, f_i))^{1-d_{ti}} dG(f_i) \end{aligned} \quad (15)$$

As the log-likelihood function includes solving several integrals, it is numerically optimized w.r.t. the unknown parameters for which several algorithms, such as the Newton–Raphson method, exist and are implemented in many statistical software packages. The integral approximation can be conducted by, for example, the adaptive Gaussian quadrature as described in [12]. Under usual regulatory conditions, the resulting estimators asymptotically exist, are consistent, and converge against normality. See [2], p.243, for a detailed discussion. Applications and estimation results can, for instance, be found in [6, 8, 13, 14]. For the extension of higher dimensional random effects, there are also some approximation methods that can be used, particularly penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL) [10].

Bayes Estimation

Finally, Bayes estimation can be used for estimation as thoroughly shown in [10]. The joint prior distribution of F_t , $\boldsymbol{\beta}$ (including a constant) and some hyperparameters θ can be given as

$$p(\boldsymbol{\beta}, F_t, \theta) = p(F_t | \theta) \cdot p(\theta) \cdot p(\boldsymbol{\beta}) \quad (16)$$

where *a priori* independence between $\boldsymbol{\beta}$ and θ is assumed. Mostly, Markov chain Monte Carlo methods are applied, which can deal with even more complex models than shown here, such as autocorrelated random effects or multifactor models. For a detailed description, we refer to [10].

References

- [1] Bohn, J. & Crosbie, P. (2003). *Modeling Default Risk*, KMV Corporation.
- [2] Davidson, R. & MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*, Oxford University Press, New York.
- [3] Duellmann, K., Kll, J. & Kunisch, M. (2008). *Estimating Asset Correlations from Stock Prices or Default Rates which Method is Superior?* Deutsche Bundesbank Discussion Paper, Series 2: Banking and Finance, Deutsche Bundesbank, Vol 04.
- [4] Duellmann, K. & Trapp, M. (2005). Systematic risk in recovery rates- an empirical analysis of U.S. corporate credit exposures, in *Recovery Risk: The Next Challenge in Credit Risk Management*, E.I. Altman, A. Resti & A. Sironi, eds. Deutsche Bundesbank.
- [5] Frey, R. (2009). Default correlation and asset correlation, *Encyclopedia of Quantitative Finance* **10**, 038.
- [6] Frey, R. & McNeil, A. (2003). Dependent defaults in models of portfolio credit risk, *Journal of Risk* **6**, 59–92.
- [7] Gordy, M.B. (2000). A comparative anatomy of credit risk models, *Journal of Banking and Finance* **24**, 119–149.
- [8] Hamerle, A. & Roesch, D. (2006). Parameterizing credit risk models, *Journal of Credit Risk* **3**, 101–122.
- [9] Kreinin, A. (2009). Large pool approximations for credit loss, *Encyclopedia of Quantitative Finance* **09**, 017.
- [10] McNeil, A.J. & Wendin, J.P. (2007). Bayesian inference for generalized linear mixed models of portfolio credit risk, *Journal of Empirical Finance* **14**, 131–149.
- [11] Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.
- [12] Pinheiro, J.C. & Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model, *Journal of Computational and Graphical Statistics* **4**, 12–35.
- [13] Roesch, D. (2005). An empirical comparison of default risk forecasts from alternative credit rating philosophies, *International Journal of Forecasting* **25**, 37–51.
- [14] Roesch, D. & Scheule, H. (2005). A multi-factor approach for systematic default and recovery risk, *Journal of Fixed Income* **15**, 63–75.
- [15] Vasicek, O.A. (1987). *Probability of Loss on Loan Portfolio*. Working paper, KMV Corporation.
- [16] Vasicek, O.A. (1991). *Limiting Loan Loss Distribution*. Working paper, KMV Corporation.

DANIEL RÖSCH

Recovery Rate

A recovery rate (RR) is the fraction of an obligor's debt that a creditor stands to recover in the event of default. Recovery rates are usually expressed as a percentage of the par value of the claim (RP). Alternatively, recovery rates can be expressed as a percentage of the market value of the claim prior to default (RMV), or as a percentage of an equivalent treasury bond (RT). Recovery rates are closely associated to the concept of loss-given-default (LGD) where $LGD = 1 - RR$. Recovery rates are not known prior to default and can vary between 0 (full loss) and 1 (full recovery). Recovery rate risk in credit portfolios exists because of the uncertainty regarding recovery rates in the event of default.

Along with the probability of default, recovery rates are important parameters in determining the loss distribution of a credit portfolio. For this purpose, the Basel II Accords expressly recommend that the calculation of regulatory capital on banking institutions include the estimated recovery rates on their credit portfolios. The most widespread methodologies for estimating recovery rates use historical averages that are conditioned on the type of credit instrument, seniority (priority of repayment), and collateral [3]. However, these estimation methods do not account for the fact that recovery rates are known to be negatively correlated to the probability of default [1, 2]. The correlation between recovery rates and default probabilities is important because it exacerbates potential losses on credit portfolios. To this effect, recent credit models have attempted to capture the endogenous nature of recovery rates [2, 4, 9, 11]. Furthermore, recent products in the credit derivatives market have enabled the extraction of recovery rates either directly [5] or indirectly [10, 14]. In addition, since 2003, major credit rating agencies have been offering recovery rate ratings based on proprietary models [8].

Historical Recovery Rates

Historical recovery rates for different types of credit securities are considered as important parameters in many credit risk models. There are various ways to estimate historical recovery rates. The most common are value-weighted mean recovery rates, issuer-weighted mean recovery rates, and issuer-weighted

median recovery rates. The value-weighted mean recovery rate is the average recovery rate on all defaulted issuers weighted by the face value of those issues. Issuer-weighted mean recovery rates and the issuer-weighted median recovery rates are the average and median, respectively, of the recovery rates on each issuer. Varma *et al.* [17] report historical recovery rates from 1982 to 2003 internationally. Globally, the value-weighted mean recovery rate for all bonds over that period was 33.8%, whereas the issuer-weighted mean and median recovery rates were 35.4% and 30.9%, respectively. In the United States, the value-weighted mean recovery rate for all bonds over that period was 35.4%, whereas the issuer-weighted mean and median recovery rates were 35.4% and 31.6%, respectively. For sovereign bonds, the value-weighted mean recovery rate for all bonds over that period was 31.2%, whereas the issuer-weighted mean and median recovery rates were 34.4% and 39.8%, respectively. Furthermore, recovery rates will differ depending on seniority and collateral of the bond. For instance, senior secured corporate bonds have a value-weighted mean recovery rate of 50.3%, compared to 22.9% for junior subordinated bonds, over 1982–2003. Carayon *et al.* [7] find that recovery rates on European bonds tend to be smaller. For instance, over 1987–2007, they find that senior secured bonds in Europe recover (issuer weighted) 61% compared to 70.6% in North America. In the Asia-Pacific (excluding Japan) region, Tennant *et al.* [16] find lower recovery rates of 35.61% on senior secured bonds over the 1990–2007 period.

Recovery Rates and Default Risk

The major problem for credit risk models is that there is a large body of empirical evidence suggesting that recovery rates are negatively correlated with default probabilities. High periods of default are associated with low recovery rates and *vice versa*. The correlation between default probabilities and recovery rates may be ascribed to at least two nonmutually exclusive reasons. First, economic downturns can simultaneously cause increases in the probability of default and lower the value of recovered assets. Second, the price at which recovered assets are sold will depend on the financial condition of peer firms [15]. Under the latter argument, in periods of high default, recovered assets are forced to be sold at “fire-sale” prices.

Acharya *et al.* [1] find both theories to be at work in explaining recovery rates. Altman *et al.* [2] empirically estimate the relationship between recovery rates (y) and default rates (x) using one linear and three nonlinear specifications:

$$\begin{aligned} y &= 0.51 - 2.61x; & R^2 &= 0.51 \\ y &= 0.002 - 0.113 \ln(x); & R^2 &= 0.63 \\ y &= 0.61 - 8.72x + 54.8x^2; & R^2 &= 0.65 \\ y &= \frac{0.138}{x^{0.29}}; & R^2 &= 0.65 \end{aligned} \quad (1)$$

All these specifications show a strong negative relationship between default rates and recovery rates.

Economic Features of Recovery Rates

There are several economic features of recovery rates that are important:

1. As described above, recovery rates are negatively correlated with default rates. This is the case when the data is examined historically as shown in [2] as well as when implied from the data, as in [10].
2. Recovery rates are highly variable and depend on regime (see [12]). They vary within rating and seniority class as well.
3. Seniority and industry are statistically significant determinants of recovery rates, as shown by Acharya *et al.* [1]. These authors also find that, in industries with high asset-specificity, recovery rates are lower.

Implied Recovery Rates

Recovery rates can also be implied from prices of certain credit derivatives. One then speaks of “implied” (or risk-neutral) recovery rates, which may not coincide with historically observed recovery rates. Recovery rate swaps are agreements to exchange a fixed recovery rate for the realized recovery rate allowing the market’s expected recovery rate to be directly recovered [5]. Digital credit default swaps (DDS) are credit default swaps (CDSs) where the recovery rates on default are prespecified, irrespective

of the final recovery rate. Arthur and Kapoor [6] show how recovery rates can be recovered using a DDS and a CDS. Finally, Pan and Singleton [14] and Das and Hanouna [10] use CDS with different maturities to extract default probabilities and recovery rates.

Approximately, if credit spreads are known, we may write the spread s as a function of default probability (λ) and recovery rate (ϕ): $s \approx \lambda(1 - \phi)$, implying that recovery may be written in a reduced-form setting as follows:

$$\phi = 1 - \frac{s}{\lambda} \quad (2)$$

More formalized and exact versions of this approximate relation may be derived from a CDS pricing model or a bond pricing model. Recovery may also be derived in the class of Merton [13] models. The expression for recovery rate is

$$\begin{aligned} E[\phi] &= E \left[\frac{V_T}{D} | V_T < D \right] = \frac{1}{D} E [V_T | V_T < D] \\ &= \frac{V_0}{D} e^{rT} \{1 - N(d_1)\} \\ d_1 &= \frac{\ln(V_0/D) + (r + \frac{1}{2}\sigma_V^2)T}{\sigma_V \sqrt{T}} \end{aligned} \quad (3)$$

where $\{V_0, \sigma\}$ are the initial value and volatility of the firm, D is the face value of debt with maturity T , and r is the risk-free interest rate. $N(\cdot)$ is the normal distribution function.

References

- [1] Acharya, V., Bharath, S.r & Srinivasan, A. (2007). Does industry-wide distress affect defaulted firms? Evidence from creditor recoveries, *Journal of Financial Economics* **85**(3), 787–821.
- [2] Altman, E., Brady, B., Resti, A. & Sironi, A. (2004). The link between default and recovery rates: theory, empirical evidence and implications, *Journal of Business* **76**(6), 2203–2227.
- [3] Altman, E., Resti, A. & Sironi, A. (2003). *Default Recovery Rates in Credit Risk Modeling: A Review of the Literature and Empirical Evidence*, working paper, New York University.
- [4] Bakshi, G., Madan, D. & Zhang, F. (2001). *Recovery in Default Risk Modeling: Theoretical Foundations and Empirical Applications*, working paper, University of Maryland.

-
- [5] Berd, A.M. (2005). Recovery swaps, *Journal of Credit Risk* **1**(3), 1–10.
 - [6] Berd, A. & Kapoor, V. (2002). Digital premium, *Journal of Derivatives* **10**(3), 66.
 - [7] Carayon, J.-M., West, M., Emery, K. & Cantor, R. (2008). *European Corporate Default and Recovery Rates, 1985–2007*, Moody's investors service.
 - [8] Chew, W.H. & Kerr, S.S. (2005). Recovery ratings: a new window on recovery risk, in *Standard and Poor's: A guide to the Loan Market*, Standard and Poor's.
 - [9] Christensen, J. (2005). *Joint Estimation of Default and Recovery Risk: A Simulation Study*, working paper, Copenhagen Business School.
 - [10] Das, S.R. & Hanouna, P. (2009). Implied Recovery, forthcoming, *Journal of Economic Dynamics and Control*.
 - [11] Guo, X., Jarrow, R. & Zeng, Y. (2005). *Modeling the Recovery Rate in a Reduced Form Model*, working paper, Cornell University.
 - [12] Hu, W. (2004). *Applying the MLE Analysis on the Recovery Rate Modeling of US Corporate Bonds*, Master's Thesis in Financial Engineering, University of California, Berkeley.
 - [13] Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *The Journal of Finance* **29**, 449–470.
 - [14] Pan, J. & Singleton, Ken (2008). Default and recovery implicit in the term structure of sovereign CDS spreads, *Journal of Finance* **63**, 2345–2384.
 - [15] Shleifer, A. & Vishny, R. (1992). Liquidation values and debt capacity: a market equilibrium approach, *Journal of Finance* **47**, 1343–1366.
 - [16] Tennant, J., Emery, K., Cantor, R., Elliott, J. & Cahill, B. (2007). *Default and Recovery Rates of Asia-Pacific Corporate Bond, Moody's Investors Service and Loan Issuers, Excluding Japan, 1990–1H200*.
 - [17] Varma, P., Cantor, Richard & Hamilton, David (2003). *Recovery Rates on Defaulted Corporate Bonds and Preferred Stocks, 1982–2003*, Moody's investors service.

Related Articles

Credit Default Swaps; Credit Risk; Exposure to Default and Loss Given Default; Recovery Swap.

SANJIV R. DAS & PAUL HANOUNA

Internal-ratings-based Approach

Within the new Basel capital rules for banks (*see Regulatory Capital*), the internal-ratings-based approach (IRBA) represents perhaps the most important innovation for regulatory minimum capital requirements. For the first time, subject to supervisory approval, banks are allowed to use their own risk assessments of credit exposures in order to determine the capital to be held against them. Within the IRBA, banks estimate the riskiness of each exposure on a stand-alone basis. The risk estimates serve as input for a supervisory credit risk model (implicitly given by risk weight functions) that provides a value for capital that is deemed sufficient to cover against the credit risk of the exposure, given the assumed portfolio diversification. In order to obtain supervisory approval for the IRBA, banks must apply for IRBA and fulfill a set of minimum requirements. Until approval is granted for the entire book or specific portfolios, banks must apply the simpler and less risk-sensitive standardized approach for credit risk, where minimum capital requirements are determined in dependence on asset class (sovereign, bank, corporate, or retail exposure) only and, if applicable, ratings by external credit assessment agencies like rating agencies or credit export agencies.

The Conception of Internal Rating Systems in Basel II

Bank internal rating systems are, in the most general sense, risk assessment procedures, which are used for the assignment of internally defined borrower and exposure categories. A rating system is based on a set of predefined criteria to be evaluated for each borrower or exposure subject to the system, and result in a final “score” or “rating grade” for the borrower or exposure. The choice and weighting of the criteria can be manifold; there are no rules or guidance on which criteria to include or exclude. The main requirement on IRBA systems is that their rating grades or scores do indeed discriminate borrowers according to credit default risk.

In practice, rating systems are often designed as purely or partly statistical tools, for example, by

identifying criteria (typically financial ratios) with good discriminatory power and combining them by means of statistical regression or other mathematical methods.

However, in order to use such tools, there must be sufficient historical data—both on defaulted and surviving borrowers and exposures—for determining the discrimination criteria and calibrating their weightings. In practice, obtaining such data often proves to be more difficult than the statistical analysis as such, either because historically borrower and exposure characteristics were not stored in a readily usable manner, or simply because for some portfolios there is not sufficient default data. In general, rating systems may include a set of quantitative and some qualitative criteria. The weighting of these criteria may also be determined by expert opinion rather than by statistical tools. In the extreme, for example, in international project finance where certain criteria are deal breakers for loan arrangements (i.e., the existence of sovereign risk coverage *via* export insurance for projects in regions with high political risk), there might be no predetermined weighting scheme at all.

Notions appear to be not entirely uniform in practice. Often, but not always, the notion of a “scoring system” or a “score card” is used for a purely statistical rating system or the statistical part of a mixed quantitative and qualitative rating system. Moreover, the notion of scores tends to be more often used for retail and small business portfolios, while for corporate, bank, and sovereign portfolios, the literature tends to speak of rating systems. From an IRBA perspective, there are no conceptual differences between these notions: they all depict different forms of IRBA systems. Likewise, there is no IRBA requirement for the number of rating systems a bank should apply. Usually, one would expect different systems for retail, small businesses and self-employed borrowers, corporates, specialized lending portfolios, sovereigns, and banks. Many of these asset classes might again see different rating systems, depending on, for example, product type (very common for retail portfolios, but not constrained to them) or sales volume and region (both common for corporate portfolios), because the different borrower and exposure categories might call for different sets of rating criteria. Within a large, internationally active and well-diversified bank, one might expect to see a large number of different rating systems.

IRBA Risk Parameters

Credit risk per rating grade is quantified by probabilities of default (PDs), which give the probabilities of borrowers to default on their obligations, with regard to the Basel default definition, within 1 year's time. The PD per rating grade is usually estimated by the use of bank internal historical default data, which may be supplemented by external default data. A specific problem within the IRBA comes from the fact that not all institutions have readily available default data according to the Basel definition. In this case, adjustments to the estimates must be made.

PDs may be estimated just for the next year (point in time (PIT)) or as long term average PDs (through the cycle (TTC)). PIT estimates take into account the current state of the economy—as a consequence, PDs per rating grade might change over time—while TTC estimates do not. The Basel Accord seems to implicate TTC estimates. Nonetheless, many supervisors might be prepared to accept PD estimates that are more of PIT type because eliminating all cyclical effects from rating systems and PD estimates might be difficult to afford in practice.

In the Basel sense, an IRBA rating system contains two additional dimensions: an exposure at default (EAD) dimension, assessing the expected exposure at the point in time when the borrower defaults, and a loss given default (LGD) dimension, measuring the expected percentage of exposure that is lost in case a borrower defaults. The EAD dimension is mainly driven by product characteristics, for example, how easily lines can be drawn by the borrower or reduced by the bank prior to default, while the LGD dimension is heavily dependent on collateral, guarantees, and other risk mitigants. Here again, Basel notions slightly differ from literature and practice: in industry, the notion of a rating system often refers to the PD dimension only, while for Basel it includes all three dimensions. Within the IRBA, banks must provide a PD, LGD, and EAD for all their exposures. While the PD must always be estimated by the bank itself, banks can choose whether they want to use supervisory LGD and EAD estimates for given product and collateral types (thus applying the so-called foundation IRBA) or whether they want to estimate these values themselves, too (thus using the so-called advanced IRBA).

The Basel II Risk Weight Functions

In order to assess the overall risk of a bank portfolio, credit portfolio risk models have to evaluate the portfolio composition and its diversification. Within the Basel II IRBA, banks are not allowed to use their own credit portfolio risk models and diversification estimates for minimum regulatory capital ratios. Rather, they must input the risk parameters PD, LGD, and EAD into a supervisory credit portfolio risk model. This model can be roughly described as Vasicek's [6] multivariate extension of Merton's [5] model of the default of a firm. In statistical terms, the model could be characterized as a one-factor probit model where the events to be predicted are the borrowers' defaults and the single factor reflects the state of the global economy. Moreover, the Basel model assumes an infinitely granular portfolio, such that all idiosyncratic single name risk is diversified away. In this sense, the Basel model is an asymptotic single risk factor (ASRF) model. For further details of the model, see [1, 4].

The assumptions of a single risk factor and of infinite granularity lead to the following characteristics of the Basel credit risk model:

1. The capital charge per exposure can be described in closed form by risk weight functions (cf [3], paragraphs 272 and 273 for corporate, sovereign, and bank exposures and paragraphs 328, 329, and 330 for retail exposures). The specifications of the risk weight function for the different exposure classes can be derived from the following generic formula for the capital requirement K per dollar of exposure:

$$K = \text{LGD} \left[N \left(\frac{G(\text{PD})}{\sqrt{1-R}} + G(0.999) \sqrt{\frac{R}{1-R}} \right) - \text{PD} \right] \frac{1 + (M - 2.5)b}{1 - 1.5b} \quad (1)$$

In this equation,

- the probability of default PD and the loss given default LGD are measured as decimals;
- the *maturity adjustment* b is given by $(0.11852 - 0.05478 \ln(\text{PD}))^2$;

- N denotes the standard normal distribution function;
- G denotes the inverse standard normal distribution function;
- the *effective maturity* M was fixed at 1 year for retail exposures, and assumes values between 0 and 5 years for other exposures as described in detail in paragraph 320 of [3].

The risk weight functions for the different exposure classes differ mostly in the specification of the *asset correlation* R . For the retail mortgage exposure class, R was fixed at 15%. For revolving retail credit, R is 4%. In contrast, in the corporate, sovereign, and bank exposure classes, R depends on PD by

$$R = 0.12 \frac{1 - e^{-50PD}}{1 - e^{-50}} + 0.24 \left(1 - \frac{1 - e^{-50PD}}{1 - e^{-50}} \right) \quad (2)$$

and also for other retail exposures, R is given as a function of PD by

$$R = 0.03 \frac{1 - e^{-35PD}}{1 - e^{-35}} + 0.16 \left(1 - \frac{1 - e^{-35PD}}{1 - e^{-35}} \right) \quad (3)$$

2. The capital charge per exposure depends only on the risk parameters PD, LGD, and EAD of the exposure, but not on the portfolio composition. Thus, the capital charge for each exposure is exactly the same, no matter which portfolio it is added to (portfolio invariance). From a supervisory point of view, portfolio invariance was an important characteristic in developing the Basel risk weight functions, as it ensures computational simplicity and the preservation of a level playing field between well diversified and specialized banks. The capital charge for the entire portfolio is the sum of the capital charges for individual exposures. The downside of portfolio invariance is that the Basel formula cannot account for risk concentrations. If it did, the capital charge for an exposure would again have to depend on the portfolio to which it is added. If banks are concerned about concentration effects

or want to measure the diversification benefits of their portfolio, they need to develop their own, fully fledged credit risk models.

3. The capital charge for each exposure, given its risk parameters, depends on the correlation with the single systematic risk factor and the so-called confidence level. The confidence level for minimum capital requirements was set by the Basel Committee to be 99.9%. As a consequence, the probability that the bank will suffer losses from the credit portfolio that exceed the capital requirements should be of an order of magnitude like 0.1%. The correlations were estimated from supervisory data bases and are assumed to decrease with decreasing creditworthiness.
4. The ASRF takes only the default event as stochastic and treats the loss in case of default as deterministic. As in practice loss amounts are stochastic as well, and potentially correlated with the drivers of the default events, banks are supposed to take account of this effect in their LGD estimates, by estimating the downturn LGDs instead of average LGDs.
5. Lastly, the ASRF is a default mode (DM) model that only accounts for losses due to defaults within a given time horizon (1 year) but not for losses due to rating migrations and future losses after 1 year. This simplification does not comply with modern accounting practice. It was therefore adjusted by introducing the maturity adjustments, which can be seen as an extension of the model toward a marked-to-market (MtM) mode.

Minimum Requirements

In order to apply the IRBA, banks must have explicit approval from their supervisors. Approval is subject to a set of minimum requirements aimed to ensure the integrity of the rating model, rating process, and thus of the risk parameters and capital charges. The minimum requirements ([3], Part 2, Section III. H) hence assemble around the following themes:

- *Rating system design.* As mentioned before, there are no regulatory requirements with regard to the rating criteria. Rating grades must, in a sensible way, discriminate for credit risk, and the onus of proof is with the bank. Moreover, there must be at least seven rating grades for performing and one grade for nonperforming

exposures in the PD dimension. No minimum grade numbers are given for the LGD and EAD dimension. Also, there is no requirement of a common master scale across all rating systems, although many banks develop such a scale for internal risk management and communication purposes.

- *Rating system operations.* By this set of minimum requirements, banks are asked to ensure the integrity of the rating process. Most notably, the rating assignments must be independent from any business units gaining from credit approval (e.g., the sales department). Moreover, there should be no “cherry picking” between rated and nonrated exposures (the latter being treated in the less risk-sensitive standardized approach), although a temporary partial use of IRBA, coupled with a supervisory approved implementation (roll-out plan) for bankwide IRBA use, and a permanent partial use for insignificant portfolios are allowed. Another important aspect is the integration of the ratings into day-to-day credit processes, including IT systems and input data availability.
- *Corporate governance and oversight.* This set of criteria requires banks to embed their rating systems into the overall governance structure of the bank. Most notably, senior management is supposed to buy into the systems and formally approve for wide use within the bank, such that the systems become accepted risk management tools at all levels in the organization. Also, the role of internal audit in regular rating audits is defined.
- *Use of internal ratings.* Banks will only receive IRBA approval if they use their ratings for a wide range of bank internal applications. Examples include credit approval, limit systems, risk-sensitive pricing and loss provisioning. Rating systems solely developed for regulatory purposes will not be recognized, as only the deep rooting into day-to-day credit risk management actions will ensure their integrity.
- *Risk quantification.* Banks need to quantify the risk parameters PD, LGD, and EAD, based on their rating grades and on the Basel default and loss definitions ([3], paragraphs 452, 453, and 460). In doing so, they should employ a variety of data sources: preferentially internal data, but

enhanced with external data sources and expert judgment if needed.

- *Validation of internal estimates.* The PD, LGD, and EAD estimates must be validated against actually observed default rates and losses. Owing to relatively short time series for the latter, validation remains one of the more difficult issues within the IRBA. For available statistical techniques see, for example, [2]. Where statistical validation is not reliable, banks should use more qualitative validation techniques, like ensuring good rating process governance, integrity of the input data, and so on.
- *Disclosure requirements.* Banks that use the IRBA must base their capital and risk disclosure requirements (the Third Pillar of Basel II) on their IRBA figures.

In practice, compliance with the minimum requirements often seems to prove much more difficult and costly than the development of the rating systems as such. The most difficult issues seem to be data availability, IT system implementation, and data feed, the actual rating of entire portfolios, which often require large amounts of data for all exposures to be fed into the systems (in the worst case manually, as often data not consistent with the rating criteria have been stored in the past) and, connected to this, the buy-in of senior management and the entire credit business into the more risk-sensitive and more transparent IRBA.

Implications for the Bank Internal Use of IRBA Figures

Risk quantification *via* IRBA can be of great use for the bank internal credit risk measurement and management. However, there are some limitations. The most important of these is surely that due to the asymptotic single risk factor model, the IRBA provides no measure of risk concentrations, be they single name, industry, or regional concentrations. If banks are concerned about concentration effects or—as the other side of the same coin—want to measure the diversification benefits of their portfolio, they need to go further and develop their own, full-fledged credit risk models with more than one risk factor and their own correlation estimates. Likewise, the asymptotic assumption needs to be given up in order to capture idiosyncratic single name risk.

The most significant benefit of the IRBA for bank internal risk management lies in the standardized assessment and measurement of stand-alone borrower and exposure credit risk. Credit risk becomes much more transparent within the organization, and there is “one common currency” for risk, expressed by the risk parameters PD, LGD, and EAD and the regulatory capital charges based on them.

References

- [1] BCBS (2004). *An Explanatory Note on the Basel II IRB Risk Weight Functions*. Basel Committee of Banking Supervision.
- [2] BCBS (2005). *Studies on the Validation of Internal Rating Systems*. Basel Committee of Banking Supervision, Working Paper No. 14.
- [3] BCBS (2006). *International Convergence of Capital Measurement and Capital Standards. A Revised Framework,*

Comprehensive Version. Basel Committee of Banking Supervision.

- [4] Gordy, M. (2003). A risk-factor model foundation for ratings-based bank capital rules, *Journal of Financial Intermediation* **12**(3), 199–232.
- [5] Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**(2), 449–470.
- [6] Vasicek, O.A. (2002). The distribution of loan portfolio value, *Risk* **15**, 160–162.

Related Articles

Credit Rating; Credit Risk; Credit Scoring; Economic Capital; Exposure to Default and Loss Given Default; Large Pool Approximations; Regulatory Capital.

KATJA PLUTO & DIRK TASCHE

Exposure to Default and Loss Given Default

In the study of credit risk, the most relevant factor has traditionally been the borrower's probability of default (or intensity of default), expressing default risk and, indirectly, migration risk. However, there are other risk profiles that significantly affect the loss experienced by the lender upon the occurrence of a default: exposure at default and loss given default. The uncertainty surrounding these variables gives rise, respectively, to exposure risk and recovery risk. These risks (captured through parameters like EAD, LGD, and RR, as explained below) have become increasingly popular thanks to the preliminary drafts of the new accord on bank capital requirements ("Basel II") that were circulated by the Basel Committee after 1999 and led to a new regulatory text in 2004 [12].

Exposure at Default and Exposure Risk

In the simplest forms of credit exposure, the amount due to the lender in the event of a default (that is, the exposure at default (EAD)) is known with certainty. This is the case, for example, of zero-coupon bonds or fixed-term loans, where the balance outstanding is predetermined in advance and cannot be modified without a formal credit restructuring.

However, the amount outstanding in the event of a default might also be uncertain, basically due to the following reasons:

1. changes in the value of the contract to which the defaulted party had committed itself (typically, an OTC derivative affected by a number of underlying variables);
2. the presence of a revolving credit line (e.g., a loan commitment) where the borrower could increase his/her credit usage before default.

While case 1, known as *counterparty risk*, can be considered as a sort of intersection between credit and market risk, case 2 represents a typical example of exposure risk. Here, the borrower's current exposure (that is, the drawn part of the credit line, DP) can increase to a larger EAD, with the increase (ΔE)

potentially as large as the current unused portion (UP) of the credit line. To account for exposure risk, banks compute credit conversion factors (CCF) as $CCF \equiv \Delta E / UP$.^a Once a set of CCFs, associated with different types of borrowers and exposures, has been estimated, a bank can forecast EAD as

$$EAD = DP + CCF \cdot UP \quad (1)$$

CCFs are usually calibrated through a statistical analysis of past defaults (see, e.g., [9, 11, 25, 27]), where the CCF is explained through the characteristics of the borrower, the exposure, and the economic environment. When past events are analyzed, the UPs must be recorded some time before the default: this can be a fixed interval ("fixed time horizon," e.g., 12 months before the default) or a fixed moment in time for all defaults that occurred in the same period ("cohort approach," e.g., January 1 for all exposures defaulted in a given year); multiple UPs can also be recorded at several different instants in time ("variable time horizon," e.g., 6, 12, 24 months before default) to assess the impact of time-to-default on exposure risk.^b

In fact, CCFs can be expected to increase with time-to-default: a study based on some 400 borrowers in the period 1995–2000 [9] has shown that one-year CCFs average 32%, while five-year CCFs average 72%; this may be due to a rating migration effect and a greater opportunity to draw down. CCFs also seem to be driven by the percent usage ratio of the credit line ($DP / (DP + UP)$): lower usage rates are usually associated with higher CCFs and with better ratings [9].

A well-known relationship also exists between ratings and CCFs: indeed, the latter have often been found to increase for borrowers with better ratings.^c In other words, exposure risk is especially significant when default risk is comparatively low. This is an expected result, given that firms with investment-grade ratings can get funds from the commercial paper market or by negotiating better terms with their suppliers, and hence tend to use a small portion of the available credit lines (which are comparatively more expensive); however, as their financial shape deteriorates and default gets closer, firms quickly resort to bank credit lines, as other sources of funds dry up.

Besides focusing on loan behavior at default, one can assess exposure risk by monitoring credit usage throughout the life of a facility, including both defaulted and performing exposures. These "usage

ratios” have been found to behave very differently for firms that eventually default, even several years later, as opposed to nondefaulting obligors. For example, a sample of about 770 000 lines of committed credit lines recorded in the Spanish central credit register^d shows that defaulting exposures have a median usage ratio of 50%, in contrast to 43% for nondefaulting facilities; this median usage ratio was found to increase (71%) in the last year before default. Usage ratios are instead lower, all other things being equal, for “seasoned” credit lines (i.e., credit lines that have been in place for a number of years); this suggests that relationship banking may play a role in preventing usage peaks in credit lines.

Other borrower characteristics may also help explain exposure risk: for example, usage ratios have been found to be higher for younger, smaller, and less profitable firms (as age, size, and profitability tend to be inversely related to PD, this is consistent with poorly rated companies being more dependent on bank credit lines).^e Other important explanatory variables are the borrower’s leverage, liquidity, and debt cushion; also, exposure risk tends to be higher for larger companies and for those having a larger share of bank debt in their liabilities mix [25]. However, generally, firm characteristics tend to have a comparatively limited impact on CCFs and usage statistics.

Exposure risk also seems to be affected by the macroeconomic cycle. For example, the gross domestic product (GDP) growth rate has been found [27] to be inversely related to credit line usage, and such a link is especially meaningful in the case of a slow-down or recession. This makes sense, as credit lines are often used to provide a liquidity buffer for borrowers in times of financial strain.

Other measures have been proposed as an alternative to CCFs: these are the *EAD factor*, $EADF = EAD/(DP + UP)$, and the *exposure multiplier*, $EM = EAD/DP$. The former can be considered as a special case of the usage ratio, recorded at the time of default; the latter cannot be computed when a credit line was totally undrawn before the borrower’s default^f.

CCFs can usually be expected to lie between 0 (if the UP is still unused at default) and 1 (if the whole UP gives rise to an extra exposure). However, the ΔE and hence the CCF could also be negative; this is likely to be the case if the credit line is revocable or has some covenant entitling the bank to claim its money back before a proper default takes

place. Curiously, however, Basel II states that CCFs cannot be set below zero, regardless of any empirical evidence that a bank may produce to its supervisors.

Apart from OTC derivatives and credit lines, exposure risk can also arise from the issuance of guarantees and other off-balance-sheet items (e.g., letters of credit, bid bonds, and performance bonds) that might be used by third parties to get relief after the default of the guaranteed entity (leading to monetary outflow for the guarantor, that is, to an EAD). In this case, the EAD can be anywhere between zero and the amount of the off-balance sheet item (OBS), and CCFs can be computed as $CCF \equiv EAD/OBS$. CCF estimates associated with different types of guarantees and OBS can then be used to forecast the EAD as $EAD = CCF \cdot OBS$

Loss Given Default and Recovery Risk

The loss rate given default—or simply loss given default (LGD^g)—is the loss rate experienced by a lender on a credit exposure if the borrower defaults. It is given by 1 minus the recovery rate (*RR*) (see equation (3)) and can take any value between 0 and 100%. Formally,

$$LGD = 1 - RR \quad (2)$$

LGD is never known when a new loan is issued, although a reasonable estimate can be produced when the default occurs, at least if there is a secondary market where the defaulted exposure can be traded. In fact, RRs can be computed based on several approaches [8,33]:

1. The *market LGD* approach uses prices of defaulted exposures as an estimate of the RR. In practice, if a defaulted bond trades at 30 cents a euro, one can infer that the market is estimating a 30% RR (hence, a 70% LGD). This approach can be used only for exposures traded on a secondary market.

A variation of this approach (*emergence LGD* approach) estimates the RR on the basis of the market value of the new financial instruments (usually, shares or long-term bonds) that are offered to lenders in exchange for their defaulted claims. These are usually issued only when the restructuring process is over and the company

emerges from default; their market price must, therefore, be discounted back in time to the moment when the default took place, using an adequate discount rate.

A third version of market LGD involves the use of spreads on performing bonds as a source of information; in fact, spreads on corporate bonds depend on both the borrower's PD and the expected RR. Assuming the PD can be estimated otherwise, one can then work out the LGD implied by market spreads (*implicit market LGD*); alternatively, by assuming that some relationship exists between PD and LGD (see below), PD and LGD can be derived jointly [13]. Note that implicit market LGD makes it possible to use a considerably larger dataset, including performing exposures, and not only defaulted ones. However, note that LGDs derived from market prices often are risk-neutral quantities; therefore, some assumption on the relationship between them and real world LGDs is needed if implicit market LGDs are to be used.

2. When market data are not available (as for most traditional banking loans, where no secondary market exists) one must turn to the *workout approach*. This is based on the actual recoveries (and recovery costs) experienced by the lender in the months (years) after the default took place. It therefore requires to set up a database, where all recoveries on defaulted exposures are filed. According to this approach, the RR (also known as *ultimate recovery*) can be computed using the following equation:

$$RR = \frac{\sum_i R_i \cdot (1+r)^{-T_i}}{EAD} \quad (3)$$

where R_i is the i th recovery flow associated with the defaulted exposure (negative R_i s denote recovery costs), r is the appropriate discount rate,^h and T_i is the time elapsed between the default and the i th recovery. Note that, based on equation (3), RR can be negative (hence LGD can exceed 100%) if recoveries do not offset recovery costs.

The determinants of RRs have been extensively investigated, mainly based on the market LGD approach, although some examples of workout LGDs exist (mainly for bank loans). Indeed, one of the first

studies estimating RRs in the 1970s [6] was based on a survey carried out among the workout departments of a number of large banks in the period 1971–1975; the average recovery on unsecured loans (based on the face value of cash flows on defaulted exposures, recorded in the first three years after default and not discounted) was found to be about 30%.

In the following years, recoveries on bank loans have been foundⁱ to be affected by many factors, including the size of the loan and different collateral types. More generally, the four main drivers of RRs and LGDs can be summarized as follows:

Exposure characteristics

These include the presence of any collateral (be it represented by financial assets of other goods, such as plants, real estate, inventories) and its degree of effectiveness (that is, how easily it can be seized and liquidated); the priority level of the exposure, which can be senior or subordinated to other exposures; any guarantees provided by third parties (like banks, holding companies of public sector entities). An important driver of recoveries is also the exposure's "debt cushion", that is, the amount of the liabilities in the borrower's balance sheet that are junior to the one being evaluated; as the volume of such junior securities increases, so does the RR on the senior exposure, as its holders are more likely to find an adequate volume of assets to be liquidated and used as a source of cash [28,34].

Borrower characteristics

These include the industry where the company operates, which may affect the liquidation process, that is, the ease with which the firm's assets can be sold and turned into cash for the creditors,^j the country of the obligor, which affects the speed and effectiveness of the bankruptcy procedures; some financial ratios, like the leverage (namely, the ratio between total assets and liabilities, which shows how many euros of assets are reported in the balance sheet for each euro of debt to be paid back) and the ratio of EBITDA (earnings before interest, taxes, depreciation, and amortization) to total turnover (which indicates whether the defaulted company is still capable of generating an adequate level of cash flow for its would-be borrowers). Another interesting variable affecting LGD is the borrower's *original* rating: indeed, "fallen angels" (i.e., investment-class obligors that were downgraded to junk) appear to behave

differently from straight speculative-grade issuers, and have been found to recover significantly more than bonds of the same seniority that were rated as speculative-grade at issuance.^k

Lender (e.g., bank) characteristics

These may include the efficiency levels of the department that takes care of the recovery process (workout department) or the frequency with which out-of-court settlements are reached with the borrowers, or nonperforming loans are spun-off and sold to third parties; in fact, sales of nonperforming loans and out-of-court settlements, while reducing the face value of the recovery (compared to what could be obtained by the bank on the basis of a formal bankruptcy procedure), also significantly shorten the duration of the recovery process. The financial effect of this shorter recovery time usually more than offsets the lower recovered amount.

Macroeconomic variables

These mainly include the level of the interest rates (higher rates reduce the present value of recoveries) and the state of the economic cycle (if the economy is in recession, the value at which the companies assets can be liquidated is likely to be lower).

During the last years, an important stream of research has addressed the relationship between PD and LGD. From a theoretical point of view, the same macroeconomic background variables that affect the default probability of the borrowers (and cause default rates to rise) may drive down the liquidation value of assets and increase LGD (so that the distribution of LGDs is different in high-default and low-default periods).^l This intuition has prompted a number of models^m generalizing the “classic” single-factor model in [17] and [22] to the case where recoveries and defaults are driven by a common component (usually systemic in nature).

From an empirical point of view, several pieces of evidence indicate that LGDs and default rates tend to increase together when the overall economic cycle deteriorates. For example, using data on US corporate bonds (Moody’s Default Risk Service database) for 1982–1997, one finds that in a severe economic downturn (when defaults are more frequent), recoveries can be expected to drop by 20–25% compared with their unconditional average [20]. Similar results are found using Standard and Poor’s Credit Pro database (bond and loan defaults) for 1982–1999

[15], as well as junk bond data for 1982–2000.ⁿ Evidence of a strong relationship between LGD and the state of the economy, including default frequencies, is also found by Moody’s KMV in its LossCalc model [23], estimated on a dataset of over 3000 recoveries on loans, bonds, and preferred stock.

The correlation between economic cycle and recoveries appears stronger if estimated at the industry level [1]. In fact, if the sector where the borrower used to operate is undergoing a recession, the lender will find it more difficult to find a buyer for the defaulted company or its assets (as competitors are likely to suffer from excess production capacity) and recoveries will be lower than expected. As recessions may occur at the industry level when the economy as a whole is doing reasonably well, moving from economy-wide to industry-specific conditions can make the empirical link between default rates and recoveries much easier to detect.

The PD/LGD correlation has wide-ranging implications for credit risk models. First, the expected loss rate can no longer be considered as the product of the expected LGD times the borrower’s unconditional PD, since a second, positive addendum must be factored in, accounting for covariance. Second, unexpected loss and Value at Risk prove to be considerably higher than they are if independence is assumed, as shown by [7]; in other words, if systematic risk plays an important role for RRs, estimates of economic capital turn out to be downward biased.^o

While most RR studies focus on mean or median values, it is also important to understand the whole probability distribution of recoveries, if extreme scenarios are to be fully understood and managed. In the case of bank loans, the probability distribution of workout LGDs is usually strongly bimodal, with peaks at 0% and 100%. In the case of bonds, unimodal distributions may be sensible, but still it is strongly advisable to use flexible distributions, such as the beta (which can be either uni- or bimodal depending on the estimated parameters, and can easily be fit to the data by the generalized method of moments).^p

Finally, it is worth emphasizing that, as with all other risks, recovery risk may also produce profits. Indeed, the price performance of defaulted bonds (estimated by comparing market LGDs to emergence LGDs) can prove extremely brilliant, although this is not always the case: while senior bonds (both secured and unsecured) have been found to perform

very well in the postdefault period (with per annum returns of 20–30%), junior bonds often show negative returns [3].

Acknowledgments

Part of this article, especially the LGD section, draws on previous work carried out with Andrea Sironi, to whom I wish to express my gratitude.

End Notes

^aCCFs are sometimes also known as *loan equivalents* (*LEQs*).

^bSee [30] for further details on fixed time horizon, cohort approach, and variable time horizon.

^cSee, for example, [11], where a sample of loan commitments in 1987–93 is analyzed, [25], based on 3281 defaulted exposures issued by 720 borrowers in 1985–2006, or [9].

^dThese are all loan commitments above €6000 issued by Spanish banks after 1984. See [27] for further details.

^eSee again [27], based on a subset of about 86 000 companies.

^fThe EM is sometimes referred to as the *CCF* (in which case, what we called CCF is indicated as *LEQ*. Note that, given the important role played by bank capital regulation in shaping credit risk measurement techniques and jargon, we chose to use the word CCF in a way that is consistent with the terminology of the new Basel accord.

^gIn principle, one should indicate the loss rate given default as LGDR (LGD rate) and use LGD for the absolute LGD (in euros or dollars). However, “LGD” is used by most practitioners (and by the new Basel accord on bank capital) to indicate the loss rate, while the absolute loss is usually indicated as $LGD \cdot EAD$.

^hThe choice of a suitable risk-adjusted r is far from trivial, and basically depends on the amount of systemic risk of the defaulted exposure. See [29].

ⁱSee, for example, [10], based on 24 years of data compiled by Citibank, or [14], using a sample of 371 loans issued by Portugal’s largest private bank during 1985–2000; both studies are based on the workout approach. A study on bank loans (large syndicated loans traded on the secondary market) based on the market LGD approach is, for example, [16].

^jSee [1] based on market LGDs observed during the United States during 1982–1999. See also [5] and the literature survey in [33].

^kSee [4], based on a sample of corporate bonds stratified by original rating and seniority: in the case of senior-secured exposures, for example, the median RR for fallen angels was 50.5% versus 33.5%.

^lA somewhat different approach has been proposed by Peura and Jovivuolle [31]. Using an option-pricing, *à la*

Merton framework, they present a model where collateral value is correlated with the value of the borrower’s assets and hence to his/her PD. This leads to an *inverse* relationship between default rates and RRs.

^mSee [19–21]. Jarrow [26] presents a model where, as in Frye’s works, RRs and PDs are correlated and depend on the state of the economy; however, his methodology explicitly incorporates equity prices in the estimation procedure, allowing the separate identification of RRs and PDs and the use of a larger dataset. Furthermore, he explicitly incorporates a liquidity premium to account for the high variability in the spreads on US corporate debt. In [32] and [15] also models are proposed that account for the dependence of recoveries on systematic risk by extending Gordy’s single-factor model.

ⁿSee [2]. Note, however, that this study finds that a single systematic risk factor—that is, the performance of the economy as a whole—is less predictive than theoretical models would suggest, while a key role is played by the supply of defaulted bonds.

^oSee also the empirical results in [15].

^pFor a more flexible approach, see [24] where a variation of the Gaussian kernel, known as *Beta kernel*, is used to fit the distribution of RRs of a sample of defaulted bonds from the period 1981–1999. See also [18], for an interesting utility-based approach to the estimation of the conditional probability distribution of RRs.

References

- [1] Acharya, V., Bharath, S. & Srinivasan, A. (2007). Does industry-wide distress affect defaulted firms: evidence from creditor recoveries, *Journal of Financial Economics*, **85**, 787–821.
- [2] Altman, E.I., Brady, B., Resti, A. & Sironi, A. (2005). The link between default and recovery rates: theory, empirical evidence and implications, *Journal of Business* **78**(6), 2203–2228.
- [3] Altman, E.I. & Eberhart, A. (1994). Do seniority provisions protect bondholders’ investments? *Journal of Portfolio Management* (Summer), 67–75.
- [4] Altman, E.I. & Fanjul, G. (2004). Defaults and returns in the high-yield bond market: the year 2003 in review and market outlook, in *Credit Risk—Models and Management*, D. Shimko, ed, RiskBooks, London.
- [5] Altman, E.I. & Kishore, V.M. (1996). Almost everything you wanted to know about recoveries on defaulted bonds, *Financial Analysts Journal* **52**(6), 57–64.
- [6] Altman, E.R.H. & Narayanan, P. (1977). ZETA analysis: a new model to identify bankruptcy risk of corporations, *Journal of Banking & Finance* **1**(1), 29–54.
- [7] Altman, E.I., Resti, A. & Sironi, A. (2005). *Recovery Risk—The Next Challenge in Credit Risk Management*, Risk Books, London.
- [8] Altman, E., Resti, A. & Sironi, A. (2005). The PD/LGD link: implications for credit risk modelling, in *Recovery*

- Risk—The Next Challenge in Credit Risk Management*, E. Altman, A. Resti & A. Sironi, eds, RiskBooks, London, pp. 253–266.
- [9] Araten, M. & Jacobs, M.J. (2001). Loan equivalents for revolving credit and advised lines, *The RMA Journal* **83**(8), 34–39.
- [10] Asarnow, E. & Edwards, D. (1995). Measuring loss on defaulted bank loans: a 24 year study, *Journal of Commercial Bank Lending* **77**(7), 11–23.
- [11] Asarnow, E. & Marker, J. (1995). Historical performance of the US corporate loan market: 1988–1993, *Journal of Commercial Bank Lending* (Spring), 13–32.
- [12] Basel Committee on Banking Supervision (2006). *International Convergence of Capital Measurement and Capital Standards—A Revised Framework—Comprehensive Version*, Bank for International Settlements, Basel.
- [13] Das, S.R. & Hanouna, P.E. (2007). *Implied Recovery*, Tratto da SSRN, <http://ssrn.com/abstract=1028612>.
- [14] Dermine, J. & Neto de Carvalho, C. (2005). How to measure recoveries and provisions on bank lending: methodology and empirical evidence, in *Recovery Risk—The Next Challenge in Credit Risk Management*, E. Altman, A. Resti & A. Sironi, eds, RiskBooks, London, pp. 101–120.
- [15] Duellmann, K. & Trapp, M. (2005). Systematic risk in recovery rates of US corporate credit exposures, in *Recovery Risk—The next Challenge in Credit Risk Management*, E. Altman, A. Resti & A. Sironi, eds, RiskBooks, London, pp. 235–252.
- [16] Emery, K. (2003). *Moody's Loan Default Database as of November 2003*, Moody's Investors Service, New York.
- [17] Finger, C. (2001). The one-factor creditmetrics model in the new Basel capital accord, *RiskMetrics Journal* **2**(1), 9–18.
- [18] Friedman, C. & Sandow, S. (2003). Ultimate recoveries, *Risk* August, 69–73.
- [19] Frye, J. (2000). Collateral damage, *Risk* (April), 91–94.
- [20] Frye, J. (2000). *Collateral Damage Detected*, Federal Reserve Bank of Chicago, Chicago.
- [21] Frye, J. (2000). Depressing recoveries, *Risk* (November), 108–111.
- [22] Gordy, M.B. (2003). A risk-factor model foundation for ratings-based bank capital rules, *Journal of Financial Intermediation* **12**, 199–232.
- [23] Gupton, G.M. & Stein, R.M. (2002). *LossCalc: Moody's Model for Predicting Loss Given Default (LGD)*, Moody's Investors Service, New York.
- [24] Hagmann, M., Renault, O. & Scaillet, O. (2005). Estimation of recovery rate densities: non-parametric and semi-parametric approaches versus industry practice, in *Recovery Risk: the Next Challenge in Credit Risk Management*, E. Altman, A. Resti & A. Sironi, eds, Risk Books, London.
- [25] Jacobs, M. (2007). *An Empirical Study of Exposure at Default*, mimeo Office of the Comptroller of the Currency, Washington, DC.
- [26] Jarrow, R. (2001). Default parameter estimation using market prices, *Financial Analysts Journal* **57**(5), 75–92.
- [27] Jiménez, G., Lopez, J.A. & Saurina, J. (2007). *Empirical Analysis of Corporate Credit Lines*, San Francisco: working paper, 2007–14, Federal Reserve Bank of San Francisco, San Francisco.
- [28] Keisman, D. (2003). *Loss Stats*, Standard & Poor's, New York.
- [29] Maclachlan, I. (2005). Choosing the discount factor for estimating economic LGD, in *Recovery Risk—The Next Challenge in Credit Risk Management*, E. Altman, A. Resti & A. Sironi, eds, RiskBooks, London, pp. 285–306.
- [30] Moral, G. (2006). EAD estimates for facilities with explicit limits, in *The Basel II Risk Parameters: Estimation, Validation and Stress Testing*, E. Bernd & R. Robert, eds, Springer Verlag, Berlin.
- [31] Peura, S. & Jovivuolle, E. (2005). LGD in a structural model of default, in *Recovery Risk—The Next Challenge in Credit Risk Management*, E.I. Altman, A. Resti & A. Sironi, eds, RiskBooks, London, pp. 201–216.
- [32] Pykhtin, M. (2003). Unexpected recovery risk, *Risk* **16**(8), 74–78.
- [33] Schuermann, T. (2005). What do we know about Loss Given Default? in *Recovery Risk—The Next Challenge in Credit Risk Management*, A. Resti & E.I. Altman, eds, Risk Books, London.
- [34] Van de Castle, K. & Keisman, D. (1999). *Recovering Your Money: Insights Into Losses from Defaults*, Standard & Poor's, New York.

Further Reading

Schleifer, A. & Vishny, R. (1992). Liquidation values and debt capacity: a market equilibrium approach, *Journal of Finance* **47**, 1343–1366.

Related Articles

Counterparty Credit Risk; Recovery Rate; Value-at-Risk.

ANDREA RESTI

Credit Portfolio Simulation^a

Portfolio Modeling

In risk management, quantitative techniques are mainly used for measuring the risk in a portfolio of assets rather than computing the prices of individual securities. The quantification of portfolio risk is traditionally split into separate calculations for market and credit risk, which are performed in different types of portfolio models. This article focuses on credit risk, more precisely on simulation techniques in structural credit portfolio models. We refer to [4] for a comprehensive exposition of Monte Carlo (MC) methods in quantitative finance including applications in market risk models.

In a typical bank, risk capital for credit risk far outweighs capital requirements for any other risk class. Key drivers of credit risk are concentrations in a bank's credit portfolio. Depending on their formulation, credit portfolio models can be divided into reduced-form models and structural (or firm-value) models (*see Reduced Form Credit Risk Models; Structural Default Risk Models*). The progenitor of all structural models is the model of Merton [13], which links the default of a firm to the relationship between its assets and the liabilities that it faces at the end of a given time period $[0, T]$. More precisely, in a structural credit portfolio model, the i th counterparty defaults if its ability-to-pay variable A_i falls below a default threshold D_i : the default event at time T is defined as $\{A_i \leq D_i\} \subseteq \Omega$, where A_i is a real-valued random variable on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $D_i \in \mathbb{R}$. The portfolio loss variable is defined by

$$L := \sum_{i=1}^n l_i \cdot \mathbf{1}_{\{A_i \leq D_i\}} \quad (1)$$

where n denotes the number of counterparties and l_i is the loss-at-default of the i th counterparty. To reflect risk concentrations, each A_i is decomposed into a sum of systematic factors $X_1, \dots, X_m$, which are often identified with geographic regions or industries, and an idiosyncratic (or firm-specific) factor Z_i , that

is A_i has the representation

$$A_i = \sqrt{R_i^2} \sum_{j=1}^m w_{ij} X_j + \sqrt{1 - R_i^2} Z_i \quad (2)$$

The idiosyncratic factors are independent of each other as well as independent of the systematic factors. It is usually assumed that the factors follow a multivariate Gaussian distribution. We refer to this class of models as *Gaussian multifactor models*.^b The impact of the risk factors on A_i is determined by $R_i^2 \in [0, 1]$ and the factor weights $w_{ij} \in \mathbb{R}$.

To quantify portfolio risk, measures of risk are applied to the portfolio loss distribution (1). The most widely used risk measures in banking are Value-at-Risk and expected shortfall: Value-at-Risk $\text{VaR}_\alpha(L)$ of L at level $\alpha \in (0, 1)$ is simply an α -quantile of L , whereas expected shortfall of L at level α is defined by

$$\text{ES}_\alpha(L) := (1 - \alpha)^{-1} \int_\alpha^1 \text{VaR}_u(L) du$$

For most practical applications, the average of all losses above the α -quantile is a good approximation of $\text{ES}_\alpha(L)$: for $c := \text{VaR}_\alpha(L)$ we have

$$\text{ES}_\alpha(L) \approx \mathbb{E}(L | L > c) = (1 - \alpha)^{-1} \int L \cdot \mathbf{1}_{\{L > c\}} d\mathbb{P} \quad (3)$$

This approximation is an exact equality unless the distribution of L has an atom at c , a situation that very rarely arises in practice.

Simulation Techniques

Since the portfolio loss distribution (1) does not have analytic form, the actual calculation and allocation of portfolio risk is a challenging problem. Saddle-point techniques have been successfully applied to certain types of portfolios; see, for example, [10] or *see Saddlepoint Approximation*. The most flexible approach, however, is based on MC simulation of the portfolio loss distribution. The following are the main steps in generating one MC sample:

1. calculation of a sample $(x_1, \dots, x_m)$ of the correlated systematic factors and a sample $(z_1, \dots, z_n)$ of the independent idiosyncratic factors;

2. calculation of the corresponding values $(a_1, \dots, a_n)$ of the ability-to-pay variables using equation (2);
3. calculation of the set of defaulted counterparties defined by $Def := \{i \in \{1, \dots, n\} \mid a_i \leq D_i\}$;
4. calculation of the portfolio loss: the sum $\sum_{i \in Def} l_i$ is a sample of the portfolio loss distribution.

The MC scenarios of the portfolio loss distribution are used as input for the calculation of risk measures. As an example, we compute expected shortfall with respect to the $\alpha = 99.9\%$ quantile based on $k = 100\,000$ MC samples $s_1 \geq s_2 \geq \dots \geq s_k$ of the portfolio loss L . Then $ES_\alpha(L)$ becomes

$$(1 - \alpha)^{-1} \int L \cdot \mathbf{1}_{\{L > c\}} d\mathbb{P} \approx \sum_{i=1}^{100} s_i / 100 \quad (4)$$

Since $ES_\alpha(L)$ is calculated as the average of 100 samples only, the MC estimate is subject to large statistical fluctuations and is numerically unstable. This is even truer for expected shortfall contributions of individual transactions. A significantly higher number of samples has to be computed, which makes straightforward MC simulation impracticable for large credit portfolios.

Different techniques have been developed that reduce the variance of MC simulations and—as a consequence—the number of samples required for stable results. We refer to [4] for a general introduction to variance reduction techniques including control variates, antithetic variables, stratified sampling, moment matching, and importance sampling. Recent research [3, 5–9, 12, 14] has shown that importance sampling is particularly efficient for stabilizing MC simulation in Gaussian multifactor models. Importance sampling attempts to reduce variance by changing the probability measure used for generating MC samples. In the above setting, the integral in equation (3) is replaced by the equivalent integral on the right-hand side of the equation

$$\int L \cdot \mathbf{1}_{\{L > c\}} d\mathbb{P} = \int L \cdot \mathbf{1}_{\{L > c\}} \cdot f d\tilde{\mathbb{P}} \quad (5)$$

where $\mathbb{P}$ is absolutely continuous with respect to the probability measure $\tilde{\mathbb{P}}$ and has (Radon–Nikodym) density f . This change of measure results in the MC

estimate

$$ES_\alpha(L)_{k, \tilde{\mathbb{P}}} := \frac{1}{k} \sum_{i=1}^k L_{\tilde{\mathbb{P}}}(i) \cdot \mathbf{1}_{\{L_{\tilde{\mathbb{P}}}(i) > c\}} \cdot f(i) \quad (6)$$

where $L_{\tilde{\mathbb{P}}}(i)$ is a realization of the portfolio loss L under the probability measure $\tilde{\mathbb{P}}$ and $f(i)$ is the corresponding value of the density function. The objective is to choose the probability measure $\tilde{\mathbb{P}}$ in such a way that the variance of the MC estimate for the integral (5) is minimal under $\tilde{\mathbb{P}}$. A general formula for the optimal importance sampling measure $\tilde{\mathbb{P}}$ is given in [15], which transforms equation (6) into a zero-variance estimator. However, since the construction requires knowledge of the integral (3) itself, the optimal measure cannot be used in the actual calculation. Nevertheless, it provides guidance on the design of an effective importance sampling strategy. Another technique for measure transformation, called *exponential tilting*, applies exponential families of distributions, which are specified by cumulant generating functions [1, 4]. As a general rule, detailed knowledge about the model (often in the form of asymptotic approximations) is indispensable for the construction of importance sampling algorithms. It is precisely this feature of importance sampling that makes the practical application more difficult but, on the other hand, increases the effectiveness of the methodology.

Importance sampling in Gaussian multifactor models utilizes the conditional independence of ability-to-pay variables by splitting the simulation of the portfolio loss distribution into two steps (compare to [11] in the more general context of mixture models). In a first step, importance sampling is used to simulate the systematic factors, and then the independence of the ability-to-pay variables conditional on systematic scenarios is exploited, for example, by another application of importance sampling or by limit theorems [7, 8].

A natural importance sampling measure $\tilde{\mathbb{P}}$ for the systematic factors is a negative shift, that is, the systematic factors have a negative mean under $\tilde{\mathbb{P}}$, which enforces a higher number of defaults and therefore increases the stability of the MC estimate. For calculating the shift, Glasserman and Li [7] minimize an upper bound on the second moment of the importance sampling estimator of the tail probability. Furthermore, they show that the corresponding importance sampling scheme is asymptotically optimal. The approach in [8, 9] utilizes the infinite

granularity approximation of the portfolio loss distribution (compare to [16]). More precisely, the original portfolio P is approximated by a homogeneous and infinitely granular portfolio $\bar{P}$. The loss distribution of $\bar{P}$ can be specified by a Gaussian one-factor model. The calculation of the shift of the systematic factors is now done in two steps: in the first step, the optimal mean is calculated in the one-factor setting and then the scalar mean is lifted to a mean vector for the systematic factors in the original multifactor model. Other importance sampling techniques [3, 6] are based on the Robbins–Monro stochastic approximation method or use large deviation analysis to calculate multiple mean shifts.

The efficiency of the proposed variance reduction schemes heavily depends on the portfolio characteristics. For example, the technique proposed in [8, 9] is tailored to large and well-diversified portfolios. For those portfolios the analytic loss distribution of the infinitely granular portfolio provides an excellent fit, which typically reduces the variance—and therefore the number of required MC scenarios—by a factor of more than 100. Smaller portfolios with low dependence on systematic factors, on the other hand, are dominated by idiosyncratic risk, which increases the relative importance of variance reduction techniques on idiosyncratic factors [7, 8], for example, importance sampling based on exponential tilting.

End Notes

^aThe views expressed in this article are those of the author and do not necessarily reflect the position of Deutsche Bank AG.

^bA survey on credit portfolio modeling can be found in [2, 11].

References

- [1] Barndorff-Nielsen, O. (1978). *Information and Exponential Families*, Wiley.
- [2] Bluhm, C., Overbeck, L. & Wagner, C. (2002). *An Introduction to Credit Risk Modeling*, CRC Press/Chapman & Hall.
- [3] Egloff, D., Leippold, M., Jöhri, S. & Dalbert, C. (2005). *Optimal Importance Sampling for Credit Portfolios with Stochastic Approximations*. Working paper, Zürcher Kantonalbank, Zurich.
- [4] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*, Springer.
- [5] Glasserman, P. (2005). Measuring marginal risk contributions in credit portfolios, *Journal of Computational Finance* **9**, 1–41.
- [6] Glasserman, P., Kang, W. & Shahabuddin, P. (2007). *Fast Simulation of Multifactor Portfolio Credit Risk*. Working paper, Columbia University, New York.
- [7] Glasserman, P. & Li, J. (2005). Importance sampling for portfolio credit risk, *Management Science* **51**, 1643–1656.
- [8] Kalkbrener, M., Kennedy, A. & Popp, M. (2007). Efficient calculation of expected shortfall contributions in large credit portfolios, *Journal of Computational Finance* **11**, 45–77.
- [9] Kalkbrener, M., Lotter, H. & Overbeck, L. (2004). Sensible and efficient capital allocation for credit portfolios, *Risk* **17**(1), S19–S24.
- [10] Martin, R., Thompson, K. & Browne, C. (2001). Taking to the saddle, *Risk* **14**(6), 91–94.
- [11] McNeil, A.J., Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools*, Princeton University Press.
- [12] Merino, S. & Nyfeler, M. (2004). Applying importance sampling for estimating coherent credit risk contributions, *Quantitative Finance* **4**, 199–207.
- [13] Merton, R. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.
- [14] Morokoff, W.J. (2004). An importance sampling method for portfolios of credit risky assets, *Proceedings of the 2004 Winter Simulation Conference*, IEEE Press, pp. 1668–1676.
- [15] Rubinstein, R.Y. (1981). *Simulation and the Monte Carlo Method*, Wiley.
- [16] Vasicek, O. (2002). Loan portfolio value, *Risk* **15**(12), 160–162.

Related Articles

Large Pool Approximations; Monte Carlo Simulation; Structural Default Risk Models; Saddlepoint Approximation; Variance Reduction.

MICHAEL KALKBRENER

Counterparty Credit Risk^a

Counterparty credit risk (CCR) is the risk that a counterparty in a financial contract will default prior to the expiration of the contract and will fail to make all the payments required by the contract. Only the contracts privately negotiated between the counterparties—over-the-counter (OTC) derivatives and securities financing transactions (SFT)—bear CCR. Exchange-traded derivatives are not subject to CCR because all contractual payments promised by these derivatives are guaranteed by the exchange.

CCR is similar to other forms of credit risk (such as lending risk) in that the source of economic loss is an obligor's default. However, CCR has two unique features that set it apart from lending risk:

- **Uncertainty of credit exposure** Credit exposure of one counterparty to the other is determined by the market value of all the contracts between these counterparties. While one can obtain the current exposure from the current contract values, the future exposure is uncertain because the future contract values are not known at present.
- **Bilateral nature of credit exposure** Since both counterparties can default and the value of many financial contracts (such as swaps) can change sign, the direction of future credit exposure is uncertain. Counterparty A may be exposed to default of counterparty B under one set of future market scenarios, while counterparty B may be exposed to default of counterparty A under another set of scenarios.

The uncertainty of future credit exposure makes managing and modeling CCR of the trading book challenging. For a comprehensive introduction to CCR, see [1, 5, 17].

Managing and Mitigating Counterparty Credit Risk

One of the most conventional techniques of managing credit risk is setting counterparty-level *credit limits*. If a new transaction with the counterparty would result in the counterparty-level exposure exceeding

the limit, the transaction is not allowed. The limits usually depend on the counterparty's credit quality: higher rated counterparties have higher limits. To compare uncertain future exposure with a deterministic limit, *potential future exposure* (PFE) profiles are calculated from exposure probability distributions at future time points. PFE profiles are obtained by calculating a quantile of exposure at a high confidence level (typically, above 90%). Some institutions use different exposure measures, such as *expected exposure* (EE) profiles, for comparing with the credit limit. It is important to understand that a given credit limit amount is meaningful only in the context of a given exposure measure (e.g., 95%-level quantile).

Future credit exposure can be greatly reduced by means of risk-mitigating agreements between two counterparties, which include *netting agreements*, *margin agreements*, and *early termination agreements*. Netting agreement is a legally binding contract between two counterparties that, in the event of default of one of them, allows aggregation of transactions between these counterparties. Instead of each trade between the counterparties being settled separately, the entire portfolio covered by the netting agreement is settled as a single trade whose value equals the net value of the portfolio. Margin agreements limit the potential exposure of one counterparty to the other by means of requiring collateral should the unsecured exposure exceed a predefined threshold. The threshold value depends primarily on the credit quality of the counterparty: the higher the credit quality, the higher the threshold.

There are two types of early termination agreements: *termination clauses* and *downgrade provisions*. Termination clause is specified at the trade level. A unilateral (bilateral) termination clause gives one (both) of the counterparties the right to terminate the trade at the fair market value at a predefined set of dates. Downgrade provision is specified for the entire portfolio between two counterparties. Under a unilateral (bilateral) downgrade provision, the portfolio is settled at its fair market value the first time the credit rating of one (either) of the counterparties falls below a predefined level.

Contract-level Exposure

Let us consider a financial institution (we will call it a *bank* for brevity) that has a single derivative

contract with a counterparty. The bank's *exposure* to the counterparty at a given future time is given by the bank's economic loss in the event of the counterparty's default at that time. If the counterparty defaults, the bank must close out its position with the counterparty. To determine the loss arising from the counterparty's default, it is convenient to assume that the bank enters into a similar contract with another counterparty in order to maintain its market position. Since the bank's market position is unchanged after replacing the contract, the loss is determined by the contract's *replacement cost* at the time of default.

If the trade value at the time of default is negative for the bank, the bank receives this amount when it replaces the trade, but has to forward the money to the defaulting counterparty, so that the net loss is zero. If the trade value at the time of default is positive for the bank, the bank pays this amount when replacing the trade, but receives nothing (assuming no recovery) from the defaulting counterparty, so that the net loss is equal to the trade value.

Summarizing this, we can write the bank's credit exposure to the counterparty at future time t as

$$E_i(t) = \max\{V_i(t), 0\} \quad (1)$$

where $V_i(t)$ is the value of trade i with the counterparty at time t from the bank's point of view and $E_i(t)$ the bank's contract-level exposure to the counterparty created by trade i at time t .

Since the contract value changes unpredictably over time as the market moves, only the current exposure is known with certainty, while the future exposure is uncertain.

Counterparty-level Exposure and Netting Agreements

If the bank has more than one trade with the counterparty and counterparty risk is not mitigated in any way, the bank's exposure to the counterparty is equal to the sum of the contract-level exposures:

$$E_c(t) = \sum_i E_i(t) = \sum_i \max\{V_i(t), 0\} \quad (2)$$

where the subscript 'c' stands for 'counterparty'.

Netting agreements allow for a significant reduction of the credit exposure. There may be several netting agreements between the bank and the counterparty, as well as some trades that are not covered

by any of the netting agreements. Counterparty-level exposure in this most general case is given by

$$E_c(t) = \sum_k \max \left[\sum_{i \in \text{NA}_k} V_i(t), 0 \right] + \sum_{i \notin \{\text{NA}\}} \max[V_i(t), 0] \quad (3)$$

The inner summation in the first term of equation (3) aggregates the values of all trades covered by the k th netting agreement (hence the notation $i \in \text{NA}_k$), while the outer summation aggregates exposures across all netting agreements. The second term in equation (3) is simply the sum of the contract-level exposures of all trades that do not belong to any netting agreement (hence the notation $i \notin \{\text{NA}\}$).

Margin Agreements and Collateral Modeling

Margin agreements can further reduce credit exposure. Margin agreements can be either *unilateral* or *bilateral*. Under a unilateral agreement, only one of the counterparties has to post collateral. If the agreement is bilateral, both counterparties have to post collateral.

Usually a margin agreement covers one or more netting agreements. We can generalize equation (3) by specifying collateral amount $C_k(t)$ available to the bank under netting agreements NA_k at time t with the convention that this amount is positive when the bank holds collateral and negative when the bank has posted collateral:

$$E_c(t) = \sum_k \max \left[\sum_{i \in \text{NA}_k} V_i(t) - C_k(t), 0 \right] + \sum_{i \notin \{\text{NA}\}} \max[V_i(t), 0] \quad (4)$$

For netting agreements that are not covered by a margin agreement, collateral is identically zero.

Unilateral Margin Agreements

Let us consider a single unilateral (in the bank's favor) margin agreement with the threshold $H_c \geq 0$

and minimum transfer amount (MTA). When the portfolio value exceeds the threshold, the counterparty must post collateral to keep the bank's exposure from rising above the threshold. As the exposure drops below the threshold, the bank returns collateral to the counterparty. MTA limits the frequency of collateral exchange. It is difficult to model collateral subject to MTA exactly because that would require daily simulation time points, which is not feasible given the long-term nature of exposure modeling. In practice, the actual threshold H_c is often replaced by the effective threshold defined as $H_c^{(e)} = H_c + \text{MTA}$. After this replacement, the margin agreement is treated as if it had zero MTA.

The simplest approach to modeling collateral is to limit the future exposure from above by the threshold (i.e., for all scenarios with portfolio value above the threshold, set the exposure equal to the threshold). However, this approach is too simplistic because it ignores the time lag between the last delivery of collateral and the time when the loss is realized. This time lag is known as the *margin period of risk (MPR)*, which we will denote by δt . While the MPR is not known with certainty, it is typically assumed to be a deterministic number that is defined at the margin agreement level. Its value depends on the contractual margin call frequency and the liquidity of the portfolio. For example, $\delta t = 2$ weeks is usually assumed for portfolios of liquid contracts and daily margin call frequency.

Applying the rules of posting collateral under the assumption of the effective threshold with zero MTA and taking into account the MPR, the collateral $C(t)$ available to the bank at time t is given by

$$C(t) = \max\{V(t - \delta t) - H_c^{(e)}, 0\} \quad (5)$$

where $V(t)$ is the portfolio value from the bank's point of view at time t .

Bilateral Margin Agreements

Under a bilateral margin agreement, both the counterparty and the bank have to post collateral: the counterparty posts collateral when the bank's exposure to the counterparty exceeds the counterparty's threshold, while the bank posts collateral when the counterparty's exposure to the bank exceeds the bank's threshold. Since we are doing our analysis from the point of view of the bank, we will keep

the counterparty's threshold H_c *nonnegative*, but will specify the bank's threshold H_b as *nonpositive*. Then, the bank posts collateral when the portfolio value (defined from the bank's point of view) is *below* the bank's threshold. The MTA is the same for the bank and the counterparty and $\text{MTA} > 0$.

Similar to the unilateral case, we will create effective thresholds for the bank and for the counterparty. Effective threshold for the counterparty, $H_c^{(e)}$, remains unchanged. From the counterparty's point of view, effective threshold for the bank must be defined in exactly the same way. After taking into account that we do not switch our point of view and $H_b \leq 0$, the definition of the effective threshold for the bank will be $H_b^{(e)} = H_b - \text{MTA}$. Now the bilateral agreement can be treated as if it had zero MTA.

Collateral available to the bank at time t under the bilateral agreement is modeled as

$$C(t) = \max\{V(t - \delta t) - H_c^{(e)}, 0\} + \min\{V(t - \delta t) - H_b^{(e)}, 0\} \quad (6)$$

The first term in the right-hand side of equation (6) describes the scenarios when the bank receives collateral (i.e., $C(t) > 0$), while the second term describes the scenarios when the bank posts collateral (i.e., $C(t) < 0$).

For more details on collateral modeling, see [9, 15, 16].

Simulating Credit Exposure

Because of the complex nature of banks' portfolios, exposure distribution at future time points is usually obtained *via Monte Carlo Simulation* process. This process typically consists of three major steps:

- **Scenario generation** Dynamics of market risk factors (e.g., interest rates, foreign exchange (FX) rates, etc.) is specified *via* relatively simple stochastic processes (e.g., geometric Brownian motion). These processes are calibrated either to historical data or to market implied data. Future values of the market risk factors are simulated for a fixed set of future time points.
- **Instrument valuation** For each simulation time point and for each realization of the underlying market risk factors, valuation is performed for each trade in the counterparty portfolio.

- **Aggregation** For each simulation time point and for each realization of the underlying market risk factors, counterparty-level exposure is obtained by applying the necessary netting and collateral rules, conceptually described by equations (3) and (4).

The outcome of this process is a set of realizations of the counterparty-level exposure (each realization corresponds to one market scenario) at each simulation time point.

Because of the computational intensity required to calculate counterparty exposures—especially for a bank with a large portfolio—certain compromises between the accuracy and the speed of the calculation are usually made: relatively small number of market scenarios (typically, a few thousand) and simulation time points (typically, in the 50–200 range), simplified valuation methods, and so on.

For more details on simulating credit exposure, see [7, 17].

Pricing Counterparty Risk—Unilateral Approach

Let us assume that the bank is default-risk-free. Then, when pricing transactions with a counterparty, the bank should require a risk premium to be compensated for the risk of the counterparty defaulting. The market value of this risk premium, defined for the entire portfolio of trades with the counterparty, is known as unilateral *credit valuation adjustment* (CVA).

A **Risk-neutral Pricing** valuation framework is used for pricing CCR. The bank's economic loss arising from the counterparty's default and discounted to today is given by

$$L_{u-l} = 1_{\{\tau_c \leq T\}} (1 - R_c) E_c(\tau_c) \frac{B_0}{B_{\tau_c}} \quad (7)$$

where τ_c is the time of default of the counterparty; $1_{\{A\}}$ is the indicator function that assumes the value 1 if Boolean variable A is TRUE and value 0, otherwise; $E_c(t)$ is the bank's exposure to counterparty's default at time t ; R_c is the counterparty **Recovery Rate** (i.e., percentage of the bank's exposure to the counterparty that the bank will be able to recover in the event of the counterparty's default); and B_t is the value of the money-market account at time t .

One should keep in mind that counterparty-level exposure $E_c(t)$ incorporates all netting and margin agreements between the bank and the counterparty, as discussed above.

Unilateral CVA is obtained by taking the risk-neutral expectation of the loss in equation (7). Under the assumption that recovery rate is independent of the market factors and the time of default, this results in

$$\text{CVA}_{u-l} = (1 - \bar{R}_c) \int_0^T EE_c^*(t) dPD_c(t) \quad (8)$$

where $EE_c^*(t)$ is the risk-neutral discounted EE at time t , conditional on the counterparty defaulting at time t , given by

$$EE_c^*(t) = E^Q \left[\frac{B_0}{B_t} E_c(t) \mid \tau_c = t \right] \quad (9)$$

and $\bar{R}_c$ is the expected recovery rate; $PD_c(t)$ is the counterparty's cumulative from today to time t , estimated today; and T is the maturity of the longest trade in the portfolio.

The term structure of the risk-neutral PDs is obtained from the **Credit Default Swaps** spreads quoted in the market [19].

We would like to emphasize that the expectation of the discounted exposure at time t in equation (9) is conditional on the counterparty's default occurring at time t . This conditioning is material when there is a significant dependence between the exposure and the counterparty credit quality. This dependence, known as *right/wrong-way risk*, was first considered in [8] and [12]. To account for it, the counterparty's credit quality must be modeled jointly with the market risk factors. For more details on modeling right/wrong-way risk, see [4, 10, 18].

In practice, the dependence between exposure and the counterparty's credit quality is often ignored and conditioning on default in equation (9) is removed. Discounted EE is calculated for a set of simulation time points $\{t_k\}$ under the exposure simulation framework outlined above. Then, CVA is calculated by approximating the integral in equation (8) by a sum:

$$\begin{aligned} \text{CVA}_{u-l} \approx (1 - \bar{R}_c) \sum_k EE_c^*(t_k) \\ \times [PD_c(t_{k-1}) - PD_c(t_k)] \end{aligned} \quad (10)$$

Since the exposure expectation in equation (10) is risk neutral, scenario models for all market risk factors should be arbitrage free. This is achieved by appropriate calibration of drifts. Moreover, risk factor volatilities should be calibrated to the available market prices of options on the risk factors.

For more details on unilateral CVA, see [3].

Pricing Counterparty Risk—Bilateral Approach

In reality, banks are not default-risk-free. Because of the bilateral nature of credit exposure, the bank and the counterparty will never agree on the fair price of CCR if they apply unilateral pricing outlined above: each of them will demand a risk premium from the other. The bilateral approach specifies a single quantity—known as bilateral CVA—that accounts both for the bank's loss caused by the counterparty's default and the counterparty's loss caused by the bank's default.

Bilateral loss of the bank is given by

$$L_{b-1} = 1_{\{\tau_c \leq T\}} 1_{\{\tau_c < \tau_b\}} (1 - R_c) E_c(\tau_c) \frac{B_0}{B_{\tau_c}} - 1_{\{\tau_b \leq T\}} 1_{\{\tau_b < \tau_c\}} (1 - R_b) E_b(\tau_b) \frac{B_0}{B_{\tau_b}} \quad (11)$$

where τ_b is the time of default of the bank; $E_b(t)$ is the counterparty's exposure to bank's default at time t ; and R_b is the bank recovery rate (i.e., percentage of the counterparty's exposure to the bank that the counterparty will be able to recover in the event of the bank's default).

The first term in equation (11) describes the bank's loss when the counterparty defaults, but the bank does not default. The second term describes the loss of the counterparty in the event of the bank's default and the counterparty's survival. From the bank's point of view, the counterparty's loss is gain arising from the bank's option not to pay the counterparty when the bank defaults, so this term is subtracted from the bank's loss. Equation (11) is completely symmetric: if we change the sign of the right-hand side, we will obtain the bilateral loss of the counterparty.

Bilateral CVA is obtained by taking risk-neutral expectation of equation (11):

$$\begin{aligned} \text{CVA}_{b-1} = & (1 - \bar{R}_c) \int_0^T \mathbb{E} E_c^*(t) \\ & \times \Pr[\tau_b > t | \tau_c = t] \text{dPD}_c(t) \\ & - (1 - \bar{R}_b) \int_0^T \mathbb{E} E_b^*(t) \\ & \times \Pr[\tau_c > t | \tau_b = t] \text{dPD}_b(t) \end{aligned} \quad (12)$$

where $\mathbb{E} E_c^*(t)$ is the discounted EE of the counterparty to the bank at time t , conditional on the counterparty defaulting at time t , defined in equation (9), and $\mathbb{E} E_b^*(t)$ is the discounted EE of the bank to the counterparty at time t , conditional on the bank defaulting at time t , defined as

$$\mathbb{E} E_b^*(t) = \mathbb{E} \left[E_b(t) \frac{B_0}{B_t} \middle| \tau_b = t \right] \quad (13)$$

If the dependence between credit exposure and the credit quality of the counterparty and of the bank can be ignored, the conditional expectations in equations (9) and (13) should be replaced with the unconditional ones. As expected, equation (12) is symmetric between the bank and the counterparty, so that the bank and the counterparty will always agree on the price of CCR for their portfolio.

One can use **Default Time Copulas; Gaussian Copula Model; Copulas: Estimation** to express the conditional probabilities in equation (12) as functions of the counterparty's and the bank's risk-neutral PDs. For example, if the normal copula model [13] is used to describe the dependence between τ_c and τ_b , the conditional probabilities in equation (12) take the form

$$\begin{aligned} \Pr[\tau_b > t | \tau_c = t] \\ = 1 - \Phi \left(\frac{\Phi^{-1}[\text{PD}_b(t)] - \rho \Phi^{-1}[\text{PD}_c(t)]}{\sqrt{1 - \rho^2}} \right) \end{aligned} \quad (14)$$

and

$$\begin{aligned} \Pr[\tau_c > t | \tau_b = t] \\ = 1 - \Phi \left(\frac{\Phi^{-1}[\text{PD}_c(t)] - \rho \Phi^{-1}[\text{PD}_b(t)]}{\sqrt{1 - \rho^2}} \right) \end{aligned} \quad (15)$$

where ρ is the normal copula correlation and $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Portfolio Loss and Economic Capital

Until now we have discussed modeling credit exposure and losses at the counterparty level. However, the distribution of the credit loss of the bank's entire trading book provides more complete information about the risk a bank is taking. Portfolio loss distribution is needed for such risk management tasks as calculation and allocation of **Economic Capital** (EC). For a comprehensive introduction to EC for CCR, see [14].

Portfolio credit loss $L(T)$ for a time horizon T can be expressed as the sum of the counterparty-level losses over all counterparties:

$$L(T) = \sum_j 1_{\{\tau_j \leq T\}} (1 - R^{(j)}) E^{(j)}(\tau_j) \frac{B_0}{B_{\tau_j}} \quad (16)$$

where τ_j is the time of default of counterparty j ; $R^{(j)}$ is the recovery rate for counterparty j ; and $E^{(j)}(t)$ is the bank's counterparty-level exposure at time t created by all trades that the bank has with counterparty j .

The economic capital $\text{EC}_q(T)$ for time horizon T and confidence level q is given by

$$\text{EC}_q(T) = Q_q[L(T)] - E[L(T)] \quad (17)$$

where $Q_q[X]$ is the quantile of random variable X at confidence level q (in risk management, this quantity is often referred to as **Value-at-Risk** (VaR)). The distribution of portfolio loss $L(T)$ can be obtained from equation (16) via joint Monte Carlo simulation of trade values for the entire bank portfolio and of default times of individual counterparties.

However, the joint simulation process is very expensive computationally and is often replaced by

a simplified approach, where simulation of counterparty defaults is completely separate from exposure simulation. The simplified approach is a two-stage process. During the first stage, exposure simulation is performed and a deterministic loan equivalent exposure (LEQ) is calculated from the exposure distribution for each counterparty. The second stage is a simulation of counterparty default events according to one of the credit risk portfolio models (see **Credit Migration Models**; **Structural Default Risk Models**; **CreditRisk+**) that are used for loan portfolios. Portfolio credit loss is calculated as

$$L(T) = \sum_j 1_{\{\tau_j \leq T\}} (1 - R^{(j)}) \text{LEQ}^{(j)}(T) \quad (18)$$

where $\text{LEQ}^{(j)}(T)$ is the LEQ of counterparty j for time horizon T .

Note that many loan portfolio models do not produce the time of default explicitly. Instead, they only distinguish between two events: "default has happened prior to the horizon (i.e., $\tau_j \leq T$)" and "default has not happened prior to the horizon (i.e., $\tau_j > T$)". Note also that, because the time of default is not known, discounting to present is not applied in equation (18).

For an infinitely fine-grained portfolio with independent exposures, it has been shown [6, 20] that LEQ is given by the EE averaged from zero to T —this quantity is often referred to as *expected positive exposure* (EPE):

$$\text{EPE}^{(j)}(T) \equiv \frac{1}{T} \int_0^T \text{EE}^{(j)}(t) dt \quad (19)$$

If one uses LEQ given by equation (19) for a real portfolio, the EC will be understated because both exposure volatility and correlation between exposures are ignored. However, this understated EC can be used in defining a scaling parameter commonly known as *alpha*:

$$\alpha_q(T) = \frac{\text{EC}_q^{(\text{Real})}(T)}{\text{EC}_q^{(\text{EPE})}(T)} \quad (20)$$

where $\text{EC}_q^{(\text{Real})}(T)$ is the EC of the real portfolio with stochastic exposures, and $\text{EC}_q^{(\text{EPE})}(T)$ is the EC of the fictitious portfolio with stochastic exposures replaced by EPE.

If α of a real portfolio can be estimated, its LEQ can be defined according to

$$\text{LEQ}^{(j)}(T) = \alpha_q(T) \text{EPE}^{(j)}(T) \quad (21)$$

Because the EC of a portfolio with deterministic exposures is a homogeneous function of the exposures, using the LEQ defined in equation (21) will produce the correct $\text{EC}_q^{(\text{Real})}(T)$. The caveat of this approach is that one has to run a joint simulation of trade values and counterparties' defaults to calculate α .

Several estimates of typical values of α for a large dealer portfolio and the time horizon $T = 1$ year are available. An International Swaps and Derivatives Association (ISDA) survey [11] has reported α calculated by four large banks for their actual portfolios to be in the 1.07–1.10 range. Theoretical estimates of α under a set of simplifying assumptions [6, 20] are 1.1 when market-credit correlations are ignored, and 1.2 when they are not.

The framework described above has found its place in the regulatory capital calculations under Basel II (see **Regulatory Capital**): a slightly modified version of equation (21) is used to calculate exposure at default (EAD) under the internal models method for CCR [2]. Basel fixes α at 1.4, but it allows banks to calculate their own α , subject to the supervisory approval and a floor of 1.2.

End Notes

^aThe opinions expressed here are those of the author and do not necessarily reflect the views or policies of the author's employer.

References

- [1] Arvanitis, A. & Gregory, J. (2001). *Credit: The Complete Guide to Pricing, Hedging and Risk Management*, Risk Books.
- [2] Basel Committee on Banking Supervision (2006). *International Convergence of Capital Measurement and Capital Standards*, A Revised Framework.
- [3] Brigo, D. & Masetti, M. (2005). Risk neutral pricing of counterparty credit risk, in *Counterparty Credit Risk Modelling*, M. Pykhtin, ed., Risk Books.
- [4] Brigo, D. & Pallavicini, A. (2008). Counterparty risk and contingent CDS under correlation, *Risk* February, 84–88.
- [5] Canabarro, E. & Duffie, D. (2003). Measuring and marking counterparty risk, in *Asset/Liability Management for Financial Institutions*, L. Tilman, ed., Institutional Investor Books.
- [6] Canabarro, E., Picoult, E. & Wilde, T. (2003). Analysing counterparty risk, *Risk* September, 117–122.
- [7] De Prisco, B. & Rosen, D. (2005). Modeling stochastic counterparty credit exposures for derivatives portfolios, in *Counterparty Credit Risk Modelling*, M. Pykhtin, ed., Risk Books.
- [8] Finger, C. (2000). Toward a better estimation of wrong-way credit exposure, *Journal of Risk Finance* 1(3), 43–51.
- [9] Gibson, M. (2005). Measuring counterparty credit exposure to a margined counterparty, in *Counterparty Credit Risk Modelling*, M. Pykhtin, ed., Risk Books.
- [10] Hille, C., Ring, J. & Shimamoto, H. (2005). Modelling counterparty credit exposure for credit default swaps, *Risk* May, 65–69.
- [11] ISDA-TBMA-LIBA (2003). *Counterparty Risk Treatment of OTC Derivatives and Securities Financing Transactions*, June.
- [12] Levin, R. & Levy, A. (1999). Wrong way exposure—are firms underestimating their credit risk? *Risk* July, 52–55.
- [13] Li, D. (2000). On default correlation: a copula approach, *Journal of Fixed Income* 9, 43–54.
- [14] Picoult, E. (2005). Calculating and hedging exposure, credit value adjustment and economic capital for counterparty credit risk, in *Counterparty Credit Risk Modelling*, M. Pykhtin, ed., Risk Books.
- [15] Pykhtin, M. (2009). Modeling credit exposure for collateralized counterparties, *Journal of Credit Risk*; to be published.
- [16] Pykhtin, M. & Zhu, S. (2006). Measuring counterparty credit risk for trading products under Basel II, in *Basel Handbook*, 2nd Edition, M. Ong, ed., Risk Books.
- [17] Pykhtin, M. & Zhu, S. (2007). A guide to modeling counterparty credit risk, *GARP Risk Review* July/August, 16–22.
- [18] Redon, C. (2006). Wrong way risk modelling, *Risk* April, 90–95.
- [19] Schonbucher, P. (2003). *Credit Derivatives Pricing Models*, Wiley.
- [20] Wilde, T. (2005). Analytic methods for portfolio counterparty risk, in *Counterparty Credit Risk Modelling*, M. Pykhtin, ed., Risk Books.

Related Articles

Default Time Copulas; Economic Capital; Exposure to Default and Loss Given Default; Monte Carlo Simulation; Risk-neutral Pricing.

MICHAEL PYKHTIN

Loan Valuation

A loan is an agreement in which one party, called a *lender*, provides the use of property, the *principal*, to another party, the *borrower*. The borrower customarily promises to return the principal after a specified period along with payment for its use, called *interest* [3]. When the property loaned is cash, the documentation of the agreement between borrower and lender is called a *promissory note*.

Although cash loans can take many forms, traditionally, banks and other financial institutions are the primary lenders of cash and businesses, organizations, and individuals are the borrowers. Most loans to corporations share a common set of structural characteristics [2, 5].

1. Interest on loans is typically paid quarterly at a rate specified relative to some reference rate such as LIBOR (i.e., $L + 250$ bp).^a Thus, loans have *floating-rate* coupons whose absolute values are not known with certainty except over the next quarter.
2. Often the firm's assets or receivables are pledged against the borrowed principal. Because of this, their recovery rates are generally higher than corporate bonds, which are most commonly unsecured.
3. Most loans are prepayable on any coupon date at par, although some agreements contain a prepayment penalty or have a noncall period. The loan prepayment feature ensures that loan prices rarely exceed several points above par.
4. Finally, unlike bonds which are public securities, loans are private credit agreements. Thus, access to firm fundamentals and loan terms may be limited and loan contracts are less standardized. It is not uncommon to find "nonstandard" covenants or other structural features catering to specific needs of borrowers or investors.

Loan valuation concerns the amount of interest that a lender requires for use of the property or an investor will charge for purchasing the loan agreement. That valuation depends on several factors, such as

1. the likelihood of failure to receive timely payments of principal, called *risk of default*;
2. the residual value of the loan in the event of default, called its *recovery value*;

3. the time by which the principal of the loan must be repaid, the *maturity*;
4. the current market rate of interest for the obligor's likelihood of default, called the *market credit spread*;
5. the likelihood of the event that a borrower will have repaid the principal at any particular date prior to maturity.

Although the bulk of the loans outstanding are rated investment-grade or better, these loans trade very infrequently because of their high credit quality and lack of price differentiation. In fact, most loans that trade after origination are those made by banks to borrowers having speculative-grade credit ratings. These loans, made to high-yield firms are typically referred to as *leveraged loans*, though the exact definition varies slightly among market participants.^b

The types of loan facilities commonly traded in secondary markets include the following:

1. **Amortizing term loans.** Usually called "term loan A", the periodic payments from these loans include partial payment of principal, similar to what a mortgage loan does. These loans are usually held by banks and are becoming less popular.
2. **Institutional term loans.** These loans are structured to have bullet or close-to-bullet payment schedules and are targeted for institutional investors. They are referred to as "*term loan B*", "*term loan C*" and so on. Institutional term loans constitute the bulk of leveraged loan market.
3. **Revolving credit lines.** These are unfunded or partially funded commitments by lenders that can be drawn at the discretion of the borrowers. The facility is analogous to a corporate credit card. It can be drawn and repaid multiple times during the term of the commitment. These commitments are traded in secondary market. They are also known as *revolvers*.
4. **Second-lien term loans.** They have cash-flow schedule similar to that of institutional term loans, except that their claims on borrowers' assets are behind first-lien loan holders in the event of default.
5. **Covenant-lite loans.** These are borrower-friendly versions of institutional term loans that have fewer than the typical stringent covenants

that restrict use of the principal or subsequent borrowing activities of the firm.

Loan Pricing

Like bonds, loans contain risk of default; an obligor may fail to make timely payments of interest and/or principal. Thus, the notion of a credit spread to LIBOR has been used to characterize the riskiness of loans, where the credit spread, s , to LIBOR is calculated as

$$V = \sum_{t=1}^{4n} \frac{c_t/4}{\left(1 + \frac{r_t + s}{4}\right)^t} + \frac{F}{\left(1 + \frac{r_{4n} + s}{4}\right)^{4n}} \quad (1)$$

where V is the market value of the loan, c_t is the coupon (LIBOR + contractual spread), r_t is the spot rate for maturity t LIBOR rates, and F is the face value of the loan to be repaid at maturity. Loan coupons are generally paid quarterly and then reset relative to LIBOR and this is reflected in equation 1. Using equation 1, we can calculate a credit spread for any loan whose market price is known.

One problem with equation 1 for loan valuation is that it fails to account for the fact that loans, unlike bonds, are typically prepayable at par on any given coupon date. The loan prepayment option creates uncertainty in the expected pattern of cash flows and complicates comparisons of value among loans based on their credit spreads. Pricing the prepayment option has proved difficult because of its dependence

on the evolution of an obligor's credit state and the changing market costs of borrowing. For example, if a firm's credit improves or the loan rate over LIBOR decreases, the likelihood of prepayment increases; the borrower can refinance at a lower rate. Conversely, if a borrower's credit deteriorates or lending rates increase, it will not be advantages for the borrower to refinance.

To account for the prepayment option, we price the loans using a credit-state-dependent backward induction method.^c To illustrate, consider pricing a term loan with face value F , intermediate floating-rate coupon payments of c_t , and a maturity at time T , to a borrower of known credit quality, J . Specifically, Figure 1 displays pricing lattices for a five-year loan to a double-B rated (i.e., $J = BB$) obligor having a coupon of LIBOR + 3%,^d and face value of 100 at maturity.^e Figure 1(a) shows how the obligor's credit state evolves over time. In the lattice, probabilities are assigned reflecting transitions from each node at time t to all nodes at $t + 1$. Thus, the probability of being at a given node will be conditional upon all the previous transitions. In practice, ratings transition probabilities are based on historical data from credit rating agencies,^f and these are typically modified by the current market price of risk to produce *risk-neutral* ratings transition matrices.^{g,h}

Having calculated transition probabilities between all future nodes, we then apply the backward induction method. At maturity T , the borrower pays the principal plus coupon, $F + c_T$, or the recovery

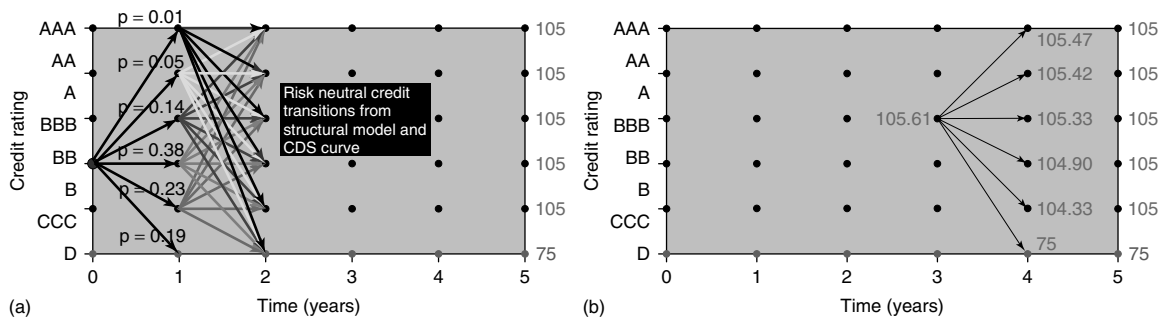


Figure 1 Credit-dependent backward induction method. (a) Double-B rated obligor, whose credit transitions are derived from historical data and incorporate market risk premiums are used to specify the likelihood of being in any credit state at future times to maturity. (b) Calculation of node values using backward induction, whereby values at each non-defaulted node are the coupon value at that node plus the sum of the conditional cash flows from the later date, discounted one period at forward LIBOR. In the example in (b), we assume a refinancing penalty of 0.5% of the principal

value in default $R * F$. Those cash flows are discounted back to each node at the previous period using forward LIBOR at $T - 1$. In other words, for each node at time $i < T$ and credit state j , and $j = (AAA, AA, \dots, CCC)$ we calculate an induced value, $v_{i,j}$, as

$$v_{(i,j)} = \min \left\{ \left(\frac{1}{\left(1 + \frac{f_{i+1,i}}{4}\right)} \sum_{k=D}^{AAA} (P_{j,k,i} * v_{i+1,k}) \right) + c_i, K_i \right\}, \quad (2)$$

where $P_{j,k,i}$ is the probability of migrating from state j to state k from time i to $i + 1$, f_i is the forward LIBOR rate from time i to $i + 1$, K_i is the terminal value of the loan at time i ,¹ and $v_{T,j} = F + c_T$. Thus, at each node i, j we compute the induced value, compare it with the terminal value, K_i , and set the value at that node, $v_{i,j}$ to the lesser of the two. In other words, if the induced value exceeds the terminal value, the loan is effectively repaid and terminates at i, j . Also, if the loan defaults at time i , the loan terminates with a value $v_{i,D} = R * F$ for all i . Finally, the value of $v_{i,j}$ at time 0 (in this example, at $v_{0,BB}$) is the *model* price of the loan.

Although equation 2 is useful for calculating prices of illiquid loans and for estimating the coupon premiums to charge for new loans, it is less useful for evaluating relative value among existing loans, which are better assessed using credit spreads. In fact, we can calculate the credit spread for a loan by discounting its expected nondefault cash flows by a constant amount over the LIBOR curve such that the discounted value matches its current market price. For all nondefault cash flows at a given time, the borrower will either prepay the principal and terminate, or pay a coupon and continue. The prepayment region in the time-and-credit-state lattice can be determined using the values of $v_{i,j}$ in equation 2. The probability of prepaying at period i is the sum of the probabilities of reaching nodes whose value of $v_{i,j}$ equal those capped at the terminal values K_i . Given the probability transition matrix and the set ω of all prepayment nodes, we can calculate the probability of prepayment at time i conditional on no prepayment before time i .

Let the conditional probability of prepayment at time be q_i ,² then the discounted cash flow is given by

$$V_J = \sum_{i=1}^T D_i * \prod_{j=1}^{i-1} (1 - q_j) * [(q_i * K_i) + ((1 - q_i) * CF_i)] \quad (3)$$

where $CF_i = c_i/4$ for $i < T$; $CF_i = (c_i/4 + F)$ for $i = T$, and the discount margin D_i is given by

$$D_i = \prod_{j=0}^i \frac{1}{\left(1 + \frac{f_{j,j-1} + \hat{s}}{4}\right)} \quad (4)$$

The credit spread, s , is determined by iteratively changing the parameter $\hat{s}$ and recalculating the discounted value of the cash flows, V_J , until V_J converges to P , the market price.

Revolving lines of credit are priced by assuming that the fraction of the loan drawn at a particular time, called the *usage*, is directly related changes to the obligor's credit quality. In other words, if a borrower's credit rating improves, it can access credit more cheaply and is also less likely to draw on existing lines of credit. Conversely, a borrower with deteriorating credit will likely draw on the credit lines it obtained when more highly rated. In this framework, usage can be interpreted as credit-dependent face value. Thus, in the equations above the face value is modified by $F \rightarrow U_j * F$ where j is the credit state and usage U_j ranges from 0 to 1.

End Notes

^aLIBOR stands for London interbank offered rate, which roughly corresponds to the interest rate charged between banks when lending large amounts of US dollars outside the United States. The coupon rate for a given quarter is set at the beginning of the period. For example, the L + 250 bp coupon in the text indicates that the borrower will pay one-quarter of 250 bp (0.625%) plus the current three-month LIBOR rate on the next coupon date.

^bAlthough some people define leveraged loans on the basis of their balance sheet leverage ratio, it is more common to use credit ratings (i.e., below BBB-) or credit spread to LIBOR above some maximum.

^cSeveral versions of the backward induction method have been proposed over the years [1, 6, 7, 9]. The version presented in equations (1–3) embodies elements that are common to most of these methodologies.

^dLoan spreads are typically quoted in basis points such as LIBOR + 300 bp, where 1% = 100 bp.

^eFor convenience, we assume LIBOR is constant at 2%, thereby generating a constant 5% coupon, and that the loan pays annually, rather than the typical quarterly coupon payment.

^fThe most well-known credit rating agencies are Fitch, Moody's, and Standard & Poor's.

^gRatings transition matrices are published regularly by the major agencies [4, 8].

^hMost models specify adjustment of physical credit transitions so that the default probabilities at each time, i , match the risk-neutral probabilities of default as implied by the bond and loan markets. For example, the risk-neutral default probability for a single risky cash flow at time t is given as $P_t^Q = \frac{1 - e^{-ts}}{(1 - R)}$ and $P_t^Q = N(N^{-1}(P_t) + \beta\lambda\sqrt{t})$ where, $P(t, Q)$ is the cumulative risk-neutral default probability to time t , s is the market credit spread, and R is the recovery rate in default. On the right, we calculate P_t^Q from P_t the physical default probability by adding a term related to the volatility of the credit relative to the market, the market price of risk, and the time to receipt of the cash flow. (For an elaboration and discussion of the derivation of this relation, see Bohn [1]. Zeng and Wen [9] describe its application to loan pricing.)

ⁱIt is common to add a refinancing premium to the principal plus coupon when defining the terminal value for evaluating prepayment as there are costs and/or penalties associated with the refinancing process.

^jThe probability of prepayment at time 1 from the initial state J is given by $q_1 = \sum_{k \in \omega} P_{J,k,0}$. For time $i > 1$, we must add the condition that the loan was not prepaid before time i ; thus, $q_i = \prod_{m=1}^{i-1} (1 - q_m) * \sum_{k \in \omega, l \notin \omega} P_{l,k,i-1}$.

References

- [1] Bohn, J. (2000). *A Survey of Contingent-Claims Approaches to Risky Debt Valuation*, Institutional Investor.
- [2] Deitrick, W. (2006). *Leveraged Loan Handbook*, Citi Markets and Banking.
- [3] Downs, J. & Goodman, J.E. (1991). *Dictionary of Finance and Investment Terms*, Barron's, Hauppauge, New York.
- [4] Emery, K., Ou, S., Tennant, J., Kim, F. & Cantor, R. (2008). *Corporate Default and Recovery Rates*, Moody's Global Corporate Finance. 1920-2007, Special Comment.
- [5] Miller, S. & William, C. (2007). *A Guide to the Loan Market*, Standard & Poor's.
- [6] Rizk, H. (1993). *GMPM Valuation Methodology: An Overview*, Citi Markets and Banking.
- [7] Rosen, D. *Does Structure Matter? (2002) Advanced Methods for Pricing and Managing the Risk of Loan Portfolios*, Algorithmics Inc.
- [8] Vazza, D., Aurora, D., Kraemer, N., Kesh, S., Torres, J. & Erturk, E. (2007). *Annual 2006 Global Corporate Default Study and Rating Transition*, Standard and Poor's Global fixed Income Research.
- [9] Zeng, B. & Wen, K. (2006). *CreditMark Valuation Methodology*, Moody's K.M.V.

Further Reading

Aguais, S., Forest, L. & Rosen, D. (2000). *Building a Credit Risk Valuation Framework for Loan Instruments*, Algo Research Quarterly.

TERRY BENZSCHAWEL, JULIO DAGRACA &
HENRY FOK

Credit Risk

Credit risk is the risk of an economic loss from the failure of a counterparty^a to fulfill its contractual obligations. For example, credit risk in the loan portfolio of a bank materializes when a borrower fails to make a payment, either the periodic interest charge or the periodic reimbursement of principal on the loan he contracted with the bank. Credit risk can be further decomposed into four main types: default risk, bankruptcy risk, deterioration in creditworthiness (or downgrading) risk, and settlement risk.

Default risk corresponds to the debtor's incapacity or refusal to meet his/her debt obligations, whether interest or principal payments on the loan contracted, by more than a reasonable relief period from the due date, which is usually 60 days in the banking industry.

Bankruptcy risk is the risk of actually taking over the collateralized, or escrowed, assets of a defaulted borrower or counterparty, and liquidating them.

Creditworthiness risk is the risk that the perceived creditworthiness of the borrower or counterparty might deteriorate. In general, deteriorated creditworthiness translates into a downgrade action by the rating agencies, such as Standard and Poor's (S&P) or Moody's, and an increase in the risk premium, or credit spread of the borrower. A major deterioration in the creditworthiness of a borrower might be the precursor of default.

Settlement risk is the risk due to the exchange of cash flows when a transaction is settled. Failure to perform on settlement can be caused by a counterparty defaulting, liquidity constraints, or operational issues. This risk is greatest when payments occur in different time zones, especially for foreign exchange transactions, such as currency swaps, where notional amounts are exchanged in different currencies.^b

Credit risk is only an issue when the position is an asset, that is, when it exhibits a positive replacement value. In that situation, if the counterparty defaults, the firm loses either all of the market value of the position or, more commonly, the part of the value that it cannot recover following the credit event. The value it is likely to recover is called the *recovery value* or *recovery rate* when expressed as a percentage; the amount it is expected to lose is called the *loss given default* (see **Recovery Rate**).

Unlike the potential loss given default on coupon bonds or loans, the one on derivative positions is

usually much lower than the nominal amount of the deal, and in many cases is only a fraction of this amount. This is because the economic value of a derivative instrument is related to its replacement, or market value, rather than its nominal or face value. However, the credit exposures induced by the replacement values of derivative instruments are dynamic: they can be negative at one point in time and yet become positive at a later point in time after market conditions have been changed. Therefore, firms must examine not only the current exposure, measured by the current replacement value, but also the profile of potential future exposures up to the termination of the deal.

Credit Risk at the Portfolio Level

The first factor affecting the amount of credit risk in a portfolio is clearly the credit standing of specific obligors (see **Rating Transition Matrices; Credit Rating**). The critical issue, then, is to charge the appropriate interest rate, or spread, to each borrower so that the lender is compensated for the risk he/she undertakes and to set the right amount of risk capital aside (see **Economic Capital**).

The second factor is "concentration risk" or the extent to which the obligors are diversified in terms of number, geography, and industry.

This leads us to the third important factor that affects the risk of the portfolio: the state of the economy. During economic boom, the frequency of default falls sharply compared with the periods of recession. Conversely, the default rate rises again as the economy enters a downturn. Downturns in the credit cycle often uncover the hidden tendency of customers to default together, with banks being affected to the degree that they have allowed their portfolios to become concentrated in various ways (e.g., customer, region, and industry concentrations) [1].

Credit portfolio models are an attempt to discover the degree of correlation/concentration risk in a bank portfolio (see **Portfolio Credit Risk: Statistical Methods**).

The quality of the portfolio can also be affected by the maturities of the loans, as longer loans are generally considered more risky than short-term loans. Banks that build portfolios that are not concentrated in particular maturities—"time diversification"—can

reduce this kind of portfolio maturity risk. This also helps reduce liquidity risk or the risk that the bank will run into difficulties when it tries to refinance large amounts of its assets at the same time.

Credit Derivatives and the ISDA Definition of a Credit Event

With the spectacular growth of the market for credit default swaps (CDSs) (*see Credit Default Swaps*), it has become necessary to be specific about what is a credit event? A credit event, usually a default, triggers the payment on a CDS. This event, then, should be clearly defined to avoid any litigation when the contract is settled. CDSs normally contain a “materiality clause” requiring that the change in credit status be validated by third-party evidence.

The new CDS market has struggled to define the kind of credit event that should trigger a payout under a credit derivatives contract. Major credit events as stipulated in CDS documentations and as formalized by the International Swaps and Derivatives Association (ISDA) are the following.

- Bankruptcy, insolvency, or payment default.
- Obligation/cross default that means the occurrence of a default (other than failure to make a payment) on any other similar obligation.
- Obligation acceleration which refers to the situation where debt becomes due and repayable prior to maturity. This event is subject to a materiality threshold of \$10 million unless otherwise stated.
- Stipulated fall in the price of the underlying asset.
- Downgrade in the rating of the issuer of the underlying asset.
- Restructuring: this is probably the most controversial credit event.
- Repudiation/moratorium: this can occur in two situations. First, the reference entity (the obligor of the underlying bond or loan issue) refuses to honor its obligations. Second, a company could be prevented from making a payment because of a sovereign debt moratorium (City of Moscow in 1998).

One of the most controversial aspects of the debate is whether the restructuring of a loan—which can include changes such as an agreed reduction in

interest and principal, postponement of payments, or change in the currencies of payment—should count as a credit event. The Conseco case famously highlighted the problems that restructuring can cause. In October 2000, a group of banks led by Bank of America and Chase granted to Conseco a three-month extension of the maturity of approximately \$2.8 billion of short-term loans, while simultaneously increasing the coupon and enhancing the covenant protection. The extension of credit might have helped to prevent an immediate bankruptcy, but as a significant credit event it also triggered potential payouts on as much as \$2 billion of CDS.

The original sellers of the CDS were not happy and were annoyed further when the CDS buyers seemed to play the “cheapest to deliver” game by delivering long-dated bonds instead of the restructured loans; at the time, these bonds were trading significantly lower than the restructured bank loans. (The restructured loans traded at a higher price in the secondary market due to the new credit-mitigation features.)

In May 2001, following this episode, ISDA issued a restructuring supplement to its 1999 definitions concerning credit derivative contractual terminology. Among other things, this document requires that to qualify as a credit event, a restructuring event must occur to an obligation that has at least three holders, and that at least two-thirds of the holders must agree to the restructuring. The ISDA document also imposes a maturity limitation on deliverables—the protection buyer can only deliver securities with a maturity of less than 30 months following the restructuring date or the extended maturity of the restructured loan—and it requires that the delivered security be fully transferable. Some key players in the market have now dropped restructuring from their list of credit events.

End Notes

^aIn the following, we use indifferently the term *borrower* or *counterparty* for a debtor. In practice, we refer to issuer risk, or borrower risk, when credit risk involves a funded transaction such as a bond or a bank loan. In derivatives markets, counterparty risk is the credit risk of a counterparty for an unfunded derivatives transaction such as a swap or an option.

^bSettlement failures due to operational problems result only in payment delays and have only minor economic consequences. In some cases, however, the loss can be quite

substantial and amount to the full amount of the payment due. A famous example of settlement risk is the 1974 failure of Herstatt Bank, a small regional German bank. The day it went bankrupt, Herstatt had received payments in Deutsche Mark from a number of counterparties but defaulted before payments were made in US dollars on the other legs of maturing spot and forward transactions.

Bilateral netting is one of the mechanisms that reduce settlement risk. In a netting agreement, only the net balance outstanding in each currency is paid instead of making payments on the gross amounts to each other. Currently, 55% of the foreign exchange (FX) transactions are settled through the CLS Bank that provides a payment-*versus*-payment (PVP) service that virtually eliminates the principal risk associated with settling FX trades [2].

References

- [1] Basel Committee on Payment and Settlement Systems (2008). *Progress in Reducing Foreign Exchange Settlement Risk*, Bank for International Settlements, Basel, Switzerland, May 2008.
- [2] Caouette, J., Altman, E., Narayanan P. & Nimmo, R. (2008). *Managing Credit Risk: The Great Challenge for Global Financial Markets*, Wiley.

MICHEL CROUHY

Credit Default Swaps

A credit default swap (CDS) is a contract between two parties, the *protection buyer* and a *protection seller*, whereby the protection buyer is compensated for the loss generated by a *credit event* in a reference instrument (see Figure 1). The credit event can be the default of the reference entity, lack of payment of coupon, or other corporate events defined in the contract. In return, the protection buyer pays a premium, equal to an annual percentage X of the notional, to the protection seller. The premium X , quoted in basis points or percentage points of the notional, is called the *CDS spread*. This spread is paid (semi)annually or quarterly in arrears until either maturity is reached or default occurs.

There are various methods for settlement at default. In a *cash settlement*, the protection seller pays the protection buyer the face value of the reference asset minus its postdefault market value. In a *physical settlement*, the protection buyer receives the initial price of the reference minus the postdefault market value, but, in turn, must make physical delivery of the reference asset or a bond from a pool of eligible assets to the protection seller in exchange for par. In both cases, the postdefault market value of the reference is typically determined by a dealer poll. The contract may also stipulate a fixed or “digital” cash payment at default, representing a fixed percentage of the notional value.

Example A protection buyer purchases 5-year protection on an issuer with notional \$10 million at an annual premium (spread) of 300 basis points or 3%. Suppose the reference issuer defaults 4 months after inception and that the reference obligation has a recovery rate (see **Recovery Rate**) of 45%. Thus, 3 months after inception, the protection buyer makes the first spread payment, roughly equal to $\$10\text{ million} \times 0.03 \times 0.25 = \$75\,000$. At default, the protection seller compensates the buyer for the loss by paying $\$10\text{ million} \times (100 - 45\%) = \5.5 million , assuming the contract is settled in cash. At the same time, the protection buyer pays to the seller the premium accrued since the last payment date, roughly equal to $\$10\text{ million} \times 0.03 \times 1/12 = \$25\,000$. The payments are netted. With these cash flows the swap expires; there are no further obligations in the contract.

CDS contracts are usually documented according to International Swaps and Derivatives Association (ISDA) standards and specify the following:

- A reference entity, whose default (the “credit event”) triggers the default payments in the CDS.
- A reference obligation, which can be a loan, a bond issued by a corporation or a sovereign nation, or any other debt instrument.
- A maturity, the common maturities being 1, 3, 5, 7, and 10 years although the majority of standardized CDSs are 5-year swaps.
- A calculation agent, responsible for computing the payouts related to the transaction. The calculation agent can be one of the counterparties of the CDS or a third party.
- A set of deliverable obligations, in case of physical settlement.

CDSs were introduced in 1997 by JPMorgan and subsequently became the most common form of credit derivative, amounting to a notional value of USD 64 trillion in 2008. With the onset of the financial crisis, this notional volume has gone down to around USD 38 trillion in the first half of 2009, but it remains large.

CDSs are over-the-counter (OTC) derivatives and are not yet exchange traded. The CDS market is a dealer market where a dozen major institutions control an overwhelming proportion of the volume and post quotes for protection premiums on various reference entities.

Uses of Credit Default Swaps

To gain exposure to the credit risk of a firm, an investor can purchase a bond issued by the corporation by paying the face value (or current price) of the bond and collect the interest paid by the issuer. Alternatively, he/she could sell protection in a credit swap referenced on the issuer’s bond. Relative to buying the reference security directly, the CDS position has the advantage of leading to the same exposure but not requiring a capital at inception. Also, if the reference entity is a foreign or sovereign entity, a CDS with a domestic counterparty might greatly simplify the legal structure of the transaction.

The protection buyer is short the credit risk associated with the reference obligation. If the buyer actually owns the reference security, then the CDS

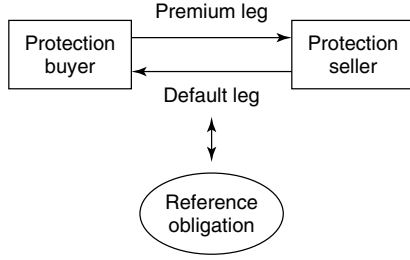


Figure 1 Structure of a credit default swap (CDS)

acts as a hedge against default. For a bank hedging its loans, this can lead to economic and regulatory capital relief. If the buyer does not have exposure to the reference security, the CDS enables him/her to take a speculative short position that benefits from a deterioration of the issuer's creditworthiness.

CDSs are often used to hedge against losses in the event of a default. Thus, CDSs can be viewed as insurance contracts against default or, more generally, as insurance against credit events. However, it is important to note that, unlike the case of insurance contracts, the protection buyer does not need to own the underlying security or have any exposure to it. In fact, an investor can speculate on the default of an entity by buying protection on a reference entity. Thus, they are more like deep out-of-the-money equity puts rather than insurance contracts.

The sheer volume of the CDS market indicates that a large portion of contracts are speculative since, in many cases, the outstanding notional of CDSs is (much) larger than the total debt of the reference entity. For example, when it filed for bankruptcy on September 14, 2008, Lehman Brothers had \$155 billion of outstanding debt, but more than \$400 billion notional value of CDS contracts had been written with Lehman as reference entity [8].

Also, unlike insurance companies, which are required to hold reserves in accordance with their issued insurance claims, a protection seller in a CDS is not required to maintain any reserves to pay off buyers. An important case is the event where a protection seller has insufficient funds to cover the default payment, thereby defaulting on its CDS payment. A famous example is the downfall of AIG, in which CDSs sold by its Financial Products subsidiary (AIGFP) played a major role.

CDSs, like many other credit derivatives, are unfunded and typically do not appear as a liability

on the balance sheet of the protection seller. This off-balance sheet nature makes them attractive to many investors, allowing them to take a synthetic exposure to a reference entity without directly investing in it. However, it can also lead to a lack of transparency and generate large exposures, which are not readily visible to regulators and market participants, and not subject to adequate capital requirements.

Valuation

A basic question is to determine the fair swap spread, or the premium, at inception. The CDS spread must equate the present value at inception of the premium payments (premium leg) and the present value of the payments at default. After inception, the swap must be marked to the market. Arbitrage-free valuation of credit default swaps can be done by using the risk-neutral pricing principle (*see Risk-neutral Pricing*): we assume a pricing measure $\mathbb{Q}$ such that the present value at t of any payout H at $T > t$ is $E^{\mathbb{Q}}[B(t, T)H]$ where $B(t, T)$ is the (risk-free) discount factor.

Consider a CDS with the notional N , payment dates $T_1, T_2, \dots, T_n = T$. Denote the (random) date of the underlying credit event as τ . A key role is played by the conditional risk-neutral survival probability $S(t, T) = \mathbb{Q}(\tau > T | \mathcal{F}_t)$ where $\mathcal{F}_t$ represents information available at date t . We denote $S(T) = S(0, T)$, its value at the inception of the contract. Denote the recovery rate by R , and $\bar{R} = E^{\mathbb{Q}}[R]$, the “implied” recovery rate (*see Recovery Swap*).

The premium leg pays a fixed annual percentage X on the notional N at dates T_i until default: the cash flow at T_i is therefore

$$XN(T_i - T_{i-1})1_{\tau > T_i} \quad (1)$$

The value at inception $t = 0$ of this stream of cash flows is therefore

$$\begin{aligned} & \sum_{i=1}^n XN(T_i - T_{i-1})B(0, T_i)E^{\mathbb{Q}}[1_{\tau > T_i} | \mathcal{F}_0] \\ &= XN \sum_{i=1}^n (T_i - T_{i-1})B(0, T_i)S(0, T_i) \\ &= XN \sum_{i=1}^n (T_i - T_{i-1})D(0, T_i) \end{aligned} \quad (2)$$

where $D(0, T) = B(0, T)S(T)$ is the risky discount factor and we have assumed independence of default times, recovery rates, and interest rates.

The protection leg (or default leg) can be modeled as a lump payment $N(1 - R)$ at T_i if default occurs between T_{i-1} and T_i (alternatively, one can consider other payment schemes such as payment at default [3, 4]). This can be represented as a stream of cash flows $N(1 - R)1_{T_{i-1} \leq \tau \leq T_i}$ paid at T_i . The value at inception ($t = 0$) of this cash-flow stream

$$\begin{aligned} & N \sum_{i=1}^n B(0, T_i) E^{\mathbb{Q}}[(1 - R)1_{T_{i-1} \leq \tau \leq T_i} | \mathcal{F}_0] \\ &= N(1 - \bar{R}) \sum_{i=1}^n B(0, T_i) \mathbb{Q}(T_{i-1} \leq \tau \leq T_i) \\ &= N(1 - \bar{R}) \sum_{i=1}^n B(0, T_i) (S(T_{i-1}) - S(T_i)) \end{aligned} \quad (3)$$

If payments are made at dates other than T_i , then accrued interest must be added. If payment dates are frequent (e.g., quarterly) the correction is small.

The fair spread for maturity T_n (or contracted spread or par spread) is defined as the spread that

equalizes at inception the values of the fixed and protection legs:

$$\begin{aligned} & XN \sum_{i=1}^n (T_i - T_{i-1}) D(0, T_i) \\ &= N(1 - \bar{R}) \sum_{i=1}^n B(0, T_i) [S(T_{i-1}) - S(T_i)] \end{aligned} \quad (4)$$

which yields

$$\begin{aligned} X &= \text{CDS}(T_n) \\ &= \frac{(1 - \bar{R}) \sum_{i=1}^n B(0, T_i) (S(T_{i-1}) - S(T_i))}{\sum_{i=1}^n (T_i - T_{i-1}) B(0, T_i) S(T_i)} \end{aligned} \quad (5)$$

Figure 2 shows the term structure of CDS spreads written on Lehman Brothers in September 2008.

To derive this formula, we have assumed that the firm's default time and recovery rate are independent, that interest rate movements are independent from

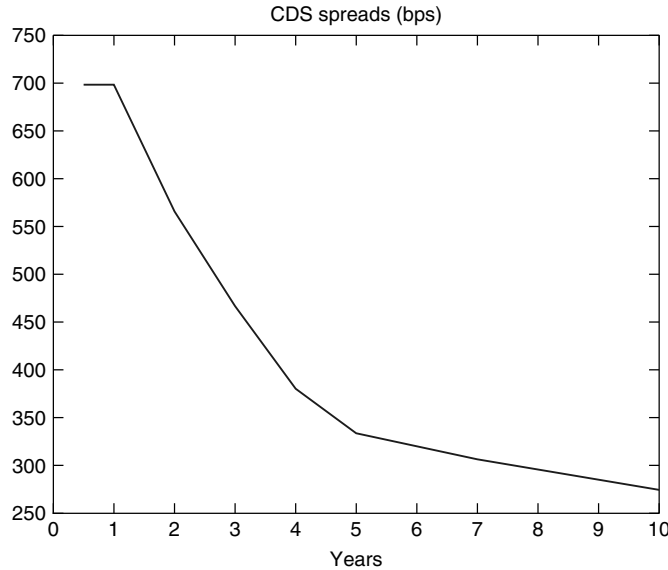


Figure 2 Term structure of CDS spreads on Lehman Brothers on September 8, 2008

default times, and that the protection seller has negligible default probability (no counterparty risk **Counterparty Credit Risk**). All these assumptions can be relaxed, especially in the context of reduced-form (see **Reduced Form Credit Risk Models; Intensity-based Credit Risk Models**) pricing models [3, 5–7]. Hull and White [6] discuss the incorporation of counterparty risk in CDSs. We note that

- CDS spreads depend on the term structure of default probabilities and on the term structure of interest rates, but only through payment dates $T_1, \dots, T_n$: two models that agree on the term structure of default probabilities will agree on CDS spreads.
- CDS spread depends on the recovery rate only through its expectation $\bar{R}$ under the pricing measure $\mathbb{Q}$. In market quotes, $\bar{R}$ has been usually chosen to be 40% for corporates, although this convention is subject to change.

Implied Default Probability

Given an estimate for the expected recovery rate $\bar{R}$ and the term structure of discount factors, one can solve equation (5) for the term structure

of default probabilities given the CDS spreads $\text{CDS}(T_1), \dots, \text{CDS}(T_n)$. The solution $S(T_i)$ is called the *implied survival probability* and $1 - S(T_i)$ is the implied default probability or the “risk-neutral” default probability implied by CDS quotes.

This procedure of inverting survival probabilities from CDS spreads is analogous to the procedure of stripping discount factors/zero coupon bond prices from bond yields (see **Yield Curve Construction**). Note that, as for yield curve construction, there are, in general, many more dates T_i (quarterly payments) than CDS maturities; hence, reconstructing $S(T)$ from CDS spreads requires interpolation or extra assumptions on survival probabilities. For example, survival probabilities are commonly parameterized as

$$S(t, T) = \exp \left[- \int_t^T h(t, u) du \right] \quad (6)$$

where $h(t, T) = -\partial_T S(t, T)/S(t, T)$ is the *forward hazard rate* (defined analogously to the forward interest rate **Heath–Jarrow–Morton Approach**).^a Reduced-form models (see **Reduced Form Credit Risk Models; Intensity-based Credit Risk Models**) lead to parametric functional forms for $h(t, \cdot)$, which can then be used to calibrate parameters to the observed CDS spreads.

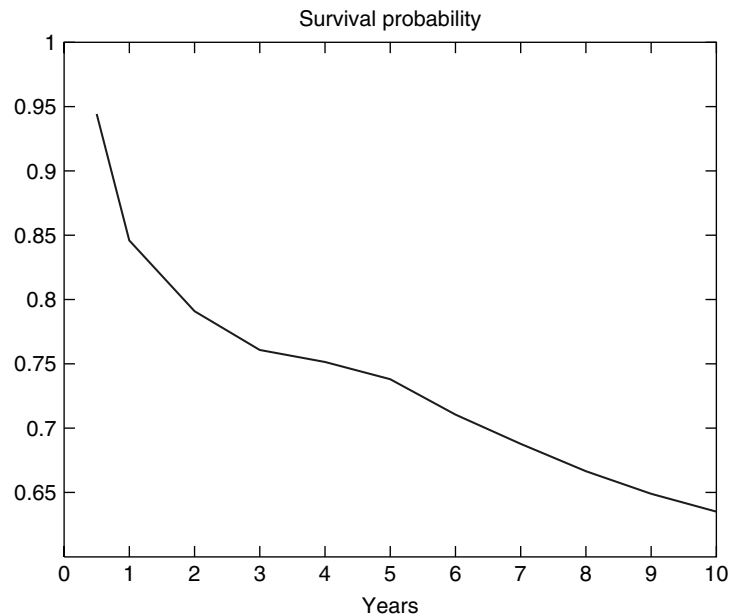


Figure 3 Risk-neutral survival probabilities implied by CDS spreads on Lehman Brothers on September 8, 2008

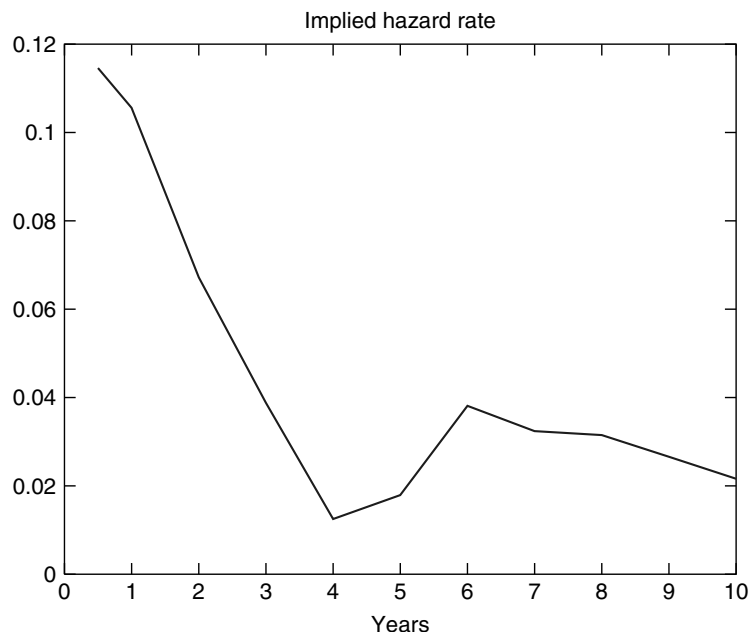


Figure 4 Hazard rates implied by CDS spreads on Lehman Brothers on September 8, 2008

Figure 3 shows survival probabilities for Lehman Brothers implied from CDS quotes on September 8, 2008, shortly before Lehman's default. Assuming that the hazard rate $h(t, T)$ is piecewise linear in T , we obtain the forward (annual) hazard rates shown in Figure 4.

This example might serve as a warning: such implied or "risk-neutral" default probabilities do not necessarily convey any information about the actual likelihood of the default of the reference entity, but they simply convey a market consensus on the premium for default protection at various maturities. Note also that the implied default probabilities and hazard rates depend on the assumption used for recovery rates.

Mark-to-Market Value of a Credit Default Swap (CDS) Position

At inception (say, $t = 0$) the mark-to-market value of a CDS position is zero for both counterparties. At a later date $t > 0$, this value is no longer zero: the mark-to-market value for the protection seller is the difference between the values of the fixed and protection legs:

$$\begin{aligned} & \text{CDS}(T_n)N \sum_{T_i > t} (T_i - T_{i-1})B(t, T_i)S(t, T_i) \\ & - N(1 - \bar{R}) \sum_{T_i > t} B(t, T_i)(S(t, T_{i-1}) - S(t, T_i)) \end{aligned} \quad (7)$$

where the sum runs over the remaining payments and the survival probabilities are now computed at time t . This quantity can be positive or negative, just as in an interest rate swap.

The mark-to-market value of the protection seller's position is the negative of the buyer's value (7). Note that the mark-to-market value (7) can be negative. This occurs when the credit quality of the reference name has improved since inception, and default protection is cheaper at current conditions, that is, available for a lower spread than that agreed upon at inception.

Triangle Formula

Consider now the simple case where the default time is described by a constant hazard rate λ (see **Hazard**

Rate):

$$S(0, T) = \exp\left(-\int_0^T \lambda dt\right) = \exp(-\lambda T) \quad (8)$$

If payments are assumed frequent, $T_i - T_{i-1} = \Delta T \ll T$, we can approximate the terms in equation (5) as

$$\begin{aligned} S(T_{i-1}) - S(T_i) &= -S'(T_i)(T_i - T_{i-1}) + o(T_i - T_{i-1}) \\ &= \lambda S(T_i) \Delta T + o(\Delta T) \end{aligned} \quad (9)$$

So

$$\begin{aligned} &\sum_{i=1}^n B(0, T_i)(S(T_{i-1}) - S(T_i)) \\ &= \sum_{i=1}^n \underbrace{B(0, T_i) S(T_i)}_{D(0, T_i)} \lambda(T_i) \Delta T \\ &\quad + o(1) \xrightarrow{\Delta T \rightarrow 0} \int_0^T dt \lambda \quad D(0, t) \end{aligned} \quad (10)$$

and

$$\sum_{i=1}^n (T_i - T_{i-1}) D(0, T_i) \xrightarrow{\Delta T \rightarrow 0} \int_0^T dt D(0, t) \quad (11)$$

Substituting in equation (5) we obtain the “triangle” relation

$$\text{CDS}(T) = (1 - \bar{R})\lambda$$

$$\text{CDS spread} = (1 - \text{recovery rate}) \times \text{hazard rate} \quad (12)$$

The assumption of a flat term structure of hazard rates is rather crude, but this formula is very useful in practice to get an order of magnitude of the risk-neutral default rate λ from CDS quotes.

Risk Management of Credit Default Swaps

Various factors affect the mark-to-market value of a CDS position. On a day-to-day basis the main concern is *spread volatility*: the value of a CDS position is primarily affected by changes in the CDS spread. Fluctuations in CDS spreads tend to exhibit heavy tails and strong asymmetry (upward moves in spreads have a heavier tail than downward moves) at daily and weekly frequencies. Figure 5 shows the daily returns in the CDS spread of CIGNA Corp. from 2005 to 2009: note the large amplitude of daily returns, which can attain 20% or 30%, especially on the upside. These tails are exacerbated by the relative *illiquidity* of many single-name CDS contracts.

Another concern is obviously the occurrence of the underlying credit event, which results in large payouts, whose magnitude is linked to the recovery rate and is difficult to determine in advance.

To provision for these risks, typically one or both parties to a CDS contract must post collateral and there can be margin calls requiring the posting of additional collateral during the lifetime of the contract, if the quoted spread of the CDS contract or the credit rating of one of the parties changes.

Additionally, as with other OTC derivatives, CDSs are exposed to counterparty risk. The counterparty risk exposure can be particularly large in a scenario where the protection seller and the underlying entity default together. This can happen, for example, if the protection seller has insufficient reserves to cover CDS payments. Counterparty risk affects the CDS spread if the default of the protection seller and the reference entity are perceived to be correlated [6]. The AIG fiasco in 2008 and the default of Lehman exacerbated the market perception of counterparty risk and has since distorted the level of CDS spreads, making it imperative to account for counterparty risk in the risk management of CDS portfolios.

To mitigate counterparty risk in the CDS market, it has been proposed by various market participants and regulators to clear CDS trades in clearinghouses. In a clearinghouse, the central counterparty acts as the buyer to every seller and seller to every buyer, thereby isolating each participant from the default of other participants. Participants post collateral with the central counterparty and are subject to daily margin calls. The introduction of a CDS clearinghouse can also reduce systemic risk resulting from CDS

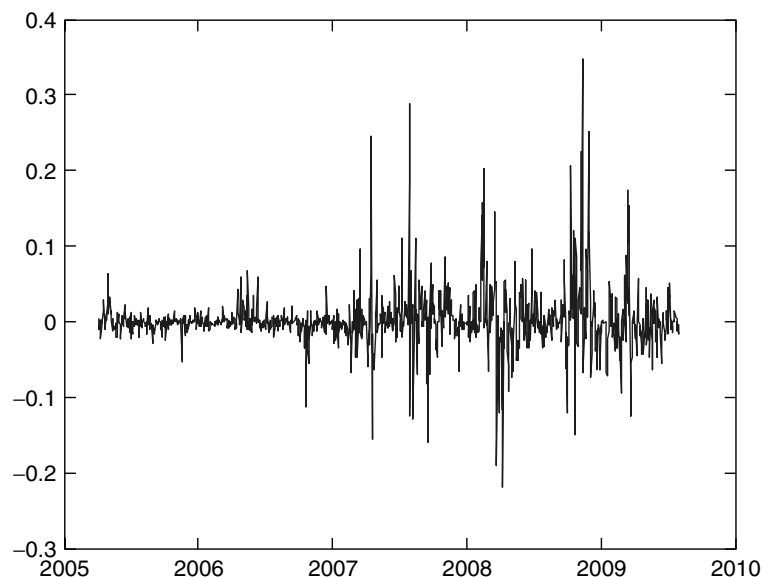


Figure 5 Daily (log-)increments of CDS spreads for CIGNA (CI), 2005–2009

transactions [1]. In the United States the first CDS clearinghouse, ICE Trust, began operating in March 2009. Other proposals to clear credit default swaps have been made by CME, NYSE Euronext, Eurex AG, and LCH Clearnet.

Credit Default Swap (CDS) Basis

An asset swap is a transaction between two parties in which the asset swap buyer purchases a bond from the other party and simultaneously enters into an interest rate swap transaction, usually with the same counterparty, to exchange the coupon on the bond for Libor plus a spread. The spread is called the *asset swap spread*. A common asset swap is the par asset swap where the buyer pays par at the inception of the deal. Unlike a CDS, an asset swap continues following bond default.

The *CDS-Bond basis* is the difference between the CDS spread and the asset swap spread on the same bond. It is an indicator of relative value of CDS *versus* the cash bond [2]. For example, when the CDS spread is higher than the asset swap spread, that is, the basis is positive, the CDS is generally considered to be more attractive than the bond. The reverse is true if the basis is negative. Negative CDS basis has been frequently observed during the recent financial crisis.

Changes in Conventions

Since 2009, the CDS market has been evolving in the direction of trading standardized single-name contracts with an upfront payment and a fixed coupon of either 100 or 500 bp and a common set of coupon payment dates (see www.cdsmodel.com). Standard maturity dates are March/June/September/December 20. Coupon payment dates are like standard maturity dates, but are adjusted to fall on the following business day. Each coupon is equal to $(\text{annual coupon}/360) \times (\text{number of days in accrual period})$. This simplifies processing and computation of coupons and cash flows. For example, every \$10 mm 100 bp standard CDS contract will pay the same 2Q09 coupon, \$26 111, on Monday, June 22, 2009, regardless of trade date, maturity, or reference entity.

The upfront payment is then set at the inception such that the buyer and seller positions have the same present value. In this convention, the dealer will quote not a spread (which is fixed) but an upfront payment. This convention applies to standardized CDS contracts on names contained in CDX and ITRAXX indices and may set the example for all other CDS contracts in the future.

End Notes

^a.Not to be confused with the (instantaneous) hazard rate or the default intensity (see **Hazard Rate**).

References

- [1] Cont, R. & Minca, A. (2009). *Credit Default Swaps and Systemic Risk*. Financial Engineering Report, Columbia University.
- [2] Davies, M. & Pugachevsky, D. (2005). Bond spreads as a proxy for credit default swap spreads, *Risk*.
- [3] Duffie, D. (1999). Credit swap valuation, *Financial Analyst's Journal* **54**(1), 73–87.
- [4] Duffie, D. & Singleton, K.J. (1999). Modeling term structures of defaultable bonds, *Review of Financial Studies* **12**, 687–720.
- [5] Hull, J. & White, A. (2000). Valuing credit default swaps i: no counterparty default risk, *Journal of Derivatives* **8**, 29–40.

- [6] Hull, J. & White, A. (2000). Valuing credit default swaps ii: modeling default correlations, *Journal of Derivatives* **8**, 897–907.
- [7] Schönbucher, P. (1998). Term structure modeling of defaultable bonds, *Review of Derivatives Research* **2**, 161–192.
- [8] VanDuyn, A & Weitzman H. (2008). Fed to hold CDS clearance talks, *Financial Times* (Oct 7).

Related Articles

Basket Default Swaps; Counterparty Credit Risk; Credit Default Swaption; Equity–Credit Problem; Exposure to Default and Loss Given Default; Hazard Rate; Intensity-based Credit Risk Models; Recovery Rate; Recovery Swap; Reduced Form Credit Risk Models.

RAMA CONT

Total Return Swap

A total return swap (TRS) is a financial contract between two counterparties to synthetically replicate the economic returns of an underlying asset. The principal mechanism and interaction are shown in Figure 1.

The reference asset still belongs to the TRS payer, who is buying protection from the TRS receiver. This reference asset contains typically a fixed interest payment and experiences a certain credit risk to be protected. The TRS payer transfers any payment made by the reference asset to the TRS receiver, who conversely pays a variable payment (typically the London interbank offered rate (LIBOR)) and a positive (or negative) spread as risk premium. Additionally, settlements for price depreciation and appreciations of the reference asset are made between the counterparties.

The TRS payer thus sells the market and credit risk of the reference asset to the TRS receiver without selling the reference asset itself. In the case of a credit event, the TRS receiver pays the difference between the value of the reference asset and the recovery value to the TRS payer. He acquires the counterparty risk of the TRS receiver instead.

Note that payments are not made continuously but rather at discrete times, that is, at given and specified reset periods. Occasionally, the reference asset consists of a whole portfolio of assets.

Reasons for Investing in a Total Return Swap

The TRS receiver explores the possibility of investing in the risk profile of the reference asset without owning it legally. Thus, insurance companies, hedge funds, and so on, count among the typical investors. They aim to work on a leveraged basis, diversify their portfolio, and achieve higher yields by taking on risk exposure. They can explore a synthetic way to make loans without having the costs and administrative burden; they explore possibilities to originating credit. Sometimes, for certain investors with capital constraints, TRS may be an effective way to leverage the use of capital.

TRS payers are typically lenders and investors who want to reduce their respective exposure to the

given borrower and potentially diversify a concentrated portfolio without removing the asset itself from their balance sheet, while maintaining the relationship with the borrower. However, TRS payers do not have to hold the asset itself on their balance sheets. If a TRS payer is taking an outright position, i.e. without holding the asset itself on the balance sheet, a TRS is an efficient way to go the asset short synthetically.

A TRS can help to activate comparative advantages of financing, depending on whether a certain market plays a role in a certain part of the market. Typically, a TRS is an off-balance-sheet deal.

Comparison with an Outright Investment in the Bond

The most striking difference from an outright investment in a bond or a loan is that with a TRS, price changes become cash flows at the predefined reset periods, at which settlements are made. For a bond, they are only accounting profits or losses and become effective at maturity or when the position is unwound. Thus, the TRS resembles a futures contract whereas the direct investment is more similar to a forward one (see, e.g., Schönbucher [3]).

Valuation and Risk Management

Schönbucher [3] gives an indication about the payoff streams of a TRS from the point of view of the TRS receiver to be counted for valuation purposes:

- Initially the TRS is closed at a fair value; hence, no cash flow is proceeded.
- If the bond does not default, the TRS receiver pays a variable coupon plus (or minus) a spread at every predefined reset point; he receives the interest from the bond and the difference in market value of the bond since the last reset is exchanged.
- If the bond defaults, the TRS receiver pays for a last time the variable coupon plus (or minus) a spread and the difference between the last market value of the bond and its recovery.

Thus, several risk factors influence the value of the TRS: the interest rate risk driven by the changing yield curve and the default probability of the

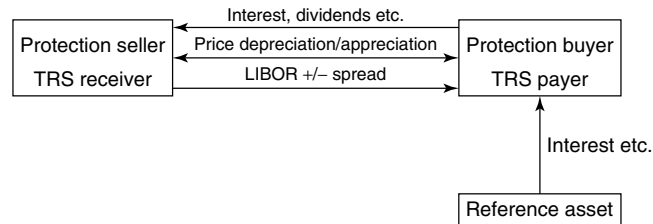


Figure 1 Mechanism of a total return swap (see Martin *et al.* [2])

reference asset (we neglect, for instance, the counterparty risk). Typical valuation models include the Duffie–Singleton model, hazard rates, and forward measure. The credit risk is reflected in the fair spread (fair means that initially there has to be no cash flow) (see also Anson *et al.* [1]).

References

- [1] Anson, M.J.P., Fabozzi, F.J., Choudry, M. & Chen, R.R. (2004). *Credit Derivatives: Instruments, Applications, and Pricing*. John Wiley & Sons.
- [2] Martin, M.R.W., Reitz, S. & Wehn, C.S. (2006). *Kreditderivate und Kreditrisikomodelle- Eine mathematische Einführung*, Vieweg Verlag. (in German).

- [3] Schönbucher, P. (2003). *Credit Derivatives Pricing Models: Models, Pricing and Implementation*, Wiley.

Further Reading

- Kasapi, A. (1999). *Mastering Credit Derivatives—A Step-by-Step Guide to Credit Derivatives and their Application*. Prentice Hall.
- Tavakoli, J.M. (1998). *Credit Derivatives. A Guide to Instruments and Applications*, Wiley.

CARSTEN S. WEHN

Recovery Swap

A recovery swap (RS), also called a *recovery lock* or a *recovery default swap* (RDS), is an agreement to exchange a fixed recovery rate R_S for the realized recovery rate ϕ , the latter being determined under prespecified contractual terms. The fixed recovery rate may be specified in terms of a recovery of par amount (RP), or as the recovery percentage of an equivalent Treasury bond, known as *recovery of Treasury* (RT), or as a fraction of the market value of the bond prior to default, also known as *recovery of market value* (RMV).

A recovery swap is no different than a forward contract at rate R_S on the underlying recovery rate ϕ . The maturity of the contract is denoted as T . If the reference credit underlying the recovery swap does not default before T , the swap expires worthless. There are no intermediate or periodic cash flows in a recovery swap. In a liquid market for recovery swaps, the quoted rate R_S is the best forecast of the expected recovery rate for default at time T . This recovery rate may then be used to price credit default swaps (CDSs).

We assume that the buyer of the recovery swap will receive R_S and pay ϕ . Hence, the buyer gains when the realized recovery rate is lower than that of the strike rate R_S . The net payoff to the contract is $(R_S - \phi)$. Recovery swaps are quoted in terms of the “strike” rate R_S . For example, a dealer might quote a recovery swap in GM at 37/40. This means the dealer is prepared to sell a recovery swap with $R_S = 37\%$ and buy at $R_S = 40\%$.

Replication and No-arbitrage

A recovery swap may be synthesized by selling a fixed recovery CDS (also known as a *digital default swap* or DDS) at a predetermined recovery rate R_D and buying a standard CDS. When the reference name defaults, the seller of the DDS pays the loss amount on default $(1 - R_D)$ and receives $(1 - \phi)$ on the CDS, thereby generating cash flow $(R_D - \phi)$. There is a triangular arbitrage restriction between the three securities: RS, DDS, and CDS. If we hold 1 unit each of the RS, CDS, and DDS, then we would need that $R_S = R_D$.

In order to state the triangular arbitrage relation more generally, consider the case when $R_S \neq R_D$, that is, when the strike recovery rates of the recovery swap and DDS are not the same. Let the premium on the CDS be c_1 and the premium on the DDS be c_2 . In order to replicate the RS, we will hold x units of the CDS and y units of the DDS. The replication has two conditions:

1. The cashflows at default must be equal for the RS and the replicating portfolio of CDS and DDS. In other words,

$$x \cdot (1 - \phi) - y \cdot (1 - R_D) = R_S - \phi \quad (1)$$

2. The premiums of the replicating portfolio must be net zero as the recovery swap does not have any intermediate cash flows. Hence the following equation must hold:

$$y \cdot c_2 - x \cdot c_1 = 0 \quad (2)$$

Set $x = 1$ in equation (1) so as to eliminate dependence on ϕ in the equation. Then we have that

$$x = 1 \quad \text{implies} \quad y = \frac{1 - R_S}{1 - R_D} \quad (3)$$

Substituting this result for x, y in equation (2) results in the following:

$$\frac{c_1}{c_2} = \frac{1 - R_S}{1 - R_D} = \frac{\mathcal{L}_S}{\mathcal{L}_D} \quad (4)$$

where $\mathcal{L}$ denotes the loss rate. We note the following:

- The no-arbitrage condition in equation (4) between the three securities implies that the ratio of the premium on the CDS to the premium on the DDS is equal to the ratio of loss rates on the recovery swap and the digital CDS. This is because the quote on the recovery swap R_S is the expected recovery rate on the CDS contract.
- It follows immediately that if R_D is specified, then equation (4) mandates a precise relation between the quotations on the three types of contracts, that is, rate R_S for the recovery swap, premium c_1 on the CDS, and premium c_2 on the DDS. Given the quotes on any two of these securities, the quote on the third security is immediately obtained.

- These no-arbitrage based results do not depend in any way on the underlying process for default or that of recovery. This makes the relationships in equation (4) very general and easy to apply in practice, as well as easy to assess empirically for academic purposes.

Applications and Uses of Recovery Swaps

Recovery swaps were first developed by BNP Paribas in early 2004 [10]. In response to market demand, banks started issuing fixed-rate recovery collateralized debt obligations (CDOs) and as a consequence were bearing recovery rate risk. In order to hedge against this recovery rate risk, market participants started selling recovery swaps.

Recovery swap markets are predominantly traded on reference entities with a high risk of default or of declining credit quality. For this reason, the largest activity in the recovery swaps market is in the auto parts and auto manufacturing sectors and geographically on North American entities [7]. Trading volumes in recovery swaps, although still small relative to the overall credit derivatives market, increased in 2005 with the defaults of Delphi Corporation and the Collins & Aikman Corporation [7]. Still, the market remains largely undeveloped and the International Swaps and Derivatives Association (ISDA), in May 2006, issued a template for the documentation on recovery swaps but the full documentation remains to be completed at this time [13].

There are two primary uses of recovery swaps. The first is to isolate the probability of default from the recovery rate. Traders may have in-house expertise in determining default probabilities but not in determining recovery and thus may wish to hedge their recovery risks through recovery swaps. The second use of recovery swaps is to eliminate recovery basis risk. Recovery basis risk occurs because of different settlement procedures between CDSs and CDOs. CDSs are often settled physically, meaning that when default occurs the seller of protection receives the defaulted bonds, whereas CDOs are almost always cash settled. The difference in settlement procedures is the source of recovery basis risk. For instance, an investor might hold a CDS Index that includes a given reference entity and have an offsetting position by selling the single-name CDS of the same entity. The investor is hedged

against the default risk of this entity, but because of differences in settlement at default between the CDS Index and the single-name CDS, the investor might get different recovery rates on the two instruments (recovery basis risk). Hence, recovery swaps can hedge against recovery basis risk by locking-in recovery rates.

Furthermore, in the case where the CDSs specify a physical settlement, it is possible that the underlying bonds might be scarce compared to the notional amount of CDS traded on the bond. This causes a “delivery squeeze” where the price, and therefore the recovery of the bond, is artificially increased because the buyers of CDS need to buy the bonds for delivery to their counterparty. For instance, in October 2005, Delphi Corporation had \$27.1 billion of outstanding CDSs against notional outstanding bonds of just \$2 billion causing the price of the defaulted bonds to surge by as much as 24% [9]. The consequence of this delivery squeeze is to reduce the profits accruing to buyers of CDSs, and recovery swaps provide a hedge against this by locking in the recovery rate ahead of time. More recently though, most CDSs are being settled in cash, thereby circumventing this problem.

Recovery Risk

There is a growing literature on recovery risk. Berd [3] provides a nice introduction and analysis of recovery swaps. DDSs are analyzed in [4]. Altman *et al.* [2] present a detailed study showing how recovery rates depend on default rates, positing and finding an inverse relationship. Chan-Lau [6] presents a method to obtain the upper bound on recovery on emerging market debt. Das and Hanouna [8] develop a methodology for identifying implied recovery rates and default probabilities from CDS spreads and data on stock prices and volatilities. Acharya *et al.* [1] provide empirical evidence that recovery rates depend on the industry, state of the economy, and specificity of assets to the industry in which the firm operates. Carey and Gordy [5, 14] show that recovery has systematic risk. Guo *et al.* [11] look at recoveries in reduced form models by explicitly modeling the postbankruptcy process of recoveries. The well-known loss given default model of Gupton and Stein [12] is well liked and used. Absolute priority rule (APR) violations are modeled in [15]. For a nice overview, see [16].

References

- [1] Acharya, V., Bharath, S. & Srinivasan, A. (2007). Does industry-wide distress affect defaulted firms? – Evidence from creditor recoveries, *Journal of Financial Economics* **85**(3), 787–821.
- [2] Altman, E., Brady, B., Resti, A. & Sironi, A. (2005). The link between default and recovery rates: theory, empirical evidence and implications, *Journal of Business* **78**, 2203–2228.
- [3] Berd, A. (2005). Recovery swaps, *Journal of Credit Risk* **1**(3), 1–10.
- [4] Berd, A. & Kapoor, V. (2002). Digital premium, *Journal of Derivatives* **10**(3), 66.
- [5] Carey, M. & Gordy, M. (2004). *Measuring Systematic Risk in Recovery on Defaulted Debt I: Firm-Level Ultimate LGDs*. Working paper, Federal Reserve Board.
- [6] Chan-Lau, J.A. (2003). *Anticipating Credit Events using Credit Default Swaps, with an Application to Sovereign Debt Crisis*. IMF working paper.
- [7] Creditflux Ltd (2006). *Jump-to-default Hedging Spurs Recovery-swap Surge*. January 1, 2006.
- [8] Das, S. & Hanouna, P. (2007). *Implied Recovery*. Working paper, Santa Clara University.
- [9] Euromoney Magazine (2006). *Why CDS Investors need to Lock in Recovery Rates Now*. May 1, 2006.
- [10] Financial Times (2004). *Capital Markets & Commodities: Investors Welcome Recovery Swap Tool*. June 18, 2004.
- [11] Guo, X., Jarrow, R. & Zeng, Y. (2005). *Modeling the Recovery Rate in a Reduced Form Model*. Working paper, Cornell University.
- [12] Gupton, G. & Stein, R. (2005). *LossCalc2: Dynamic Prediction of LGD*. Working paper, Moodys.
- [13] Investment Dealers' Digest (2006). *New ISDA Documentation Boosts Recovery Swaps*. May 22, 2006.
- [14] Levy, A. & Hu, Z. (2006). *Incorporating Systematic Risk in Recovery: Theory and Evidence*. Working paper, Moodys KMV.
- [15] Madan, D., Guntay, L. & Unal, H. (2003). Pricing the risk of recovery in default with APR violation, *Journal of Banking and Finance* **27**(6), 1001–1218.
- [16] Schuermann, T. (2004). *What do we know about Loss Given Default?* 2nd Edition, Working paper, Federal Reserve Bank of New York, forthcoming in *Credit Risk Models and Management*.

Related Articles

Credit Default Swaps: Exposure to Default and Loss Given Default; Recovery Rate.

SANJIV R. DAS & PAUL HANOUNA

Constant Maturity Credit Default Swap

A constant maturity credit default swap (CM CDS) is a credit derivative with payments linked to periodic fixings of a standard CDS rate with a fixed tenor (e.g., 5 years) on a particular credit entity. In business practice, CM CDS is usually presented as an elaboration of a plain CDS suitable for use in trading strategies expressing certain views on the steepening or flattening of credit spreads. From the point of view of quantitative modeling, CM CDS is best understood as a simple representative of a family of structured credit exotics, more complicated members of which would depend on a nonlinear combination of CDS spreads of more than one maturity and/or refer to more than one credit entity. Both business practice and quantitative modeling of CM CDS are based on ideas and techniques originally developed for CMS-linked fixed income exotics and adapted to credit modeling.

Instrument Structure

A CM CDS trade usually consists of two legs, one of which is a CM CDS leg, which is a sequence of payments $P_1, P_2, \dots, P_n$ made on a schedule of payment dates $T'_1, T'_2, \dots, T'_n$ (which is typically set following conventions of a standard CDS schedule) and is computed by the following formula:

$$P_i = \Delta_i \min[C, a \cdot S(T_i, T_{i+M})] \quad (1)$$

Here $S(T_i, T_{i+M})$ is the rate of a CDS spanning M payment periods and observed at the fixing date T_i associated with the payment date T'_i , Δ_i is the daycount fraction for the accrual period i , a is the multiplier called *participation rate*, and C is the fixed rate reset cap.^a Notional amount of the trade is set to 1. The fixing normally happens at the beginning of the accrual period, so that $T'_i = T_{i+1}$.^b In the case of a default of the reference credit, the fraction of the last payment accrued before the default is paid and further payments stop.

The second leg of a CM CDS is normally a standard CDS protection leg; however, structures where the second leg is either a standard CDS premium leg or another CM CDS leg corresponding

to a different tenor are also possible. The quote for a CM CDS is usually given in terms of the participation rate, a , of the CM CDS leg. With the second leg being the standard CDS protection leg, a participation rate of less than 100% reflects the expectation of rising credit spreads, whereas a participation rate of more than 100% corresponds to the expectation of decreasing credit spreads.

Trading Aspects

Descriptions of CM CDS structures started circulating in research communications of securities firms in early 2004, [3, 5, 6]. On November 21, 2005, ISDA provided a publication [4] setting a standard for the terms and condition of the CM CDS leg, including a mechanism for the determination of CDS rate resets. Establishing undisputable resets for CDS rates presents a problem because no standard source of information on CDS rates similar to Telerate pages of interest rates has emerged.

The primary mechanism stipulated by ISDA compels the seller of protection in a CM CDS to make a binding bid for the fixed rate premium that the seller is willing to pay in exchange for protection in a standard CDS on the same credit entity. This bid will be used as a CDS rate reset and is expected to be a good proxy for the true CDS rate because the seller has no incentive to quote a value that is too high (in which case the seller can be forced to buy overpriced CDS protection) or too low (in which case the seller will lose on the coming CM CDS payment). A fallback mechanism, which comes into effect if the receiver fails to provide a bid, puts on the buyer the burden of obtaining CDS quotes from a set of fallback dealers, subsequently using the highest quote as the rate reset.

At the time of this study, the volume of CM CDS transactions remains limited, taming optimistic predictions for the development of multilayered structures, such as tranches of CM CDS portfolios, but not destroying such prospects completely. We also note that more exotic structures with payoffs linked to nonlinear combinations of CDS rates of different maturities were observed in the market and have potential for further growth.

Quantitative Modeling

The present value of a CM CDS leg is given by the sum of expectations of the payments (1) discounted

under risk-neutral measure:

$$V = \sum_i \Delta_i E[\min[C, a \cdot S_{T_i}(T_i, T_{i+M})] P_{T'_i} D(0, T'_i)] + A \quad (2)$$

$$A = \sum_i \sum_{T'_{i-1} < \tau'_k \leq T'_i} \Delta(T'_{i-1}, \tau'_k) \times E[\min[C, a \cdot S_{T_i}(T_i, T_{i+M})] \times (P_{\tau'_{k-1}} - P_{\tau'_k}) D(0, \tau'_k)] \quad (3)$$

Here $P_{T'_i}$ is the indicator of survival of the underlying name until time T'_i , and $D(0, T'_i)$ is the stochastic discount factor from 0 to T'_i . We equipped the notation $S_{T_i}(T_i, T_{i+M})$ for the CDS rate reset with a subscript T_i to indicate that this rate is modeled as a value at time T_i of a certain stochastic process. The contribution A of the interest accrued within the period of time $\Delta(T'_{i-1}, \tau'_k)$ between the last coupon payment and the time of default is written in a discretized form using a sufficiently frequent subdivision $\{\tau'_k\}$ of the segment $[T'_{i-1}, T'_i]$.

We consider major types of modeling approaches in the order of increasing sophistication, including the nonstochastic approximation, convexity adjustments, instantaneous hazard rate modeling, and forward credit spread modeling.

Nonstochastic Approximation

We can obtain a simple approximation assuming that the CDS rates are nonstochastic. In this approximation, each process $S_t(T_i, T_{i+M})$ of the expectation of the time T_i CDS rate reset conditional on the information accumulated until time t is frozen at $t = 0$. As a result, the relevant CM CDS rate at date T_i is equal to the forward CDS rate $F(T_i, T_{i+M})$, which can be expressed in terms of the survival probability function, $p_s(t) = E[P_t]$, and the deterministic discount function, $D_0(t) = E[D(0, t)]$,

$$F(T_i, T_{i+M}) = \frac{(1 - R) \sum_{T_i < \tau_j \leq T_{i+M}} (p_s(\tau_{j-1}) - p_s(\tau_j)) D_0(\tau_j)}{\sum_{j=i+1}^{i+M} (\Delta_j p_s(T_j) D_0(T_j) + A_j)} \quad (4)$$

$$A_j = \sum_{T_{j-1} < \tau_k \leq T_j} \Delta(T_{j-1}, \tau_k) \times (p_s(\tau_{k-1}) - p_s(\tau_k)) D_0(\tau_k) \quad (5)$$

Here R is the recovery rate and $\{\tau_j\}$ is a subdivision of the segment $[T_i, T_{i+M}]$, sufficiently frequent to enable an accurate calculation of the default leg and the contributions A_j of the premium accrued in the time $\Delta(T_{j-1}, \tau_k)$ between the last CDS coupon date and the event of default.^c

The value of a CM CDS leg obtained by setting $S_{T_i}(T_i, T_{i+M}) \rightarrow F(T_i, T_{i+M})$, $P_{T'_i} \rightarrow p_s(T'_i)$, and $D(0, T'_i) \rightarrow D_0(T'_i)$ in equations (2–3) gives a quick estimate sufficient for a qualitative understanding of the relationship between the term structure of CDS spreads and the CM CDS participation rate but missing the essential effects of the dynamics of credit spreads. Indeed, even apart from the obvious problem of not handling the cap condition, the expectation $E[S_{T_i}(T_i, T_{i+M}) P_{T'_i} D(0, T'_i)]$ can be very different from $F(T_i, T_{i+M}) p_s(T'_i) D_0(T'_i)$ when credit spread volatility is taken into account.

Convexity Adjustments

In this section a discussion on volatility-dependent corrections to the results obtained in the nonstochastic approximation is provided. Such correction can be either derived from a fully consistent model capable of computing $E[S_{T_i}(T_i, T_{i+M}) P_{T'_i} D(0, T'_i)]$ or introduced in an *ad hoc* manner. We begin by looking at the convexity adjustments in the more limited sense of instrument-specific adjustments without building a full-fledged model. In the remainder of this article, we omit the discussion of the accrued interest term A and focus on the main term in equation (2).^d

The first step is to switch to a new measure in which the process $S_t(T_i, T_{i+M})$ is a martingale. The numeraire $N_t(T_i, T_{i+M})$ of the desired measure is known as the *risky basis point value (RBPV)* and is given as

$$N_t(T_i, T_{i+M}) = P_t \sum_{j=i+1}^{i+M} \left(\Delta_j B_t(T_j) + \sum_{T_{j-1} < \tau_k \leq T_j} \Delta(T_{j-1}, \tau_k) H_t(\tau_{k-1}, \tau_k) \right) \quad (6)$$

where $B_t(T_j)$ is the time t value of a risky unit payment at T_j and $H_t(\tau_{k-1}, \tau_k)$ is the time t value of a unit payment at τ_k conditional on a default event in the interval $(\tau_{k-1}, \tau_k]$. We used a discretized form of the accrued interest consistent with equation (3).

The existence of the required measure follows from a representation of the CDS rate as a ratio of two tradable assets:

$$S_t(T_i, T_{i+M}) = L_t(T_i, T_{i+M})/N_t(T_i, T_{i+M}) \quad (7)$$

where $L_t(T_i, T_{i+M})$ is the time t expectation of the CDS default leg,

$$L_t(T_i, T_{i+M}) = (1 - R)P_t \sum_{T_i < \tau_k \leq T_{i+M}} H_t(\tau_{k-1}, \tau_k) \quad (8)$$

(For a rigorous discussion of the mathematics of measure change involving risky basis point value as a numeraire, see [8].) After the measure change, the contribution of each individual coupon payment to the CM CDS leg can be written as

$$\Delta_i N_0(T_i, T_{i+M}) E \left[\frac{B_{T_i}(T'_i)}{N_{T_i}(T_i, T_{i+M})} f(S_{T_i}(T_i, T_{i+M})) \right] \quad (9)$$

where $f(X) = \min(C, aX)$. The next step is to assume that S_t follows a lognormal martingale, $F \exp(\sigma W_t - 0.5\sigma^2 t)$, and to replace the true value of the ratio $B_{T_i}(T'_i)/N_{T_i}(T_i, T_{i+M})$ by the value of a suitable increasing function $g(S)$ at $S = S_{T_i}$. Imposing the condition $N_0(T_i, T_{i+M})g(F(T_i, T_{i+M})) = p_s(T'_i)D_0(T'_i)$ ensures that the calculation of the average (9) for $\sigma = 0$ brings us back to the nonstochastic valuation. A nonzero volatility $\sigma > 0$ leads to a positive correction due to the convexity of the product $g(S)f(S)$ in the region of values of S close to $F(T_i, T_{i+M})$ and distant from the cap C .

This approach has an advantage of relative simplicity and potential ability of calibrating the model volatility σ to CDS options. The disadvantage is an uncontrollable assumption in the choice of the function $g(S)$.

Instantaneous Hazard Rate Modeling

A more systematic modeling of CM CDS is possible in the framework of stochastic instantaneous hazard rates. This approach starts with postulating a stochastic differential equation (SDE) for the stochastic default intensity $\lambda(t)$. A reasonable choice is a lognormal process (similar to the Black–Karasinski model of interest rates) or an affine process (similar to the Cox–Ingersoll–Ross model of interest rates). A normal process (similar to the Hull–White model of interest rates) was also used despite a conceptual problem posed by positive probabilities of negative hazard rates. Multifactor models for joint stochastic evolution of hazard rates and instantaneous interest rates are also possible.

An exact analytical solution for a CM CDS is not available in any of these models because of a two-layered structure involving inner expectations for CDS rates fixings conditional on the state achieved on the fixing dates. The machinery of trees, lattices, or partial differential equation (PDE) solvers, however, can be accommodated to handle CM CDS structures. The key element is a construction of a slice of values of CM CDS rate fixings on the set of model states achieved on the fixing date. This is done using a representation of the CDS rate in terms of conditional expectations of elementary instruments $B_t(T)$ and $H_t(T_1, T_2)$ provided by equations (7), (6), and (8). We refer to Chapter 7 of the book [7] for the details of a possible realization of a tree-based construction.

An advantage of hazard rate modeling is its consistency that allows to price a wide range of credit instruments of different maturities, including CDS options, asset swaps, bond options, and credit linked notes using the same model. A notable disadvantage is the difficulty of calibration.

Forward Credit Spread Modeling

As drawbacks of short-rate models of interest rates led to the invention and development of swap and LIBOR market models, similar progression is taking place in the space of structured credit models. We refer to the work [1] and Chapter 23 of [2] for details of a model in which the CDS rates $S_t(T_i, T_{i+M})$ are chosen as primary variables.

An advantage of this approach is the ease of calibration and ability to derive efficient analytical

approximations under minimal additional assumptions. At present, the disadvantage is the paucity of relevant market data, leaving a large freedom in specifying the structure of volatilities and correlations. A full payback from this level of model sophistication cannot be expected until the market for structured products develops enough to provide liquid quotes for CDS option volatilities for a dense set of maturities, similarly to caplet and swaption volatility matrices in the interest rate markets.

End Notes

^a. The structure can obviously be extended to admit a fixed rate reset floor, which, however, is not included in the standard ISDA template.

^b. The actual payment dates T_i' usually have a delay of at least one business day and are rolled forward or backward to fall on a valid business day in accordance with currency-dependent conventions. In the practice of quantitative modeling, proper care is taken to make sure that correct discount factors reflecting the actual payment dates are used.

^c. These expressions are often written in terms of integrals obtained in the limit of an infinitely frequent discretization. The same remark applies to equation (3).

^d. A rigorous calculation of the convexity correction to accrued interest term is technically involved and can be avoided by using a proportionally adjusted correction to the main term.

References

- [1] Brigo, D. (2006). CMCDS valuation with market models, *Risk* June, 78–83.
- [2] Brigo, D. & Mercurio, F. (2007). *Interest Rate Models—Theory and Practice, with Smile, Inflation, and Credit*, 2nd Edition, Springer.
- [3] Calamaro, J.-P. & Nassar, T. (2004). *CMCDS: The Path to Floating Credit Spread Products*, Deutsche Bank, Global Markets Research.
- [4] ISDA (2005). *Additional Provisions for Constant Maturity Credit Default Swaps*, International Swaps and Derivatives Association, November 21, 2005. Available at www.isda.org.
- [5] Pedersen, C. & Sen, S. (2004). *Valuation of Constant Maturity Default Swaps*, Lehman Brothers, Quantitative Research Quarterly.
- [6] Renault, O. & Ratul, R. (2007). Constant maturity credit default swaps, in *The Structured Credit Handbook*, A. Rajan, G. McDermott & R. Ratul, eds, Wiley Finance, pp. 57–77.
- [7] Schönbucher, P. (2003). *Credit Derivatives Pricing Models*, Wiley Finance.
- [8] Schönbucher, P. (2004). Measure of survival, *Risk* August, 79–85.

Related Articles

Constant Maturity Swap; Convexity Adjustments; Credit Default Swaps; Credit Default Swaption; Forward and Swap Measures; Hazard Rate; Intensity-based Credit Risk Models; Swap Market Models; Term Structure Models.

TIMUR S. MISIRPASHAEV

Credit Default Swaption

Credit default swap (CDS) options, also known as *single name credit default swaptions*, allow an investor to buy protection on a reference name by entering a CDS (*see Credit Default Swaps*) at a previously set CDS spread. CDS options may knock out if the reference entity defaults before the exercise date. Usually, the buyer of protection will also receive the default payment in that case. Such a cash flow simply corresponds to the default leg of a CDS with maturity equal to the exercise date. We will thereafter assume cancellation of the contract if the underlying name defaults before the exercise date. We refer to **Credit Default Swap Index Options** for the extensions to the portfolio case.

Denote the default time associated with the underlying name by τ . This is usually the date of a credit event; we refer to ISDA master agreement for further details, given that this notion may vary through time and differ across geographical regions. For $t \leq s$, we denote by $\tilde{B}(t, s)$ the time t price of a defaultable discount bond, paying one at time s if $\tau > s$ and zero otherwise. Clearly, $\tilde{B}(t, s)$ collapses to zero at the default of the underlying name if it occurs before s , since no payment will eventually be received by the discount bond holder. $T_1, \dots, T_N$ denote the payment dates on the premium leg of the underlying CDS. For simplicity, we will further neglect effects of accrued premiums. T is the exercise date of the European credit spread call option and p the strike. We can write the payoff at time T as $(p_T - p)^+ \sum_{T_k > T} \tilde{B}(T, T_k)$. p_T is the CDS par spread at time T . There is also usually a multiplicative adjustment to take into account the premium payment frequency, which is quarterly in most cases and which is not dealt with here for notational simplicity. For the same reason, we do not account for a possible up-front payment associated with the CDS, which is likely to be applied after the implementation of the “big bang” CDS protocol. That will result in some small adjustments to the payoff function and thus to the pricing formulas, which will be neglected thereafter. Clearly, p_T is not defined if the option has already cancelled out, that is, if $\tau < T$, but the option payoff is equal to zero in that case.

Pricing Approaches

With some adaptations due to the cancellation feature of the CDS option, the pricing methodology parallels the well-known approaches in interest rate swaptions. One may consider a suitable distribution of the forward CDS spread under an appropriate risk-neutral measure. This readily leads to a Black-type pricing formula and is dealt with first. In another approach, one can rather model the “instantaneous” CDS spread, which is related to the “intensity” of the default time.

Black Formula for Credit Default Swap Options

As usual r denotes the default-free short rate and Q is the usual risk-neutral probability associated with the savings account. We consider $G = (G_t)$, a filtration such that τ is a stopping time and r is an adapted process.

The idea of using “survival measures” was introduced in [9] and further developed in [4, 6, 7, 10], among others (*see also Credit Default Swap Index Options*). We will denote by $\sum_{T_k > T} \tilde{B}(t, T_k)$ the “risky level” which corresponds to the time t price of a unitary premium leg associated with the forward CDS starting at T . We consider the probability measure $\hat{Q}$ associated with the previous risky level numéraire (*see also Change of Numéraire* about change of numéraire techniques) defined by

$$\frac{d\hat{Q}}{dQ} = \frac{\sum_{T_k > T} \tilde{B}(T, T_k)}{\sum_{T_k > T} \tilde{B}(0, T_k)} \exp - \int_0^T r(u) du \quad (1)$$

Let us remark that $\frac{d\hat{Q}}{dQ} = 0$ on the set $\{\tau \leq T\}$. Thus $\hat{Q}$ is absolutely continuous but not equivalent to Q . $\hat{Q}(\tau > T) = 1$ which leads to the terminology “survival measure”. We can then readily express the value of the credit spread option at $t = 0$ as

$$\left(\sum_{T_k > T} \tilde{B}(0, T_k) \right) E^{\hat{Q}} [(p_T - p)^+] \quad (2)$$

We can get around the issue of the CDS premium p_T not being defined *after default* by considering $p_T 1_{\{\tau > T\}}$, which we assume to be a random variable

with respect to G_T . This does not change the computation in the previous equation since $\hat{Q}(\tau > T) = 1$.

Let us first remark that for the forward CDS to be priced normally, we must have $E^{\hat{Q}}[p_T 1_{\{\tau > T\}}] = p_{0,T}$ where $p_{0,T}$ denotes the forward CDS premium. In the case where $p_T 1_{\{\tau > T\}}$ is lognormal under $\hat{Q}$, with volatility parameter σ , we readily get a Black formula for the price of the CDS option:

$$\left(\sum_{T_k > T} \tilde{B}(0, T_k) \right) \times (p_{0,T} N(d_1) - p N(d_2)) \quad (3)$$

where $d_1 = \frac{\ln\left(\frac{p_{0,T}}{p}\right) + \frac{\sigma^2}{2}T}{\sigma\sqrt{T}}$ and $d_2 = d_1 - \sigma\sqrt{T}$.

Intensity Approaches

Another approach consists in specifying the intensity of the default time. This is the path followed in [2–4]. To circumvent the difficulty with default intensity dropping down to zero after default and the various mathematical issues related to enlargement of filtrations, the easiest way is to model the default time through a Cox process. We thus define the default time τ associated with the underlying name as

$$\tau = \inf \left\{ t, \int_0^t \lambda(s) ds \geq -\ln U \right\} \quad (4)$$

where λ is a positive process adapted to some filtration $\mathfrak{F} = (\mathfrak{F}_t)$ and U follows a standard uniform variable independent of $\mathfrak{F}$. For simplicity, we will further assume that $(\mathfrak{F}, Q)$ is a Brownian filtration. Following [1] or [8], we define as $H = (H_t)$ the filtration generated by the counting process $N_t = 1_{\{\tau \leq t\}}$ and we denote by $G_t = \mathfrak{F}_t \vee H_t$, the relevant information at time t , incorporating knowledge about occurrence of default prior to t and current and past values of financial variables such as interest rates or credit spreads of the reference entity (see **Filtrations** for mathematical details about filtrations in finance). Up to default time, $\lambda(t)$ is the default intensity of τ (we refer to **Point Processes** regarding point processes and to **Compensators** about compensators and intensities). While the default intensity drops to zero after τ , we can remark that $\lambda(t)$ is still well defined, thanks to the above Cox modeling framework. For instance, one can consider shifted Cox–Ingersoll–Ross (CIR) processes for the

short rate r and the pseudodefaut intensity λ (see **Cox–Ingersoll–Ross (CIR) Model**).

$p_{t,T}$ will further denote the time t forward CDS premium. Though $p_{t,T}$ has a financial meaning only on the set $\{\tau > t\}$, its computation can be extended to the complete set of events in the previous Cox modeling framework (see [4] for further discussion). $p_{t,T}$ solves for the following equation where, once again, we do not take into account accrued premium or up-front payments effects

$$\begin{aligned} p_{t,T} \times \sum_{T_k > T} E \left[\left(\exp - \int_t^{T_k} (r + \lambda)(u) du \right) | \mathfrak{F}_t \right] \\ = \int_T^{T_N} E \left[\left(\exp - \int_t^s (r + \lambda)(u) du \right) \right. \\ \left. \times (1 - \delta)\lambda(s) | \mathfrak{F}_t \right] ds \end{aligned} \quad (5)$$

Prior to default, the left-hand term corresponds to the value at time t of the premium leg of the underlying forward default swap, while the right-hand term is associated with the default leg. Clearly $p_{t,T}$ is $\mathfrak{F}_t$ -measurable and we can prove that it is both a $(\mathfrak{F}, \hat{Q})$ and a $(G, \hat{Q})$ martingale. Thus, the forward default swap premium shares the properties of a “true” price. It can be checked that $p_{T,T} = p_T$.

Using an extended version of Girsanov theorem (see **Equivalence of Probability Measures**) for point processes (see **Point Processes**), it can be shown that

$$\frac{dp_{t,T}}{p_{t,T}} = \sigma d\hat{W}_t \quad (6)$$

where $\hat{W}$ is a $(\mathfrak{F}, \hat{Q})$ Brownian motion.

Let us also assume that there exists some specification of r and λ such that the volatility σ is constant. Then, the forward CDS spread has a lognormal dynamics under $\hat{Q}$. This readily leads to the already stated Black formula for the price of the CDS option. The most obvious advantage is the simplicity of the outcome. The drawbacks are also rather obvious. The lognormal assumption for the forward spreads is questionable since jumps are often included in the dynamics of λ as in the affine specification within [5].

The intensity approach is easy to understand and is consistent across strikes, maturity of the option, and maturity of the CDS. However, it entails dealing with extra parameters and is numerically more involved. In the more general setting involving correlation between r and λ , Monte Carlo simulation is usually required. In special cases, such as deterministic default-free rates, analytical formulas can be derived. Fortunately enough, in most examples, the correlation parameter has little impact on option prices and analytical approximations of the implied volatility in the Black formula can be derived. Let us remark that in these approximations σ depends on the exercise date and the maturity of the underlying CDS.

Acknowledgments

The author thanks A. Cousin, L. Cousot, A. Godet and C. Pedersen and the editors for helpful remarks. The usual disclaimer applies.

References

- [1] Bielecki, T.R. & Rutkowski, M. (2002). *Credit Risk: Modeling, Valuation and Hedging*, Springer.
- [2] Brigo, D. & Alfonsi, A. (2005). Credit default swap calibration and derivatives pricing with the SSRD stochastic intensity model, *Finance and Stochastics* **9**(1), 29–42.
- [3] Brigo, D. & Cousot, L. (2006). The stochastic intensity SSRD implied volatility patterns for credit default swap options and the impact of correlation, *International Journal of Theoretical and Applied Finance* **9**(3), 315–339.
- [4] Brigo, D. & Matteotti, C. (2005). *Candidate Market Models and the Calibrated CIR++ Stochastic Intensity Model for Credit Default Swap Options and Callable Floaters*. Working paper, Credit Models, Banca IMI.
- [5] Duffie, D. & Gârleanu, N. (2001). Risk and valuation of collateralized debt obligations, *Financial Analysts Journal* **57**(1), 41–59.
- [6] Hull, J. & White, A. (2003). The valuation of credit default swap options, *Journal of Derivatives* **10**(3), 40–50.
- [7] Jamshidian, F. (2004). Valuation of credit default swaps and swaptions, *Finance and Stochastics* **8**(3), 343–371.
- [8] Jeanblanc, M. & Rutkowski, M. (2000). Modelling of default risk: an overview, in *Mathematical Finance: Theory and Practice*, J. Yong & R. Cont, eds, Higher Education Press, Beijing, pp. 171–269.
- [9] Schönbucher, P.J. (2000). *A Libor Market Model with Default Risk*. Working paper, University of Bonn.
- [10] Schönbucher, P.J. (2003). *A Note on Survival Measures and the Pricing of Options on Credit Default Swaps*. Working paper, ETH Zurich.

Related Articles

Change of Numeraire; Compensators; Cox–Ingersoll–Ross (CIR) Model; Credit Default Swap Index Options; Credit Default Swaps; Filtrations; Point Processes.

JEAN–PAUL LAURENT

Credit Default Swap (CDS) Indices

Credit markets have shown tremendous growth in the last 10 years. In particular, the telecom bubble and corporate scandals of the early 2000s increased the interest of market participants in products such as credit default swaps (CDS) (*see Credit Default Swaps*), which provide protection against credit events. In response to this demand for credit protection, credit indices were introduced in 2003, increasing the liquidity of CDS markets. These indices are, in essence, standardized baskets of CDS written on investment-grade and high-yield corporate issuers, or emerging-market governments. Table 1 shows the basic composition criteria of the main indices (more stringent criteria apply too, in particular, those concerning liquidity of the individual CDS). The specific constituents for each index are posted at www.markit.com.

In most indices, issuers are equally weighted. A new series of a given index is issued semiannually, excluding from the basket those issuers who no longer match selection criteria (e.g., downgraded issuers) and adding new ones. In case of a default event, the defaulting issuer is removed from the basket, but the weights remain and the index continues to trade. The reduced basket is referred to as a *new version* of the same series. The loss payment for a default event is determined through the same settlement auction as for single-name CDS (*see Credit Default Swaps*).

Credit indices are commonly issued with initial maturities of 3–10 years. Similar to CDS, a credit index is a contract which entails that the protection buyer pays a spread (or coupon) at a regular frequency (usually quarterly according to International Swaps and Derivatives Association (ISDA) dates) in return for default protection on some notional amount. In case of a default of one of the referenced issuers, the protection seller pays the non-recovered part of the protected notional times the weight of the issuer in the index. The contract does not terminate, but the protected notional is reduced accordingly. Importantly, the index trades with a fixed spread for each series; changes in market pricing are reflected in the upfront payment required to enter the contract. In contrast, in a standard CDS

(for all but distressed credits), the spread is set on any given day such that no upfront payment is required.^a

A standard market practice is to roll index positions so as to maintain a position in the on-the-run (i.e., most recent) series and version, in order to guarantee maximum liquidity. From an investor's point of view, in addition to enabling credit diversification, credit indices introduce the possibility of leverage without significant liquidity concerns, as several derivatives on these indices exist today (*see Collateralized Debt Obligations (CDO); Credit Default Swap Index Options*).

Pricing Framework

Credit indices are routinely priced through the standard CDS model. Indeed, though the index contracts trade with a fixed spread, the convention is to quote a theoretical fair spread (i.e., the coupon that the index would need to pay in theory in order to require no upfront payment) and use the CDS model to convert this fair spread to an upfront payment for the index. The issuers in the basket are assumed to be homogeneous in credit quality and recovery rate. When deriving the common hazard credit curve for the issuers, the convention is to assume a flat curve (*see Hazard Rate*). The expected losses are computed from the credit curve, assuming that losses are paid at the end of a coupon period, and given a particular recovery rate. The present value for the index contract is then the difference between the discounted expected losses and the discounted spread payments weighted by the survival probability (since premium is only paid on the remaining protected notional).

The contract can alternatively be valued by using information on the individual constituents, thus relaxing the homogeneity assumption. We can theoretically replicate the index by considering a basket of individual CDS that pay the same spread. We compute the expected losses on the index by aggregating the individual-constituent expected losses, each of which is derived from the full-term structures of credit spreads for the constituent. Similarly, the payment side aggregates survival probabilities over all issuers. It is worth noting that the dependence structure between the issuers does not play a role here as the whole basket is considered.

Table 1 Main credit indices

Index name	Number of constituents	Region	Credit quality
CDX.NA.IG	125	North America	Investment grades
CDX.NA.IG.HVOL	30	North America	Low-quality investment grades
CDX.NA.HY	100	North America	Noninvestment grades
CDX.NA.HY.B		North America	B rated
iTraxx Europe	125	Europe	Investment grades
iTraxx Europe HiVol	30	Europe	Low-quality investment grades

Imperfect Replication

Pricing the index through the constituents is appealing, but it is not surprising to observe significant differences with the quoted index prices. The replicating strategy is not perfect, as the mechanics behind credit indices are slightly different from those for CDS. We mentioned earlier that an index trades with a floating upfront and a fixed spread, whereas most CDS trade with a floating spread and no upfront. This implies that we cannot, in general, enter into a basket of CDS contracts that pay the same spread as the index. And while the basket can be composed without initial capital, the index requires a nonzero upfront investment. After a default event, new differences between the credit index and the basket appear. On the index, the reduction in spread payments is independent of the defaulting issuer; the spread is fixed and only the protected notional changes. On the other hand, the spread reduction for the basket is proportional to the spread on the CDS for the specific defaulting issuer. The index and the basket consequently exhibit different behaviors through time, and offer different sensitivities to interest rates.

Fair Spread Decomposition

As contracts in their own right, credit indices are subject to specific demand and supply effects, and have their own distinct risk profile. A simple, standard way of analyzing their risk is to observe the quoted index fair spread. This approach, though, cannot distinguish the risk due to specific issuers from the risk due to demand for the index as a whole.

A useful decomposition is to break the index fair spread into three components: the average fair CDS spread across the constituent issuers, the nonlinear

component, and the basis. The first two components constitute the theoretical fair spread of the index, as replicated through a basket of (market-traded) issuer CDS. The nonlinear portion of this fair spread accounts for the heterogeneity in credit quality among the issuers, and increases both with the level of the average fair spread and the dispersion of the individual fair spreads. The nonlinear component is very sensitive to an increase in default likelihood of a single issuer. The basis—defined as the difference between the observed fair spread and the theoretical fair spread—contains a risk premium rewarding the index dealer for the small portion of risk that cannot be perfectly hedged through the replicating basket, and embeds a liquidity premium as well.

End Notes

^aNote that changes to the conventional CDS protocol were instituted in early 2009. Among other things, the new protocol stipulates that single-name CDS trade with a fixed coupon of 100 or 500 bp, and settle *via* an upfront payment (*see Credit Default Swaps* for further discussion).

Further Reading

Couderc, F. (2006). Measuring risk on credit indices: on the use of the basis, *Risk Metrics Journal* Winter 2007, 61–87.
Zhang, H. (2005). Instant default, upfront concession and CDS index basis, *Journal of Credit Risk* 1(2), 79–89.

Related Articles

Basket Default Swaps; Collateralized Debt Obligations (CDO); Credit Default Swap Index Options; Credit Default Swaps.

FABIEN COUDERC & CHRISTOPHER C. FINGER

Basket Default Swaps

Basket default derivatives or swaps are more sophisticated credit derivatives that are linked to several underlying credits. The standard product is an insurance contract that offers protection against the event of the k th default on a basket of n , $n \geq k$, underlying names. It is similar to a plain credit default swap (CDS) but the credit event to insure against is the event of the k th default, and it is not specified to a particular name in the basket. A premium, or spread, s is paid as an insurance fee until maturity or the event of k th default. We denote by $s^{k\text{th}}$ the fair spread in a k th-to-default swap, that is, the spread making the value of this swap equal to zero at inception. For the basic product description, we refer, for example, to [2, 3, 12].

If the n underlying credits in the basket default swap are independent, the fair spread $s^{1\text{st}}$ of a first-to-default swap (FtD) is expected to be close to the sum of the fair individual default swap spreads s_i over all underlying credits $i = 1, \dots, n$. For exponential waiting times, this follows since the minimum of exponentially distributed waiting times has itself an exponential distribution with an intensity that equals the sum of the intensities of the individual waiting times. If, on the other hand, the underlying credits are in some sense “totally” dependent the first default will be the one with the worst spread; therefore $s^{1\text{st}} = \max_i(s_i)$.

For the exact determination of the fair spread of basket default swaps, multivariate modeling of the default times of the credits in the basket is necessary. This dependency modeling can be classified into three different approaches, which are also used in collateralized debt obligations (CDO)-modeling (*see Collateralized Debt Obligations (CDO)*):

- Copula approach (*see Default Time Copulas; Gaussian Copula Model; Copulas: Estimation*)
- Asset-value approach (*see Structural Default Risk Models; also Merton, Robert C.*)
- Reduced-form, spread-based approach (*see Multiname Reduced Form Models; Hazard Rate; Intensity-based Credit Risk Models; Duffie–Singleton Model; Jarrow–Lando–Turnbull Model*)

Modeling Approaches

Copula Approach

As contingent payments are only triggered in case of default, copula modeling focuses on the multivariate distribution of default times.

The copula approach was first applied in this context in [13, 14]. The challenge is to specify a function C such that with given marginal distribution F_i , we have that

$$\begin{aligned} \text{Prob}\{\tau_1 \leq t_1, \dots, \tau_n \leq t_n\} &= F(t_1, \dots, t_n) \\ &= C(F_1(t_1), \dots, F_n(t_n)) \end{aligned} \quad (1)$$

Basically, the set of Copula functions coincides with the set of all multivariate distribution functions whose marginal distributions are uniform distributions on $[0, 1]$, since under certain regularity assumptions

$$C(u_1, \dots, u_n) = F(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n)) \quad (2)$$

One of the most elementary copula functions is the normal copula (or Gauss copula), which is derived by this approach from the multivariate normal distribution (*see Gaussian Copula Model*). Clearly, there are various different copulas generating all kinds of dependencies, for example, in [3, 12]. The advantage of the normal copula, however, is that it relates to the one period version of certain asset-value models used in credit portfolio risk modeling. But note that since the asset-value approach can only model defaults up to a single time horizon T , the calibration between the two models can only be done for one fixed horizon. Dynamic extensions of this in the asset-value context are exit time models.

Asset-value Models

In asset-value models we are looking for stochastic processes (Y_t^i) called *ability-to-pay processes* and (nonstochastic) barriers $K_i(t)$ such that the default time τ_i for credit i can be modeled as the first hitting time of the barrier $K_i(t)$ by the process (Y_t^i) :

$$\tau_i = \inf\{t \geq 0 : Y_t^i \leq K_i(t)\} \quad (3)$$

First, successful models of this class are reached when Y^i are either Brownian motions with drift or time changed Brownian motions; see [9, 15], where also some numerical calibration results are shown. Exit times of more general stochastic processes, including stochastic volatility models, are applied to default modeling in [8].

Reduced-form Modeling

Here we start from the classical single-name CDS approach, where the default time is a double stochastic Poisson process (or Cox-process); see **Hazard Rate; Multiname Reduced Form Models** and [5, 6, 11]. In this approach, it is assumed that conditional on a realization of a path of the default intensity, the default time is distributed like the time of the first jump of a time-inhomogeneous Poisson process with this intensity. Typically, the dynamics of resulting credit spreads are closely tied to the dynamics of the default intensity in this approach.

The main challenge here is the incorporation of default dependence. One either has to model common jumps in the spread processes or applies the copula approach exogenously to the default times given from the spread and hazard rates [4, 17]. Recently, an even more reduced approach was developed [1, 7, 18, 19] in which the accumulated losses $(L_t)_{t \geq 0}$ are modeled directly as a stochastic process. The

most general construction (as e.g., in [7]) is to view L as an increasing cadlag pure Jump process with absolute continuous compensator $\nu(dt, dx) = g(t, dx)dt$; see, for example, [10] for the underlying stochastic analysis. This is particularly useful, if one considers options on the spread s^{kth} of a basket swap. Here, the modeling attempt is on L and the single-name modeling is not considered.

Pricing

In order to price basket default swaps, we need the distribution $F_{(k:n)}(t)$ of the time τ^{kth} of the k th default. The k th default time is, in fact, the order statistic $\tau_{(k:n)}$, $k \leq n$, and, in general, we can derive the distribution of the k th order statistics from the multivariate distribution functions [3]. For pricing we also need the survival function:

$$S_{(k:n)}(t) = 1 - F_{(k:n)}(t) \quad (4)$$

The fair spread s^{kth} for maturity T_m is then given by

$$s^{kth} \sum_{i=1}^m \Delta_i B(T_0, T_i) S_{(k:n)}(T_i) = \sum_{i=1}^n (1 - REC_i) \int_{T_0}^{T_m} B(T_0, u) F_{(k:n)}^{kth=i}(du) \quad (5)$$

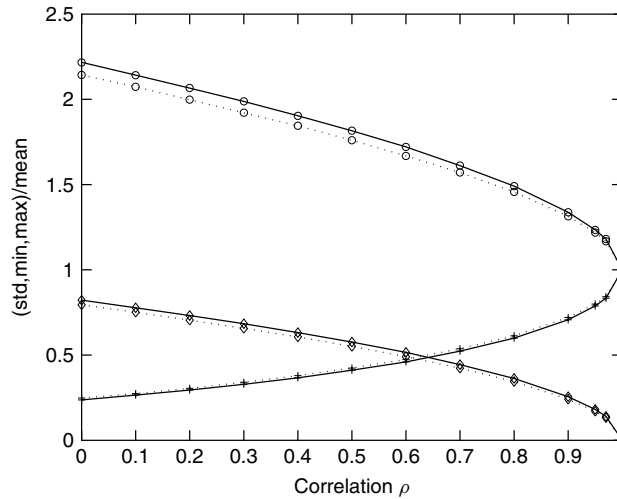


Figure 1 k th-to-default spread versus correlation for a basket with three underlyings: (solid) s^{1st} , (dashed) s^{2nd} , (dashed-dotted) s^{3rd}

The first part is the present value of the spread payments, which stops at τ^{kth} . The second part is the present value of the payment at the time of the k th default. Since the recovery rates might be different for the n underlying names, we have to sum up over all names and weigh with the probability that the k th default happens around u and that the k th defaulted name is just i (we assume that there are no joint defaults at exactly the same time). So $F_{(k:n)}^{kth=i}$ is the probability distribution of the k th order statistic of the default times and that $kth = i$. Figure 1 [3] shows the k th-to-default spreads for a basket of three underlyings with fair spreads $s_1 = 0.009$, $s_2 = 0.010$, and $s_3 = 0.011$, and pairwise equal normal copula correlation on the x -axis. In [16], it was already observed that the sum of the k th-to-default swap spreads is greater than the sum of the individual spreads, that is, $\sum_{k=1}^n s^{kth} > \sum_{i=1}^n s_i$. Both sides insure exactly the same risk, so this discrepancy is due to a windfall effect of the first-to default swap. At the time of the first default, one stops paying the huge spread s^{1st} on the one side but on the plain-vanilla side one stops just paying the spread s_i of the first defaulted obligor i .

References

- [1] Bennani, N. (2005). *The Forward Loss Model: A Dynamic Term Structure Approach for the Pricing of Portfolio Credit Derivatives*. Working paper.
- [2] Bluhm, C., Overbeck, L. & Wagner, C. (2002). *An Introduction to Credit Risk Modeling*, CRC Press/Chapman & Hall.
- [3] Bluhm, C. & Overbeck, L. (2006). *Structured Credit Portfolio Analysis, Baskets and CDOs*, CRCpress/Chapman & Hall.
- [4] Duffie, D. & Gârleanu, N. (2001). Risk and valuation of collateralized debt obligations, *Financial Analysts Journal* **57**, 41–59.
- [5] Duffie, D. & Singleton, K. (1998). *Simulating Correlated Defaults*. Working paper, Graduate School of Business, Stanford University.
- [6] Duffie, D. & Singleton, K. (1999). Modeling term structures of defaultable bonds, *Review of Financial Studies* **12**, 687–720.
- [7] Filipovic, D., Overbeck, L. & Schmidt, T. (2008). *Dynamic Term Structure of CDO-losses*. Working Paper.
- [8] Fouque, J.P., Wignall, B.C. & Zhou, X. (2008). Modeling correlated defaults: first passage model under stochastic volatility, *Journal of Computational Finance* **11**(3), 43–78.
- [9] Hull, J. & White, A. (2001). Valuing credit default swaps II: modeling default correlations, *The Journal of Derivatives* Spring, 12–21.
- [10] Jacod, J. & Shiryaev, A.N. (1987). *Limit Theorems for Stochastic Processes*, Springer.
- [11] Jarrow, R.A., Lando, D. & Turnbull, S.M. (1997). A Markov model for the term structure of credit risk spreads, *Review of Financial Studies* **10**, 481–523.
- [12] Laurent, J. & Gregory, J. (2005). Basket default swaps, cdo's and factor copulas, *Journal of Risk* **7**, 103–122.
- [13] Li, D.X. (1999). The valuation of basket credit derivatives, *CreditMetricsTM Monitor* April, 34–50.
- [14] Li, D.X. (2000). On default correlation: a copula function approach, *Journal of Fixed Income* **6**, 43–54.
- [15] Overbeck, L. & Schmidt, W. (2005). Modeling default dependence with threshold models, *Journal of Derivatives* **12**(4), 10–19.
- [16] Schmidt, W. & Ward, I. (2002). Pricing default baskets, *Risk* **15**(1), 111–114.
- [17] Schoenbucher, P. (2003). *Credit Derivatives Pricing Models: Models, Pricing, Implementation*, Wiley Finance.
- [18] Schönbucher, P. (2005). *Portfolio Losses and the Term Structure of Loss Transition Rates: A New Methodology for the Pricing of Portfolio Credit Derivatives*. Working paper.
- [19] Sidenius, J., Piterbarg, V. & Andersen, L. (2005). *A New Framework for Dynamic Credit Portfolio Loss Modelling*. Working paper.

Related Articles

Collateralized Debt Obligations (CDO); Copulas; Estimation; Copulas in Insurance; Credit Default Swaps; Credit Default Swap (CDS) Indices; Default Time Copulas; Duffie–Singleton Model; Gaussian Copula Model; Hazard Rate; Jarrow–Lando–Turnbull Model; Multiname Reduced Form Models; Intensity-based Credit Risk Models; Reduced Form Credit Risk Models; Structural Default Risk Models.

LUDGER OVERBECK

Collateralized Debt Obligations (CDO)

Collateralized debt obligations (CDOs) can be generically defined as structured products using tranching^a and securitization technology to repackage and redistribute credit risks. Figure 1 symbolically depicts the mechanics of a CDO.

The first forms of CDOs appeared during the 1980s with the repackaging of high-yield bonds such as collateralized bond obligations (CBOs), following hot on the heels of the first collateralized mortgage obligations (CMOs) pioneered by First Boston in the United States in 1983. This technique was later extended to other asset classes such as bank loans (especially leveraged loans). With the advent of the credit derivatives market and the surge in credit default swap (CDS) trading at the beginning of the decade, CDOs became one of the fastest growing segments of the credit market (the so-called structured credit market) and a crucible of financial innovation. In the limited time frame of a few years (2001–2007), CDOs and structured credit arguably became the hottest areas in capital markets and among the greatest fee and trading income generators for investment banks, asset managers, and hedge funds, until the 2007 subprime crisis marked the (temporary?) end of the party.

This article first provides definitions and a typology of CDOs, based on their main characteristics. The second section deals with the main modeling techniques for CDOs. We then dwell upon the impact of the 2007 subprime crisis on the CDO business and look at the evolution of the market and structures in the aftermath of this watershed. Our concluding remarks deal with the future for CDOs in a post-credit crisis world.

Definitions and Typology of CDOs

CDOs cover a large variety of products and structures. The following parameters can be used to define the different types of CDOs.

The Nature of Collateral Assets

The common denominator of CDO transactions was, until 2002–2003, the application of securitization techniques to (credit) assets sourced in the financial markets, such as bonds (CBOs), or from financial institution balance sheets, such as bank loans (collateralized loan obligations (CLOs)). Theoretically, any asset generating recurrent cash flows can be securitized and therefore be used as a collateral to a CDO transaction. What distinguishes CDOs from securitization transactions (asset-backed securities (ABSs)),^b which deal with extended pools of small credit exposures, is that CDO underlying assets can be construed as unitary credit risks and analyzed as such (each CDO underlying asset usually carries an individual credit rating).

In this decade, the range of instruments used as CDO collateral has considerably increased, including securitization issues (CDOs of ABS), other CDOs (CDOs of CDOs), trust-preferred securities (TRUPs), and going as far as hedge funds^c or private equity participations.

In parallel, the rise of credit derivatives (CDSs) has led to the emergence of a new type of products, the synthetic CDOs.^d Instead of “cash” securities, these instruments reference a pool of CDSs, which replicate the risk and cash-flow profile of a bond portfolio. The credit risk is transferred to the special-purpose vehicle (SPV) using CDS technology, which then issues securities backed by this “synthetic” portfolio. What makes synthetic CDOs attractive to structurers and managers is that they avoid the logistics and financial risk of buying in and warehousing securities while a CDO is being constructed and sold to investors. The use of CDSs as reference “assets” for CDOs opened the door to innovative structures and management techniques, which led part of the structured credit business away from traditional securitization and closer to exotic derivative trading as discussed later.

Risk Transfer Mechanism

One must first distinguish credit risk transfer from the collateral portfolio to SPV and, second, from the SPV to capital market investors.

Credit risk transfer from the collateral portfolio to the SPV may happen *via* the following:

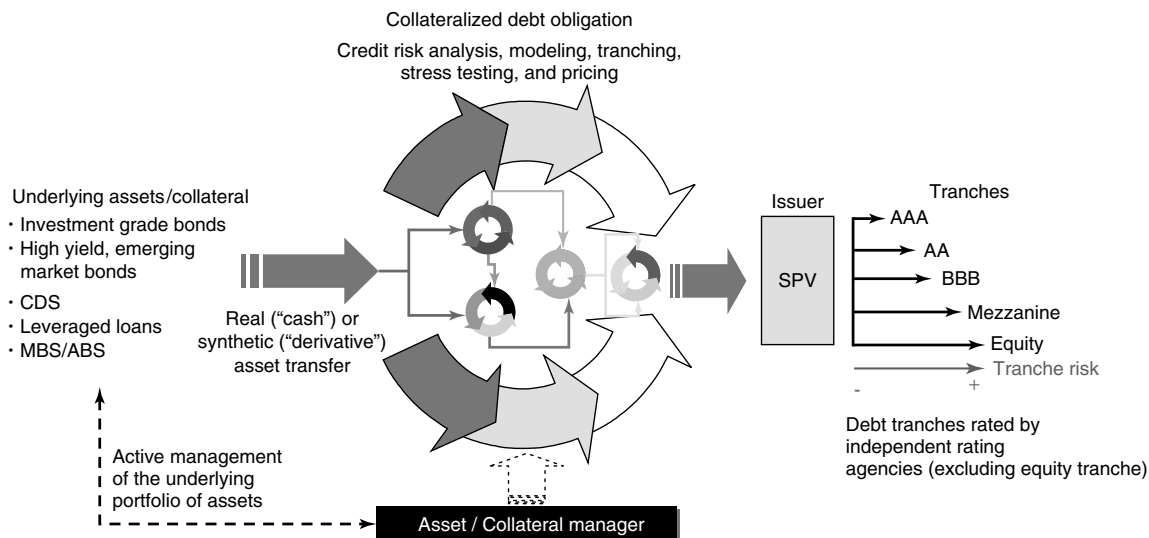


Figure 1 Mechanics of a CDO (Bruyere *et al.* 2005)

- “real” asset acquisition (true sale): “cash CDO” or
- credit derivative technology (or other, e.g., insurance): “synthetic CDO” or collateralized synthetic obligation (CSO).

Risk transfer from the SPV to capital market investors can take the following forms:

- SPV credit-linked note issuance: “funded CDO”;
- credit derivatives (CDSs) sold by the investor to the SPV: “unfunded CDO”; and
- a combination of the above-mentioned: “partially funded CDO”. Most whole capital structure CDOs fall into that category.

Objective of the Transaction

Most CDOs are structured for arbitrage purposes. Arbitrage CDOs are tailor-made investment products, using cash or synthetic technology, created for the benefit of capital market investors. In these transactions, collateral assets are usually sourced in the fixed-income cash or credit derivative markets.

However, a significant part of the CDO market was also driven with the purpose of bank balance sheet management. In such a transaction, the

objective for the sponsor bank is to obtain regulatory or economic capital relief using CDO technology to transfer credit risk to investors. In these transactions, assets or credit risk exposures are typically sourced from the sponsor bank’s own balance sheet.

Static or Managed CDOs

“Static CDOs” are characterized by the fact that the composition of the reference portfolio does not change over the life of the transaction (but for substitutions in a limited number of cases).

At the opposite end of the spectrum, “managed CDOs” (*see Managed CDO*) allow for the dynamic management of the portfolio of collateral assets within a predetermined set of constraints. CDOs are usually managed by a third-party asset manager with credit management expertise. In a managed arbitrage CDO, the asset manager’s objective may be the following:

- to avoid default and ensure timely payment of interest and repayment of principal (“cash-flow CDO”) or
- to optimize the market value of the underlying collateral pool through active management (“market-value CDO”).

“Self-managed CDOs” enable investors themselves to manage the reference portfolio of the CDO they have underwritten.

The following section provides an analysis of the main CDO modeling techniques.

Analysis of CDO Modelling Techniques

Cash-flow CDOs

On the basis of securitization techniques, cash-flow CDOs usually aim at exploiting an arbitrage opportunity between the yield generated by a portfolio of credit assets and that required by investors on the securitized debt, the great majority of which (80–90%) is rated investment grade due to the various credit enhancement mechanisms:

- *Tranching and waterfall*
The creation of several layers of risk (“tranches”) and the sequential allocation of income generated by the collateral portfolio in order of tranche seniority.
- *Subordination*
Losses are absorbed by all junior tranches to a given tranche, thus providing a protection “cushion” (when the CDO is liquidated, the senior creditors have priority over the mezzanine investors, who have priority over the equity holders).
- *Overcollateralization (O/C) and interest cover (I/C) tests*
These act as CDO covenants, leading to the diversification of cash flows toward the early repayment of the most senior tranche if they are breached, thus strengthening the level of subordination.
- *Diversification*
Reference portfolios are diversified in terms of obligor geography and sector, thus limiting the risk of correlated defaults.

Risks and sources of performance in cash-flow CDOs include the following:

- *Default risk*
Underperformance of the underlying portfolio (defaults) leads to a decrease in the amount of assets (and therefore the amount of capital, the equivalent of a write-off in accounting terms) and

in future income streams (since the coupon is no longer being paid on the asset in default) and therefore in the dividend amounts ultimately paid to the equity tranche investors.

- *Portfolio management*
Active trading by the CDO manager may generate losses (which have the same impact as a default) or gains (which are then paid out in dividends or incorporated into the CDO capital, thereby, increasing the subordination level). Generally, the CDO manager is only able to modify the portfolio for a given period (5–7 years, the so-called reinvestment period). He/she must comply with a set of criteria (quality of the portfolio, sector diversification, maturity profile, maximum annual trading allowance, etc.) defined in accordance with the rating agencies.
- *Ramp-up risk*
When a cash CDO is launched, the underlying portfolio cannot be immediately constituted by the manager (essentially to avoid disturbing market liquidity). The portfolio is, therefore, built up over 3–6 months (the ramp-up period). During that time, asset prices may go up and the initial average coupon target for the portfolio might not be attained. In addition, the bank arranging the transaction carries the credit risk of the collateral during the ramp-up period (the so-called “warehousing” risk). To avoid taking too much risk on their balance sheets and allocate capital, banks have been using off-balance sheet vehicles (such as conduits and structured investment vehicles (SIVs)) to park the assets during the ramp-up period. However, as witnessed during the 2007 credit crisis, these defense structures backfired as liquidity dried up and banks were forced to consolidate the vehicles and the security warehouses on their balance sheets.
- *Reinvestment risk*
During the life of the transaction, the manager is regularly led to replace assets and therefore to reinvest part of the portfolio. Market conditions may change and the average coupon level might not be attained. To manage this risk, the manager and the other equity investors usually have an early termination option on the CDO.

Synthetic CDOs: Correlation Products

In the synthetic space, tranching and securitization techniques can also be applied to a portfolio of

CDSs (the so-called whole capital structure (WCS) synthetic CDOs).

However, a watershed appeared with the creation of single-tranche technology, fed by the rise in CDS trading liquidity and advances in credit-modeling expertise. For an investor, any CDO tranche can be considered as a “put spread”^e on the losses of the reference portfolio (the attachment and the detachment points of the CDO tranche being equivalent to the two strike prices of this option combination). Thus, the pricing of a CDO tranche ($x\%$ to $y\%$) can be deduced from the value of the portfolio (i.e., the losses from 0 to 100%) from which the value of the equity tranche (0% to $x\%$) and that of the senior tranche ($y\%$ to 100%) are subtracted. These techniques led to the development of the exotic credit market, which trades on the basis of “correlation”, not unlike the equity derivative market, and volatility.

In a bespoke single-tranche CDO, the arranger usually retains the unsold tranches on its books and dynamically manages their credit risk by selling a fragment (delta) of the notional amount determined for each reference credit entity in the portfolio. This delta must then be readjusted dynamically depending on the changes in the credit spreads. The objective of delta hedging is to neutralize the price variations in the tranche that are linked to changes in the spread of the entities in the underlying portfolio. The delta of a tranche depends upon its seniority and residual maturity. Since deltas are being determined using marginal spread variations, a significant change in spreads will lead to a profit or loss depending upon the convexity of the tranche price (“gamma” in option language). Synthetic CDO arrangers, therefore, not only manage first-order risk levels but also need to monitor their convexity positions.

These hedging mechanisms, however, are not perfect, since they do not deal with second-order risks:

- *Recovery rate in the event of default*
This parameter cannot be inferred from market data. Thus, it is necessary for the dealers to set aside appropriate levels of reserves^f to cover this risk.
- *P&L in the event of default*
Tranche convexity properties are magnified in the event of default and bank positions must be managed accordingly.

- *Correlation (“rho”)*

The pricing and risk management of CDOs are based on correlation rate assumptions. Correlation is determined on the basis of a smile (or skew), which depends mainly on the subordination of the tranche considered. Different correlation rates can thus be given to the attachment and detachment points x and y . This approach by “correlation pairs” is commonly referred to as *base correlation*.

- With the rise of the credit index market,^g CDO arrangers have benefited from new methods for managing their correlation books. Standard tranches are now traded on the main indices in the interbank market, thus providing a benchmark level for the correlation parameters. Until the 2007 credit crisis, liquidity had significantly increased in the CDO tranche market, enabling arrangers to rebalance their books with credit hedge funds and other sophisticated investors.

The Impact of the Subprime Crisis on the Evolution of the CDO Market

With the 2007 subprime and credit crisis, CDOs have come to epitomize the evil of financial innovation. The credit-risk-dispatching mechanism implicit in CDO structures has been broadly accused of fostering the wide spread of poorly understood risks among mainstream capital market investors lured by attractive yields in a low interest-rate environment. To what extent does that charge stand?

ABS CDOs and Subprime Crisis: How Did It Happen?

A key driver for the subprime residential mortgage-backed securities (RMBS) market was the strong demand for subordinated bonds (aka mezzanine tranches, in particular, BBB and BB) from ABS CDO managers.

The reason these bonds were so attractive is that the rating agencies assumed that a portfolio of subordinate mezzanine bonds from various ABS issues would not be highly correlated (much as they assume that corporate bonds from various industries are not highly correlated). Because of this low-correlation assumption, pooling subprime mezzanine bonds into

a CDO structure enabled the CDO manager to create, in essence, new AAA-rated CDO bonds, using only BBB subprime RMBS.

The assumed diversification benefit drove the capital structure of the CDO and explains a large of part of the enormous “misrating” of subprime CDO risk by rating agencies (let alone the rating of the underlying subprime RMBS risk itself).

ABS CDO—A Key Driver of the “Subprime” Demand. The demand from ABS CDOs allowed RMBS originators to lay off a significant portion of the risk. We estimate that \$70 billion of mezzanine subprime RMBS were issued in 2005–2007 *versus* \$200 billion of mezzanine ABS CDOs over the same period. Such notional amount of mezzanine ABS CDOs roughly represents an implied capacity of \$90 billion for mezzanine subprime RMBS investments (over the vintages 2005–2007).^h

This excess demand was filled by synthetic risk (CDS) buckets. The creation of the ABS CDS market multiplied credit risk in the system, allowing for the creation of far more CDOs than the available cash “CDOable” assets. For example, one tranche of a subprime RMBS securitization (nominal \$15.2 million) was referenced in at least 31 mezzanine ABS CDOs (total notional of \$240.5 million).

High-grade ABS CDOs also need to be taken into account. Although the “subprime” demand from these CDOs (roughly \$85 billion) was lower than the nominal of high-grade subprime actually issued (\$230 billion), they fueled the issuance of mezzanine ABS CDOs through the feature of the “inner CDO bucket”. Such a bucket typically had an average size of 20%, allowing CDO arrangers to channel a significant portion of ABS CDO risk. Such “resecuritization” was also facilitated by the existence of CDS on CDOs, further multiplying the credit risk in the system: one tranche of a mezzanine ABS CDO (\$7.5 million nominal) was referenced in at least 17 high-grade ABS CDOs (\$154 million total notional).

At first sight, it would, therefore, be fair to conclude that, since 2005, ABS CDOs have globally absorbed almost every cash-subordinated bond created in the subprime world (and have sold significant protection in synthetic form as well), while traditional cash buyers were largely absent. However, does this mean that the credit risk was effectively transferred to “mainstream” capital market investors?

Anatomy of the ABS CDO Market: Where Did It All Go?. About \$430 billion of ABS CDOs were issued between 2005 and 2007. However, the amount of risk transferred outside the banking system was actually limited because of the following factors:

- investment banks retaining a significant part of super-senior risk, either directly (\$85 billion for the most affected: Citigroup, UBS, Merrill Lynch, Morgan Stanley) or indirectly (by taking on counterparty risk on monoline insurers; \$120 billion notional amounts);
- resecuritization effect though CDO bucket (\$40 billion notional);
- off-balance sheet vehicles, for which banks retained all potential losses (conduits) or part of the losses (SIV, ~\$15 billion of ABS CDO investments); and
- “quasi-”off-balance sheet vehicles, such as money market funds that were subsequently supported by bank capital.

Outside the main banking sector, the most notable “CDO” casualties were either sophisticated insurers (such as AIG) or medium-sized banks (IKB, SachsenLB, and other German Landesbanken).

As a result, it appears that CDOs were primarily a repackaging tool. The main roots of the “subprime” demand stem from abusive off-balance sheet structures (SIVs, conduits) and regulatory capital arbitrages (negative basis trades, long/short badly captured by Value-at-Risk (VaR) models, etc.), both of which resulted in maintaining most of the risk within the banking system while “masking” its true price/value.

One could argue that there was no “real” CDO market for RMBS where rational investors could have sent earlier warning signals (by reducing demand, refusing incestuous features such as CDO buckets within ABS CDOs) and acted as stabilization agents (long-term demand, different investor base than in the underlying RMBS market).

In addition, the derivative market did not perform up to its objectives, as it was created too late (the ABX index, which effectively introduced a greater price transparency) and actually magnified the effects of the mispricing/misrating of RMBS risk.

In conclusion, if the ABS CDO market effectively drove the demand for mezzanine subprime RMBS, its impact on mainstream investors has been limited. In

that respect, it is worth noting that the vast majority of RMBS risk (approximately 82 cents on the dollar) ended up being rated AAA and acquired not by CDOs but by institutions taking advantage of very cheap funding.

How did Other CDO Markets Fare?

Leveraged Loan CLOs. CLOs have suffered from pressure on both the asset and the liabilities sides. Prices of leveraged loans fell in line with the overall credit market, due to technical factors (significant loan overhang resulting from warehouses at the major investment banks) and fundamental fears (increase in default rates, weakly structured leveraged buy-out (LBO) deals). On the liability side, we estimate that negative basis buyers represented 50% of the AAA CLO buyer base, while banks and SIVs/CDOs accounted for 25% and 15%, respectively. The CLO market suffered from the disappearance of such “cheap funding”.

Even though we witnessed an LBO “bubble” (private equity houses taking advantage of the strong CLO bid), the impact of the burst has not been as significant as for the ABS CDO market:

- CLOs were not the sole buyer of leveraged loans.
- They did not suffer from misrating.
- New AAA CLO buyers stepped in (Asian institutions, unaffected banks, insurance companies).

Most of the CLO deals issued in 2008 have been balance sheet driven (cleaning up of warehouses), with simple two-tier structures (AAA and equity), where the AAA tranche (or the equity) is retained by the originating bank.

As the full capital structure execution is challenging and as the sourcing of cash asset is difficult (illiquidity, no warehouse providers for ramp up), the development of single-tranche synthetic CLOs, supported by the growth of the Loan CDS market (ISDA documentation, launch of LCDX and LevX indices), is a key feature of the forthcoming years.

Corporate Synthetic CDOs. With the huge growth in synthetic CDOs, what is commonly referred to as the *structured bid* became a dominant driver of credit spreads. While a combination of mark-to-market losses, rating downgrade risk, and headline risk could have caused investors to unwind positions

in synthetic CDOs, this market segment actually held up well in line with the underlying asset quality (corporate earnings) further supported by the liquidity provided by banks (correlation desks).

Even though the market avoided the “great unwind”, the buying base for these products has essentially gone away, and while some prop desks and hedge funds are still active, the institutional money that provided the liquidity backbone has vanished.

Conclusion: Where Next for CDOs?

The postcrisis CDO market will probably be characterized by a convergence trend toward the mechanics of the corporate synthetic market, which has proved more efficient and resilient for the distribution of credit risk:

- the development of index and index tranches (transparent and traded correlation) fueling liquidity;
- less reliance on rating agencies and more in-house due diligence on assets; and
- a return to balance-sheet-driven transactions.

The main challenges for the CDO market include the following:

- restoring investor confidence in the benefit of structured products by providing better transparency and liquidity;
- addressing the AAA funding issue (now that SIVs and conduits have been dissolved); and
- overcoming the discrepancies in accounting treatment.¹

Once the dust has settled, we expect securitization and CDO transactions to come back on the basis of more transparent and rational fundamentals.

End Notes

^aTranching is the operation by which the cash flows from a portfolio of assets are allocated by order of priority to create various layers (“tranches”), from the less risky (“senior” tranche) to the most risky (“first loss” or “equity” tranche). Tranching technology is usually performed using rating agency guidelines in order to ensure that the senior tranche attracts the most favorable rating (triple-A).

^b Asset-backed securities are securities representing a securitization issue. The ABS market covers mortgage-backed securities (residential and commercial), consumer (credit card, student loans, auto loans), and commercial loans (trade receivables, leases, small business loans, etc.).

^c Collateralized fund obligations.

^d “Synthetic” in as far as the mechanism for transferring risk is synthetic, using a derivative.

^e Combination of two put options on the same underlying asset, at two different strike prices.

^f Usually in the form of bid–ask spreads.

^g iTraxx for the European market and CDX.NA for the US market.

^h On the basis of the following assumptions: 50% of the portfolio allocated to subprime, of which 60% to the precedent vintage.

ⁱ While a cash CDO (or any cash bond) can be accounted for as “available for sale” by banks and insurers (meaning that its price volatility will directly impact the equity base of the investor), the valuation of an equivalent synthetic products impacts the income (P&L) of the investor.

Reference

- [1] Bruyere, R., Cont, R., Copinot, R., Jaeck, Ch., Fery, L. & Spitz, T. (2005). *Credit Derivatives and Structured Credit: A Guide for Investors*, Wiley.

Related Articles

Base Correlation; Basket Default Swaps; CDO Square; CDO Tranches; Impact on Economic Capital; Collateralized Debt Obligation (CDO) Options; Credit Default Swaps; Default Barrier Models; Forward-starting CDO Tranche; Managed CDO; Multiname Reduced Form Models; Nested Simulation; Random Factor Loading Model (for Portfolio Credit); Reduced Form Credit Risk Models; Special-purpose Vehicle (SPV); Total Return Swap.

RICHARD BRUYERE & CHRISTOPHE JAECK

Forward-starting CDO Tranche

At the core of any CDO pricing model is a mechanism for generating dependent defaults. If a simple factor structure is used to join their marginal distributions, the default times of the underlying credits are independent conditionally on the realization of the common factor(s). This conditional independence of defaults is very useful because it allows one to use quasi-analytical algorithms to compute the term structure of expected tranche losses, which is the fundamental ingredient for the valuation of a synthetic CDO.

Because of their analytical tractability, conditionally independent models have become a standard in the synthetic CDO market. In the next section, we review the one-factor Gaussian-copula model, which has played a dominant role since the early days of single-tranche trading.

The Gaussian-copula Model

In the one-factor Gaussian-copula framework, the dependence of the default times is Gaussian, and is therefore completely specified by their correlations. In this model, given a particular realization of a normally distributed common factor Y , the probability that the j th credit defaults by time t is equal to

$$\pi_{j,t}(Y) = N\left(\frac{D_{j,t} - \beta_j \cdot Y}{\sqrt{1 - \beta_j^2}}\right), \quad j = 1, 2, \dots, M \quad (1)$$

where $N(\cdot)$ denotes the standard Gaussian distribution function, the vector $\{\beta_j\}$ determines the correlations of the default times, $\{D_{j,t}\}$ are free parameters chosen to satisfy

$$p_{j,t} = \int_Y \pi_{j,t}(Y) dN(Y) \quad (2)$$

and $p_{j,t}$ are the (unconditional) probabilities that name j defaults by time t . Importantly, for the CDO model to price the underlying credit default swap (CDS) correctly, $p_{j,t}$ must be backed out from the term structure of observable CDS spreads.

Given a realization of the Gaussian factor Y , the M individual credits are independent, and a simple recursive procedure [2] can then be employed to recover the conditional loss distribution of the underlying portfolio, as well as the loss distribution of any particular tranche of interest. Once we know how to compute the loss distribution of a tranche for a given realization of the common factor, it is straightforward to take a probability-weighted average across all possible realizations of Y and thus recover the unconditional loss distribution of the tranche.

Repeating this procedure for a grid of horizon dates and interpreting the expected percentage loss up to time t as a “cumulative default probability”, we can price the tranche using exactly the same analytics that we would use for pricing a CDS. More precisely, we can define the “tranche curve” as the term structure of expected surviving percentage notional of the tranche, that is,

$$Q(t) = 1 - E\left[\frac{[L_t - U]^+ - [L_t - (U + V)]^+}{V}\right] \quad (3)$$

where L_t is the number of loss units experienced by the reference portfolio by time t , U is the number of loss units that the tranche can withstand (attachment), and V is the number of loss units protected by the tranche investor. Then the two legs of the swap can be priced using

$$\text{Premium} = cN \sum_{i=1}^T \Delta_i Q(t_i) B(t_i) \quad (4)$$

$$\text{Protection} = N \sum_{i=1}^T B(t_i) (Q(t_{i-1}) - Q(t_i)) \quad (5)$$

where c is the annual coupon paid on the tranche, N is the notional of the tranche, t_i , $i = 1, 2, \dots, T$ are the coupon dates, Δ_i , $i = 1, 2, \dots, T$ are accrual factors, and $B(t)$ is the risk-free discount factor for time t . Notice that, for ease of notation, we have used the coupon dates t_i , $i = 1, 2, \dots, T$ to discretize the timeline for the valuation of the protection leg.

Pricing of Reset Tranches

Let us define a reset tranche as a path-dependent tranche whose attachment and/or width are reset at

a predetermined time (the reset date) as deterministic functions of the random amount of losses incurred by the reference portfolio up to that time. Notice that forward-starting tranches and tranches whose attachment point resets at a future date both belong to this class.

Pricing a Reset Tranche

Let t_s denote the reset date, λ_j , $j = 1, 2, \dots, M$, the number of loss units produced by the default of the j th name, $\lambda = \sum \lambda_j$ the maximum number of loss units that the portfolio can suffer, $p(\omega)$ the probability today that the reference portfolio incurs exactly ω loss units by the reset date t_s .

A reset tranche can be defined by the vector $\{t_T, t_s, U, V, U(\omega), V(\omega)\}$ where $U(\omega) \geq \omega$ is the attachment point of the tranche (in loss units) after the reset date, and $V(\omega)$ is the number of loss units protected by the tranche investor after the reset date. We can price the two legs of this swap as follows:

Premium

$$= cN \sum_{\omega=0}^{\lambda} p(\omega) \sum_{i=1}^T \Delta_i Q(t_i; \omega) B(t_i) \quad (6)$$

Protection

$$= N \sum_{\omega=0}^{\lambda} p(\omega) \sum_{i=1}^T B(t_i) (Q(t_{i-1}; \omega) - Q(t_i; \omega)) \quad (7)$$

where we have defined the *conditional* tranche curve $Q(t; \omega)$, $t_0 \leq t \leq t_T$, as

In words, the conditional tranche curve $Q(t; \omega)$ represents the (risk-neutral) expected percentage surviving notional of the tranche at time t , conditional on the event that the reference portfolio experiences a cumulative loss of ω units up to the reset date.

Equally, we can write down the valuation in terms of the *unconditional* tranche curve

$$Q(t) = \sum_{\omega=0}^{\lambda} p(\omega) \cdot Q(t; \omega) \quad (9)$$

and thus obtain the familiar equations

$$\text{Premium} = cN \sum_{i=1}^T \Delta_i Q(t_i) B(t_i) \quad (10)$$

$$\text{Protection} = N \sum_{i=1}^T B(t_i) (Q(t_{i-1}) - Q(t_i)) \quad (11)$$

However, while the unconditional tranche curve for $t_0 \leq t \leq t_s$ reduces to the standard tranche curve defined in the section The Gaussian-copula Model,

$$\begin{aligned} Q(t) &= \sum_{\omega=0}^{\lambda} p(\omega) \cdot Q(t; \omega) = 1 - \sum_{\omega=0}^{\lambda} p(\omega) \\ &\times E \left[\frac{[L_t - U]^+ - [L_t - (U + V)]^+}{V} \middle| L_{t_s} = \omega \right] \\ &= 1 - E \left[\frac{[L_t - U]^+ - [L_t - (U + V)]^+}{V} \right] \end{aligned} \quad (12)$$

$$\begin{aligned} Q(t; \omega) &= T(t, \omega) \cdot \left(1 - E \left[\frac{[L_t - U(t; \omega)]^+ - [L_t - (U(t; \omega) + V(t; \omega))]^+}{V(t; \omega)} \middle| L_{t_s} = \omega \right] \right), \\ T(t, \omega) &= 1 - 1_{\{t > t_s\}} \frac{[\omega - U]^+ - [\omega - (U + V)]^+}{V}, \\ U(t; \omega) &= \begin{cases} U, & t \leq t_s \\ U(\omega), & t > t_s \end{cases}, \\ V(t; \omega) &= \begin{cases} V, & t \leq t_s \\ V(\omega), & t > t_s \end{cases} \end{aligned} \quad (8)$$

the unconditional tranche curve for $t_s < t \leq t_T$

$$Z_{v_1, v_2}^0 = 0 \text{ otherwise} \quad (15)$$

$$\begin{aligned} Q(t) &= \sum_{\omega=0}^{\lambda} p(\omega) \cdot Q(t; \omega) \\ &= \sum_{\omega=0}^{\lambda} p(\omega) \cdot T(t, \omega) \cdot \left(1 - E \left[\frac{[L_t - U(\omega)]^+ - [L_t - (U(\omega) + V(\omega))]^+}{V(\omega)} \middle| L_{t_s} = \omega \right] \right) \end{aligned} \quad (13)$$

incorporates the added complexity of the path-dependent valuation.

Deriving the Conditional Tranche Curve

Our discussion so far leaves open the problem of constructing the conditional tranche curve. From the previous discussion, it should be clear that to achieve this goal we need to be able to compute conditional expectations of the form $E[f(L_{t_u}, \omega) | L_{t_s} = \omega]$ for some function f . In this section, we present a two-dimensional recursive algorithm for computing the joint distribution of cumulative losses at two different horizons, which in turn allows us to compute the conditional expectations that we need. The methodology is conceptually similar to the one introduced by Baheti *et al.* [3] for pricing “squared” products.

As anticipated, we assume that the underlying default model exhibits the property of conditional independence. We exploit this by conditioning our procedure on a particular realization of a common factor Y . We first discretize losses in the event of default by associating each credit with the number of loss units that its default would produce: we indicate by λ_j the integer number of loss units that would result from the default of name j . Next, we construct a square matrix (Z_{v_1, v_2}) whose sides consist of all possible loss levels for the reference portfolio, that is, $(0, 1, \dots, \lambda)$. In this matrix, we store the joint probabilities that the reference portfolio incurs v_1 loss units up to time t_s and v_2 loss units up to time t_u , with $t_u \geq t_s$. By definition of cumulative loss, the matrix must be upper triangular, that is,

$$Z_{v_1, v_2} = 0 \text{ if } v_2 < v_1 \quad (14)$$

For the nontrivial elements where $v_2 \geq v_1$, we set up the following recursion. We first initiate each state (recursion step $j = 0$) by setting

$$Z_{v_1, v_2}^0 = 1, \text{ if } v_1 = 0 \text{ and } v_2 = 0$$

We preserve the notation adopted during our description of the Gaussian-copula model and denote by $\pi_{j,t}(Y)$ the probability that name j defaults by time t , conditional on the market factor taking value Y . Now we feed one credit at a time into the recursion and update each element according to the following:

If $v_1 \geq \lambda_j$, then

$$\begin{aligned} Z_{v_1, v_2}^j &= (1 - \pi_{j,u}(Y)) \cdot Z_{v_1, v_2}^{j-1} \\ &\quad + \pi_{j,s}(Y) \cdot Z_{(v_1 - \lambda_j), (v_2 - \lambda_j)}^{j-1} \\ &\quad + (\pi_{j,u}(Y) - \pi_{j,s}(Y)) \cdot Z_{(v_1), (v_2 - \lambda_j)}^{j-1} \end{aligned} \quad (16)$$

If $v_2 < \lambda_j$, then

$$Z_{v_1, v_2}^j = (1 - \pi_{j,u}(Y)) \cdot Z_{v_1, v_2}^{j-1} \quad (17)$$

If $v_1 < \lambda_j \leq v_2$, then

$$\begin{aligned} Z_{v_1, v_2}^j &= (1 - \pi_{j,u}(Y)) \cdot Z_{v_1, v_2}^{j-1} \\ &\quad + (\pi_{j,u}(Y) - \pi_{j,s}(Y)) \cdot Z_{(v_1), (v_2 - \lambda_j)}^{j-1} \end{aligned} \quad (18)$$

After including all the issuers, we set

$$(Z_{v_1, v_2}) = (Z_{v_1, v_2}^M) \quad (19)$$

The matrix (Z_{v_1, v_2}) now holds the joint loss distribution of the reference portfolio at the two horizon dates t_s and t_u , conditional on the realization of the market factor Y , and we can numerically integrate over the common factor to recover the unconditional joint loss distribution. Using the joint distribution of losses at different horizons, it is then straightforward, for any function $f(\cdot)$, to compute conditional expectations of the form $E[f(L_{t_u}, \omega) | L_{t_s} = \omega]$, which is how we construct the conditional tranche curve.

Comments

We have presented a simple methodology for quasi-analytically pricing a class of default-path-dependent

tranches. The proposed methodology is general in the sense that it can be easily applied to any model with conditionally independent defaults, including “implied copula” models fitted to liquidly traded tranches as in the Hull–White [4] model. The algorithm is useful because fast pricing of reset tranches allows one to obtain a variety of Greeks that are essential for effective risk management.

As observed by Andersen [1], however, some caution is necessary when pricing instruments whose valuation is sensitive to the joint distribution of cumulative losses at different horizons. Liquidly traded tranches only contain information about marginal loss distributions and tell us nothing about their dependence. Implying a default time copula from these prices, therefore, implicitly contains an arbitrary assumption about intertemporal dependencies, and it is easy to verify that different implied copulae

that fit observable prices equally well may produce significantly different valuations for path-dependent instruments.

References

- [1] Andersen, L. (2006). Portfolio losses in factor models: term structures and intertemporal loss dependence, *Journal of Credit Risk* **4**, 71–78.
- [2] Andersen, L., Sidenius, J. & Basu, S. (2003). All your hedges in one basket, *Risk* November, 67–72.
- [3] Baheti, P., Mashal, R., Naldi, M. & Schloegl, L. (2005). Squaring factor copula models, *Risk* June, 73–76.
- [4] Hull, J. & White, A. (2006). *The Perfect Copula*. Working Paper, University of Toronto.

PRASUN BAHETI, ROY MASHAL & MARCO
NALDI

CDO Square

A CDO-of-CDO, or CDO square (CDO²), is a type of collateralized debt obligation (CDO) that has CDO tranches as reference assets. The CDO² market is a natural extension of the CDO market. The concept of CDO² was pioneered by the ZAIS group, when they launched ZING I in 1999, focusing on Euro CDO assets [2]. Recognizable growth of the CDO² market, particularly in the United States, was fueled by the excessive volume growth of the CDO market in the new millennium. In 2004, the situation of tightening credit spreads and a stable credit outlook shifted investor and dealer interest toward more structured credit basket products in the search for yield and tailored risk-return profiles. The repackaging of CDO tranches *via* the CDO² technology allows the dealer to manage the risk and capital cost of (residual) trading book positions. As in the case of CDOs, the dealer can exploit the rating alchemy, that is, the difference between traded and historical default probabilities as well as default correlations, to generate positive carry strategies. The investor will benefit from a more diversified reference portfolio, generally higher yield than similarly rated corporate debt, and the double-layer subordination effect.

We distinguish three main CDO² transaction types: cash CDO², synthetic CDO², and hybrid CDO². All transaction types can either refer to a static portfolio of CDOs or can be combined with an active management of the reference CDO portfolio. In a cash or cash-flow CDO², the reference assets are existing cash CDOs, which typically provide the funds to pay CDO² investors. In a synthetic CDO², the credit risk is generated synthetically, for example, *via* unfunded CDO tranche swaps. Hybrid CDO²s appeared with the rise of the structured finance CDO market and comprise both elements. Typically, in such transactions, the major portion (ca. 80–90%) of the reference portfolio is cash asset-backed security (ABS) exposure and the remainder is synthetic CDOs. It is quite common to use a special purpose vehicle (SPV) overlay for cash CDO² and hybrid CDO².

Mechanics of a Synthetic CDO²

A synthetic CDO² tranche follows the same mechanics as an ordinary CDO tranche (*see Collateralized Debt Obligations (CDO)*), with the only difference that its reference portfolio is made up of CDO tranches. This portfolio is called the *outer or master portfolio*. CDO tranches are determined by their corresponding reference portfolio plus an attachment (or subordination) level and detachment (or exhaustion) level with regard to aggregated credit losses. Hence, we refer to the loss attachment and detachment of the outer portfolio as *outer attachment* and *outer detachment*. Similarly, each CDO tranche of the outer portfolio is described by a corresponding inner reference portfolio and an inner attachment and detachment level (compare with Figure 1 for a schematic description). The inner reference portfolios often overlap and include some of the same reference assets.

Inner attachment and detachment levels as well as the reference notional of assets in inner portfolios are quite often of comparable size. Typically, we find ca. 50–150 assets per inner portfolio and ca. 5–10 inner CDO tranches, which generally translate into a total of ca. 250–500 different reference assets.

CDO² investors benefit from two layers of subordination. First, we must have a considerable number of default events with associated loss rates to exceed the subordination of at least one inner CDO tranche. This will trigger losses on the outer portfolio. However, only if the subordination of the outer CDO tranche is exhausted, we will recognize CDO² losses. The mathematical description of the aggregated CDO² loss $L_{\text{out}}(t)$ for any future date t during the contract term reflects the double-layer effect.

First, the inner portfolio losses $L_j(t)$ have to be determined *via*

$$L_j(t) = \sum_{i=1}^N N_{ij} \cdot (1 - R_i) \cdot 1_{\{\tau_i \leq t\}} \quad (1)$$

where N_{ij} is the notional of assets $i = 1, \dots, N$ in the inner reference portfolios $j = 1, \dots, M$, R_i denotes the asset-specific recovery rate and $1_{\{\tau_i \leq t\}}$ is the (stochastic) default indicator function for the default time τ_i of reference asset i . Second, the inner portfolio losses $L_j(t)$ have to be transformed into

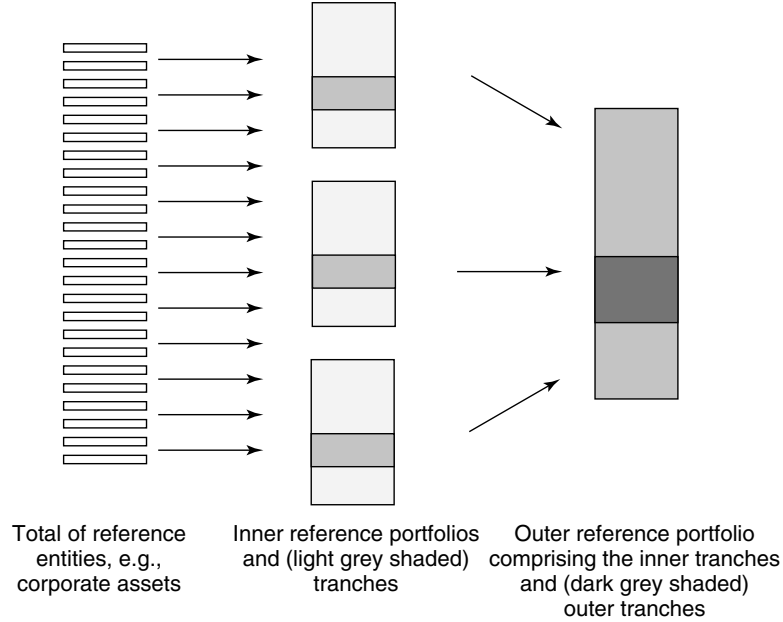


Figure 1 Schematic CDO² description

inner CDO tranche losses $L_{\text{inn},j}(t)$ (see **Collateralized Debt Obligations (CDO)**)

$$L_{\text{inn},j}(t) = \min[D_j - A_j, \max[L_j(t) - A_j, 0]] \quad (2)$$

where A_j and D_j denote the inner attachment and exhaustion level of the corresponding inner reference portfolio j . Third, the outer tranche or CDO² tranche loss can be computed as

$$L_{\text{out}}(t) = \min[D_{\text{out}} - A_{\text{out}}, \max[L_{\text{tot}}(t) - A_{\text{out}}, 0]] \quad (3)$$

where A_{out} and D_{out} denote the attachment and exhaustion points of the outer tranche and $L_{\text{tot}}(t) = \sum_{j=1}^M L_{\text{inn},j}(t)$ the sum of inner tranche losses.

Risk Analysis and Pricing

The limited universe of liquid and actively traded reference assets naturally yields overlaps in inner reference portfolios; in other words, reference assets tend to occur in more than one real-life inner reference portfolio. This causes the CDO² loss distribution to display fatter tails on both ends, since the

(non)occurrence of an isolated default event might simultaneously affect several inner reference portfolios, thereby displaying a leveraged effect [1]. This impact is even more pronounced in case of thin tranches and understood as cliff risk of CDO²s. Moreover, the double-layer tranche technology generally amplifies correlation sensitivities: an increase in the asset correlation yields a higher increase of correlation between affected inner CDO tranches. In summary, overlap and correlation are the main risk drivers of a CDO² tranche. In addition, the described effects considerably increase the impact of other risk drivers such as changing credit spreads (respectively, changing default probabilities) and changing recovery rates.

The key ingredient to pricing is the stochastic evaluation of the accumulated CDO² tranche loss $L_{\text{out}}(t)$ as determined in the previous paragraph. This requires the consistent use of a multivariate credit (default) model. Since no market standard has been developed yet owing to the lack of truly observable correlation information, the necessity and benefit of appropriate scenario models are highlighted in this article. The rating agencies Moody's, Standard & Poor's, and Fitch have consistently adapted their

CDO rating technology to the CDO² case. In particular, the rating technology comes with a look-through capability to underlying assets of inner reference portfolios. However, the look-through capacity stops with ABS-type assets that are modeled as a single asset.

References

- [1] Kakodkar, A., Galiani, S., Jonsson, J.G. & Gallo, A. (2006). *Credit Derivatives Handbook 2006—Vol. 2 A Guide to the Exotics Credit Derivatives Market, Credit Derivatives Strategy*, Merrill Lynch, New York.

- [2] Smith, D. (2003). CDOs of CDOs: art eating itself? in *Credit Derivatives: The Definite Guide*, J. Gregory, ed, Risk Books, London, pp. 257–279.

Related Articles

Collateralized Debt Obligations (CDO); Managed CDO; Structured Finance Rating Methodologies.

HANS-JÜRGEN BRASCH

Leveraged Super-senior Tranche

A leveraged super-senior (LSS) note is a structure that allows investors to take a leveraged exposure to the super-senior (SS) part of a collateralized debt obligation (CDO). This provides an enhanced level of return while typically maintaining a AAA rating. Leverage is achieved by posting an initial collateral amount that is less than the notional of the underlying SS tranche. All credit losses to the investor are capped at the collateral amount, but the coupon is paid on the full notional. Early unwind clauses are typically included in the trade to mitigate the risk to the issuer that losses exceed the collateral amount. Compared to a standard SS tranche the investor is exposed to mark-to-market (MTM) risk as well as credit risk, that is, for certain market moves his/her principal could be reduced even if no credit losses have occurred owing to a forced unwind. The issuer faces a so-called gap risk, the risk that the MTM of the tranche will fall below the initially posted collateral amount before a trade unwind can take place.

Super-senior Swap

In an (unleveraged) SS swap transaction an investor (protection seller) will take exposure to credit losses on the SS tranche of a CDO (see **Collateralized Debt Obligations (CDO)**). This means that in return for a regular fee (or coupon), the investor will make good all losses on the underlying reference portfolio that exceed the attachment amount, but are below the detachment amount. For an SS tranche the attachment point will be higher than what is required for achieving a AAA rating (see **Credit Rating**). The spread of the tranche is the fee that values the swap at 0.

Leveraged Super-senior (LSS) Swap

Since the risk of experiencing any losses on an SS tranche is remote, the spread for a standard SS tranche is low. In particular, it is lower than for other AAA-rated securities, which limits the attractiveness of the transaction to investors. LSS structures became

popular in 2005, when spreads on the 22–100%, 5-year, iTraxx index (see **Credit Default Swap (CDS) Indices**) tranche were around 5 bps.

Issuers of LSS notes are typically investment banks. When arranging a CDO transaction the issuer needs to be able to sell the entire capital structure, as otherwise he/she will be left with the remaining risk. For the reasons mentioned above, it can be harder to sell SS risk than risk on mezzanine and senior tranches. The LSS transaction allows the issuer to repackage the SS risk in a way that is attractive to the investor.

In an LSS transaction, the investor will invest in an SS tranche with notional N (referred to as the *reference tranche*). However, he/she will only post an initial collateral amount X (also referred to as the *participation amount*) that is less than the notional amount of the tranche. Any credit losses to the investor will reduce the collateral and losses are thus capped at the collateral amount, X . However, the investor still receives a coupon on the full notional, N , of the reference SS tranche. The ratio, N/X , is referred to as the *leverage*. If losses reach the collateral amount, the trade will terminate without any further cash flows. The typical structure of the LSS trade can be seen in Figure 1. We compare credit losses on the LSS and a standard SS in Figure 2.

The LSS structure provides an increased coupon to the investor compared to investing the collateral amount X in a standard SS tranche. The reason for this is that the investor takes on MTM risk over and above his/her credit risk as is described below.

The issuer of an LSS note (protection buyer) is only covered for losses up to the collateral amount X . However, he/she will typically hedge his/her position by selling protection on the corresponding standard SS tranche (cf. the section Hedging and Risks) where he/she is liable for all the losses on the tranche up to the full notional amount N . Thus, the issuer needs to mitigate the risk that losses exceed the collateral level. This is done *via* the inclusion of a trigger mechanism: As soon as a predefined trigger level is reached the trade will unwind at the MTM of the reference tranche capped at X . The trigger should thus be set such that the MTM on unwind will be less than X .

We note that in some transactions the investor has the option of posting more collateral upon trigger to avoid an unwind. In this case, the investor will continue to be paid the coupon of the original

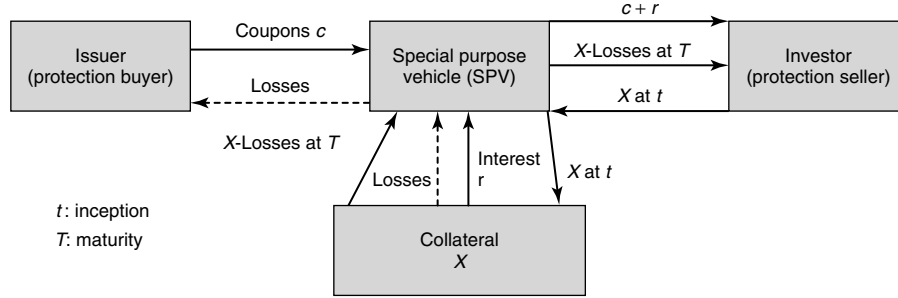


Figure 1 This figure shows the standard structure of cash flows in an LSS transaction

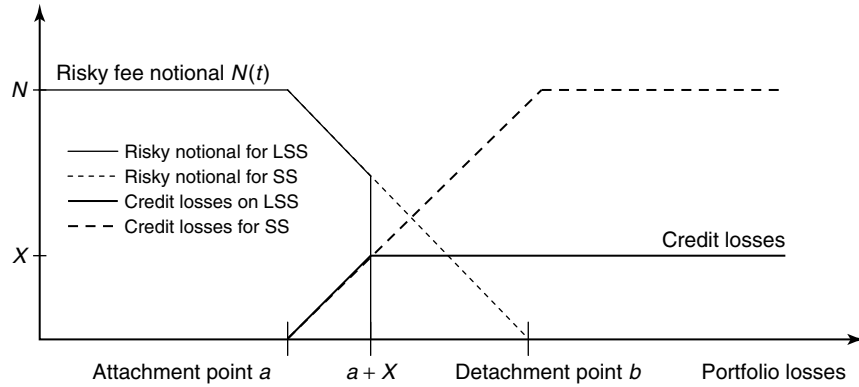


Figure 2 This figure shows the credit losses to the investor and the risky fee notional, $N(t)$, as a function of the portfolio losses. The coupon amount paid to the investor at a coupon date t_i is $s \cdot N(t_i)$, where s is the spread. For comparison we include both the behavior of the LSS and the reference unleveraged super-senior (SS)

transaction. This means that it is never optimal for the investor to deliver since after posting more collateral he/she will be liable for more losses without receiving a higher coupon in compensation. It is more favorable to reinvest in a new LSS transaction.

We describe the three main types of trigger mechanisms below. There is a trade-off between how well the trigger can approximate the MTM of the trade and how easy it is to objectively assess whether the trigger has been breached.

• Loss Trigger

A loss trigger is breached when the amount of portfolio notional lost owing to defaults exceeds the trigger level. This is the easiest trigger to monitor as the loss amounts can be objectively determined. However, the loss provides an imperfect approximation for the MTM of the tranche. In particular, if spreads

widen, the value of the LSS can drop severely from the point of view of the investor even in the absence of any defaults. This poses a risk to the issuer since at the time of trigger the MTM of the tranche could have dropped below the collateral amount.

• Spread Trigger

Spread triggers are based on the average spread of the underlying portfolio. Trigger levels can be defined as a function of the time to maturity and the level of losses in the portfolio. This provides a much better proxy to the MTM of the tranche than the loss trigger. For some standard portfolios, for example, iTraxx or CDX (*see Credit Default Swap (CDS) Indices*), the value of the average spread can also be assessed using publicly available information and is hence unambiguous. Often, however, the LSS is based on bespoke portfolios. In this case, the valuation of the

SS spreads will have to rely on models, for which there is no universally agreed methodology.

• Mark-to-market Trigger

The MTM trigger is based on the MTM of the reference (unleveraged) SS tranche. Clearly, if the MTM trigger is set below the collateral level the issuer ensures that the collateral will cover the unwind payment (up to gap risk, cf. the section Hedging and Risks). The disadvantage is that the MTM trigger is the hardest to assess objectively. Typically, the MTM for a tranche is not quoted, and hence one has to rely entirely on (complex) models for valuation.

Hedging and Risks

If the trigger mechanism guaranteed that upon unwind the issuer will receive the full MTM of the reference swap then this swap would provide a perfect hedge for the LSS. The coupon amount the investor would receive would be the same, as if he/she had invested the full notional amount N in an SS swap transaction.

However, there are two reasons why the trade can unwind without recovering the full MTM of the hedge:

- Typically, there will be a delay between a trigger breach and the actual unwind of the hedge. In this period, there is the risk that the MTM will drop below the collateral amount. The issuer then has to make good the difference ($MTM - X$). This is the so-called *gap risk*: The issuer is exposed to large and sudden increases in the value of SS protection or equivalently increases in the spread.
- Even in the absence of a trigger breach, the LSS will unwind if the SS tranche losses have wiped out the collateral. Since the collateral is 0 in this case, the issuer will have to pay the full MTM of the hedge to unwind his/her position. However, this scenario is unlikely, since the trigger should be set so that a trigger event occurs before the collateral has been reduced to 0.

The investor in an LSS transaction faces MTM risks as well as credit risks associated with SS tranche losses. In the case of a trigger event, the investor will be forced to unwind his/her position and will lose part or all of his/her principal as he/she realizes his/her

MTM losses (unless he/she posts more collateral). Unless dealing with a loss trigger, a trigger breach can happen even if the investor has not incurred any actual credit losses, for example, if there is a dramatic rise in spreads.

Valuation

The valuation of LSS transactions poses additional challenges to that of pricing a standard SS tranche. This is because the unwind feature means that we need to be able to value the risk of possible MTM losses to the investor and the issuer. Hence, we need to model the joint behavior of MTM and the portfolio losses. This is a dynamic problem that requires more than knowledge of the marginal loss distributions needed for standard tranche pricing.

There are two main candidates for dynamic credit models that can, in principle, be used to value an LSS transaction:

- low-dimensional models of the portfolio loss process;
- dynamic models of all single name spreads in the portfolio (*see Duffie–Singleton Model; Multiname Reduced Form Models*).

Modeling and valuation of the LSS product is not only important for the issuer and the investor but also for assessing the rating of the note. This depends not only on the probability of experiencing credit losses but also on the probability of having a trigger event. Rating agencies (*see Credit Rating*) use in-house models for this as described in, for example, [1, 2].

Model-independent Bounds

Some model-independent bounds for the value of the LSS can be derived. We discuss this from the perspective of the issuer who is long protection. Let us denote the spread of a standard tranche with attachment point a and detachment point b by $S_{a,b}$.

The spread of the corresponding leveraged tranche with collateral amount x will be denoted by $S_{a,b,x}$. Note that the leverage amount α is given by $\alpha = (b - a)/x$. The most basic bound we can then write down is

$$S_{a,b} \leq S_{a,b,x} \leq \alpha \cdot S_{a,b} \quad (1)$$

This means the following:

- The spread of the leveraged tranche is less than the leverage amount times the spread of the unleveraged tranche. This is because the issuer has additional unwind and gap risk. The difference

$$\alpha \cdot S_{a,b} - S_{a,b,x} \quad (2)$$
 is the *gap charge*.
- The spread of the leveraged tranche is greater than that of the unleveraged tranche. This is because of the trigger mechanism, which allows the issuer to recover the MTM of the unleveraged tranche up to the collateral amount.

We can also give a more stringent floor for the LSS value. To this, we introduce the fee leg value $F_{a,b}$ and contingent (or loss) leg value $C_{a,b}$ of a tranche (see **Collateralized Debt Obligations (CDO)**). The (positive) value of the fee leg is the expected value of all coupon payments paid on the risky notional. The contingent leg is the (positive) expected value of any loss payments. We now have the following bounds:

$$C_{a,a+x} \leq C_{a,b,x} \quad (3)$$

This is because on the contingent leg, the issuer will at least recover losses up to $a + x$. He can effectively recover more on unwind since he/she receives the MTM of the unleveraged tranche.

We also have

$$F_{a,b} \geq F_{a,b,x} \quad (4)$$

This corresponds to the fact that on the fee leg the issuer will at most pay the fees of the unleveraged reference tranche. He/she might effectively pay less if there is an unwind not due to the trigger such that no MTM exchange takes place.

For a more rigorous discussion we refer to [3].

References

- [1] Chandler, C., Guadagnuolo, L. & Jobst, N. (2005). CDO spotlight: Approach to rating super senior CDO notes, in *Standard & Poors Structured Finance*, Standard & Poor's a Division of the McGraw-Hill Companies, Inc., New York.
- [2] Osako, C., Perkins, W. & Kissina, I. (2005). Leveraged super-senior credit default swaps, Fitch Ratings Structured Finance July.
- [3] Gregory, J. (2008). A trick of the credit tail, *Risk* **21**(3), 88–92.

Related Articles

Forward-starting CDO Tranche.

MATTHIAS ARNSDORF

Managed CDO

General Definition

A managed (or “active”) *collateralized debt obligation* (CDO) is a large-scale securitization transaction that is actively arranged and administered to unbundle, transform, and diversify financial risks from a dynamic reference portfolio of one or more credit-sensitive asset classes (associated with different creditors, countries, and/or industry sectors). Although the type(s) of asset(s) in the reference portfolio are known and fixed through the life of the CDO, the underlying collateral of a managed CDO is variable.

Types of Managed CDOs

In general, CDO Managers operate under investment guidelines that are defined in the governing documents of the CDO transaction. Managers adjust the investment exposure over time to meet a pre-specified risk–return profile and/or achieve a certain degree of diversification in response to changes in risk sensitivity, market sentiment, and/or timing preferences. These guidelines specify parameters for the initial portfolio (during the “ramp-up phase”, see below) but not the exact composition, for example, a minimum average rating, a minimum average yield, a maximum average maturity, and a minimum degree of diversification. As opposed to a static CDO, managers monitor and, if necessary, trade assets within the reference portfolio in order to inform decisions about asset purchases and sales that protect the collateral value from impairment due to deterioration in credit quality [6]. For further references, see [1–3, 5, 7, 8]. “Lightly managed” reference portfolios allow for some substitution of assets in the context of a defensive management strategy, while a “fully managed” CDO suggests a more active role of managers subject to limits and investment guidelines that are determined by the issuers, rating agencies, and different levels of risk tolerance of investors at inception. In the event of the issuer’s insolvency or default, managers are charged with maximizing recoveries on behalf of investors. However, investors in managed CDOs do not know what specific assets CDO managers will invest in, recognizing that those assets will change over time as managers alter the composition

of the reference portfolio. Thus, investors face both credit risk and the risk of poor management.

The majority of CDOs are managed and, in many instances, involve compounded structured finance claims. While standard CDOs use the same off-balance sheet structuring technology as *asset-backed securities* (ABSs) (e.g., securities that are themselves repackaged obligations on mortgages, consumer loans, home equity lines of credit, and credit card receivables), their reference portfolios typically include a wider and more diverse range of assets, such as senior secured bank loans, high-yield bonds, and *credit default swaps* (CDSs). In particular, the variable portfolio structure of managed CDOs is particularly amenable to refinance ABSs, emerging market bonds, or even other CDOs (to produce CDOs of CDOs, also called *CDO*²), as collateral assets in the so-called “pools-of-pools” structures.

Managed CDOs Have an Arbitrage Proposition

Managed CDOs are structured for *arbitrage* purposes. As opposed to *balance sheet* CDOs, where issuers unload defined asset exposure to third parties in order to change their balance sheet composition or debt maturity structure, in *arbitrage* transactions, the ability to trade a dynamic reference portfolio helps managers focus on the pool’s prospects for appreciation with the view of realizing economic gains while limiting downside risks. These gains result from the pricing mismatch between investment returns from reference assets (in the case of a *cash flow* structure) or credit protection premia on exposures (in the case of a *synthetic* structure) and lower financing cost of generally higher rated liabilities in the form of issued CDO securities. While cash flow CDOs, the most common type of CDOs, pay off liabilities with the cash generated from interest and principal payments, synthetic CDOs sell credit protection (together with various third-party guarantees) to create partially funded and highly leveraged investment on the performance of designated credit exposures (without actually purchasing the reference assets).

The Life of a Managed CDO

The life of a managed CDO can be divided into three distinct phases (Figure 1). During the “ramp-up

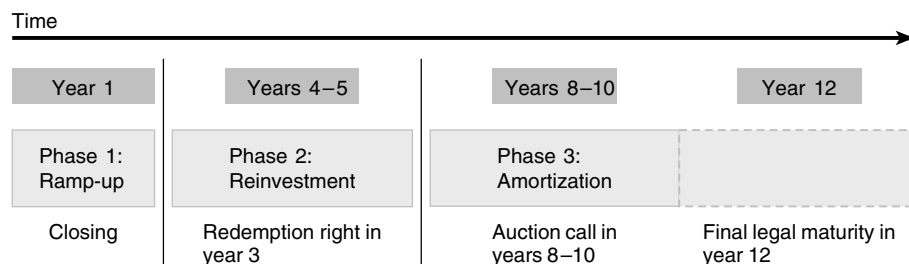


Figure 1 The phases in the life of a managed CDO

phase” (which lasts about one year), asset managers invest the proceeds from CDO placement (possibly after an initial warehousing period when the sponsor finances the buildup of the asset portfolio before securitizing). During the subsequent “reinvestment phase” (up to five years or longer), managers reinvest cash flows as well as trade the reference portfolio within the prescribed guidelines. Cash flows generated by the assets are used to pay back investors generally in sequential order from the senior investors, who hold the highest rated (typically “AAA”-rated) securities, to the “equity investors” who bear the first-loss risk and generally hold unrated securities. In transactions with *revolving* pools, portfolio assets can be replaced (e.g., credit card and trade receivables, corporate bonds) and balances are adjustable up to maximum limits without amortization schedule of principal. In contrast, managers of *substituting* pools incorporate new assets (within defined credit parameters) as original liabilities are paid down (e.g., corporate bonds, some residential mortgages, and consumer loans), but balances remain fixed. In the “amortization phase”, the reference portfolio matures (or is prepaid/sold) and investors receive some or all of their principal investment back according to the seniority of their claim.

Lessons from the Credit Crisis

Although rating agencies have developed stress tests to evaluate the resilience of dynamic portfolio structures, the 2007 subprime mortgage crisis demonstrated that managed CDOs might create incentive problems [4]. Existing quality and coverage tests on the underlying collaterals are designed to trigger amortization scenarios if asset performance deteriorates. However, CDO managers can manipulate these

tests to avoid early amortization. In response to a general repricing of risk, dwindling investor demand increased risk premia and curtailed the capacity of CDO managers to offset higher funding costs. Faced with rising liability pressures and without real buyers available, managers of “blind pools” could “double up” by opting for riskier positions and greater leverage to preserve own arbitrage gains within predefined investment guidelines, which were gradually undermined by the disassociation of ratings and structured asset performance. In principle, if transaction costs are ignored, risk-neutral managers would not benefit from dynamic asset allocation by substituting badly performing assets. Under worsening credit conditions, better asset performance comes at a premium, making it more expensive to weed out distressed assets. Therefore, CDO managers are no better off than before once they divert funds to safer but more costly assets (or accept higher hedging costs).

References

- [1] Cousseran, O. & Rahmouni, I. (2005). The CDO market – functioning and implications in terms of financial stability, *Banque de France Financial Stability Review* June (6), 43–62.
- [2] Duffie, D. & Gârleanu, N. (2001). Risk and valuation of collateralized debt obligations, *Financial Analysts Journal* 57(1), 41–59.
- [3] Goodman, L.S. & Fabozzi, F.J. (2002). *Collateralized Debt Obligations: Structures and Analysis*, John Wiley & Sons Inc., Hoboken, NJ.
- [4] Jobst, A. (2005). Risk management of CDOs during times of stress, *Derivatives Week*, Euromoney, London. (28 November), pp. 8–10.
- [5] Jobst, A. (2007). A primer on structured finance, *Journal of Derivatives and Hedge Funds* 13(3), 199–213.

- [6] Jobst, A. (2008). What is securitization? in *Finance and Development*, Vol. 47(3), (September), p. 48f.
- [7] Punjabi, S. & Tierney, J.F. (1999). *Synthetic CLOs and their Role in Bank Balance Sheet Management*, Deutsche Bank Research, Fixed Income Research.
- [8] Schorin, C. & Weinreich, S. (1998). *Collateralized Debt Obligation Handbook*. Working Paper, Fixed Income Research, Morgan Stanley Dean Witter.

Related Articles

Collateralized Debt Obligations (CDO); CDO Tranches: Impact on Economic Capital; Forward-starting CDO Tranche; Special-purpose Vehicle (SPV).

ANDREAS A. JOBST

Collateralized Debt Obligation (CDO) Options

The synthetic collateralized debt obligation (CDO) tranche activity is still a relatively new business, which started in late 2000. Initially seen as an eccentricity in the securitization market, it finally got a life on its own. Contrary to the rest of the securitization market, it grew as an arbitrage business, where the bank will not gain from structuring fees (like for cash CDOs) but from an arbitrage between two different markets: single-name credit default swaps *versus* synthetic single-tranche CDO. Its evolution was then marked by multiple borrowings from equity derivatives market, using its terminology (single tranche viewed as call spread on credit losses) and its technology (correlation smiles and the type of derivatives on it). This helps explain why the article focuses exclusively on synthetic CDO tranches, because to our knowledge there do not exist derivatives based on cash CDO notes.^a

For a more comprehensive CDO framework, *see Collateralized Debt Obligations (CDO)*, and for an introduction to CDO tranche pricing, *see Intensity-based Credit Risk Models*. However, we introduce those two important notions that are used in this article:

- A credit default swap (CDS) is a bilateral contract where the protection buyer will pay a quarterly premium (expressed as a proportion of the CDS notional) and receive from the protection seller, if specific events took place (related to the default or bankruptcy of a specific corporate entity), a payment corresponding to one minus the price of a bond of the defaulted entity, that payment being called the *severity of default* or the *credit loss*.
- A synthetic CDO tranche is a bilateral contract where the protection buyer will pay a quarterly premium (the premium leg of the swap) and receive from the protection seller the increment in loss on the tranche (the loss leg of the swap), where the loss on the tranche is contractually defined as a function of the sum of credit losses on a portfolio of single-names CDS, or more accurately as $\min(d - a, \max(L_t - a))$ with the following:

- L_t represents cumulative credit losses on that portfolio up to date t , that is, the sum of different severities of defaulted entities that are in that portfolio;
- a is defined as the attachment point of the CDO tranche in proportion of the initial pool size; and
- d is defined as its detachment point.

Definitions

Here, we list the different derivatives that we have seen in the synthetic CDO markets created between 2001 and 2007. Those derivatives can be categorized in two groups: derivatives on CDO tranches where the behavior of the derivative is conditioned by losses realization, or default-path-dependent derivatives, and derivatives conditioned by a spread/market value evolution. Some derivatives like leveraged supersenior (LSS) (*see Leveraged Super-senior Tranche*) can be in the two categories depending on their variations.

The first category of default-path-dependent structure is also known as *reset tranches* (as defined in [4]): those are CDO tranches where the attachment point and/or the width of the tranche are modified at a future reset date as a predetermined function of the portfolio losses up to that date.

- *Forward-starting CDO* (*see Forward-starting CDO Tranche and* [2]): this is a CDO tranche where the contract becomes effective at a future date, where any entities defaulting between the entry into the forward and its effective date will be considered to have a recovery rate at 100%, or in other words at the effective date, the CDO tranche will have an attachment point equal to the sum of the cumulative losses up to that date t_1 and the pre-fixed initial attachment point (i.e., $a + L_{t_1}$), its width being unchanged in dollar amount (i.e., $d - a$). There is also a variation on that contract [7] where the forward CDO is the obligation to enter into a CDO tranche at a future date, taking into consideration the erosion of subordination due to losses up to the effective date, and the decrease in the width, but not the losses on the tranche, thus a subordination of $\max(0; a - L_{t_1})$ and width of $\min(d; \max(a; L_{t_1}))$.
- *Subordination step-up*: This is a standard CDO tranche except that at the reset date, if losses have

not started to touch the tranche the subordination will be increased by a fixed amount. Multiple variations of those contracts exist with several reset dates or increase in subordination linked to losses being in a specific band.

- *Leveraged supersenior* (see **Leveraged Supersenior Tranche**): This is a synthetic CDO tranche, with a large attachment point, thus its supersenior nature, which is initially partially collateralized by the protection seller. Owing to that partial collateralization, when a loss trigger or mark-to-market (MtM) trigger is breached, the protection seller has the obligation of either providing additional collateralization or unwinding its contract at market value. When the trigger is based on loss level, this can be viewed as a reset tranche.

The second category encompasses all derivatives in the classical sense, that is, derivatives based on the market value of the underlying asset.

- *Call on CDO tranches*: This is an option giving the option holder the possibility to buy protection on a synthetic CDO tranche at a predetermined spread on one of several future dates, being either European for one single date or Bermudan for a set of future dates. The strike is defined as a spread level (and not as a value of the tranche). The synthetic CDO tranche, that is, portfolio composition, attachment/detachment points, and maturity, is defined initially, and akin to the differentiation done on forward-starting CDO, losses up to the exercise date of the option may or may not affect the attachment point.
- *Put on CDO tranches*: Contrary to the call option, this gives the option holder the possibility to sell protection on a CDO tranche at a predetermined spread.
- *Callable structure*: This is an option that gives the protection seller (or the protection buyer) the right to terminate the transaction at no additional costs during its life. If the option is for the protection seller, this is, in fact, a Bermudan call on the CDO tranche itself with a strike equal to its initial spread level. Here the attachment point of the underlying synthetic CDO tranche will be eroded by losses up to the exercise date of the option.
- *Rating guarantee*: We know of one investment bank that worked on the possibility to issue a guarantee on the CDO tranche rating, giving the

option to investors to put their CDO tranche at par to the issuer of such guarantee if the rating was downgraded below a prespecified threshold. This is in effect a callable structure conditional on the tranche being downgraded by a rating agency.

Purpose and Market

The purpose of those innovations is, in most cases, related to issues encountered by the bank's desk working on synthetic CDO tranches. The innovations in that market were always caused not by a need of the investors but by potential arbitrages to exploit. Synthetic CDO tranche from 2001 to 2007 was a booming market with several success stories. However, this was a very competitive market, where the competitive advantage was due to the endless creation of new structural features. Indeed, as soon as an innovation was introduced to the market by one player, several others tried to imitate it, soon depleting the potential gains evidenced by such innovation (Figure 1).

Each innovation was triggered by either an arbitrage to exploit or a specific problem encountered by the desk:

- *Forward-starting CDO*: Those products were created to exploit discrepancies between the terms structure of spread as the five-year maturity spread was depleted because of the wave of five-year synthetic CDO tranches. Synthetic CDO tranches were starting to be structured at 10 years or even as forward starting 5–10 years to benefit from the tightening at 5 years. Indeed, a 5–10 years forward starting CDO can be seen as a combination of a 10 years synthetic tranche and a 5 years synthetic tranche, selling protection for 10 years but buying protection for the first 5 years.
- *Leveraged supersenior*: When the correlation desk sold synthetic CDO tranches, they sold mainly equity and mezzanine tranches, and thus either delta-hedged them or kept the most senior tranches on their book. The supersenior exposures were very hard to sell due to their low spreads compared to their notional amounts, that is, the amount of cash needed to invest in those tranches. The creation of LSS allowed those desks to buy protection on supersenior synthetic CDO tranches by broadening the investor base outside of its

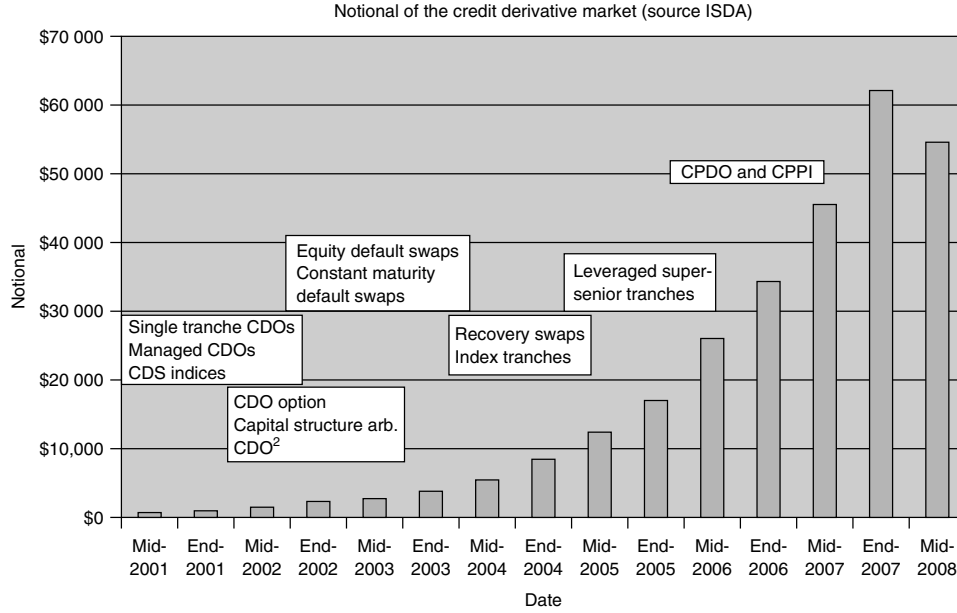


Figure 1 Evolution of the notional of the credit derivatives market with several innovations

initial clients (monolines or (re)insurance companies), the LSS having a higher spread for an “assumed” low credit risk.

Valuation

The valuation of a CDO tranche, whether initially or during its life, relies on the knowledge of the loss distribution of the underlying portfolio through time or, in other words on the law governing the random path L_t representing the cumulative losses up to t . The knowledge of the loss distribution at different future date (thus a loss distribution surface) is required^b to price a CDO tranche, that is, to value the two legs of that tranche swap: $P[L_T \leq l | \mathcal{F}_t]$ the probability that losses up to time T will exceed threshold l knowing the losses at time t .^c If the existing information in the market consists of the credit index tranches prices, in the arbitrage pricing theory framework, from those prices we will extract constraints on the “spot” loss distribution surface being $P[L_T \leq l | \mathcal{F}_0]$.

Some CDO derivatives can be valued with that “spot” loss distribution surface: the forward-starting CDO as described in [7] (the second variation as described above) can be understood as a long–short

position on maturities: long, the CDO tranche at the longest maturity and short, the same CDO tranche at the effective date. Indeed, on comparing the two positions—a t_1/t_2 forward-starting CDO tranche *versus* a long CDO tranche at t_2 and a short CDO tranche at t_1 with the same attachment/detachment points:

- if losses are always below the attachment point, no CDO tranches will be touched;
- if $L_{t_1} < a$ and $L_{t_2} > a$, the forward-starting CDO will lose $\min(d - a, L_{t_2} - a)$ and the long CDO tranche will lose the same amount; and
- if $L_{t_1} > a$ and $L_{t_2} > a$, the forward-starting CDO will lose $\min(d - L_{t_1}, L_{t_2} - L_{t_1})$ and the long–short CDO tranches will lose/gain $\min(d - a, L_{t_2} - a) / \min(d - a, L_{t_1} - a)$, which gives the same aggregate amount.^d

However, to value the other reset tranches, we need an additional information: the intertemporal dependence of losses, that is, the dependence between losses at different dates. For a forward-starting CDO—first variation—we need the law of (L_{t_1}, L_{t_2}) to be able to price it.

In addition, for options on tranche, dependent on future spreads, the knowledge of the “spot” loss

distribution surface is not sufficient to value those options. An additional assumption related to spread volatility is needed: this can be an *ad hoc* assumption directly on the volatility [7] or can be embedded into a stochastic deformation of the loss distribution surface $P[L_T \leq l | \mathcal{F}_t]$ through time. This leads researchers to introduce the class of models known as a *dynamic losses model*. The dynamic losses model defined so far relies on the standard CDO models, which are classified according to two broad categories [8]:

- *Top-down models*: The top-down approach will only look at the evolution of the losses on the portfolio and model its dynamics. The seminal paper describing a general framework for such dynamic of the “forward” loss distribution surface is [11], where the distribution of losses in the portfolio is represented as a Markov chain with stochastic transition rates. Andersen *et al.* [3] explore the same road in a less general manner. Those approaches are tractable, flexible but they do not capture information from the single-names CDS market.
- *Bottom-up models*: The approach starts with a representation of the credit risk of the underlying single names in order to build a loss distribution surface. Starting from the modeling of individual defaults, they use classical credit modeling:
- *Structural models (see Default Barrier Models)*: A structural model computes the default as the breaching by a random process of a barrier (in the initial Merton seminal article the first represents the assets of a company and the second its indebtedness). That class of model incorporates a dynamic for the probabilities of losses naturally, introducing default dependencies through the random process, with linear combination of random process (Brownian motion or Gamma process, see [9]). A related class of models looks at a discrete evolution of creditworthiness, generally with a Markov chain, where the use of stochastic transition rates can also be applied (a related example is in [1]).
- *Reduced-form models (see Intensity-based Credit Risk Models)*: A reduced-form model uses hazard rates to represent the risk of

default of individual companies; a portfolio can be analyzed with correlated hazard rates. A natural extension of those models to address dynamic losses is to use stochastic hazard rate for each company, which may be linked through common jumps (as first introduced by Duffie and Gârleanu [5]), correlated Brownian motion, or even introduction of stochastic time process mapping calendar time to business time [10].

Following the financial crisis of 2008, the landscape for synthetic CDO tranches has seen a change in paradigm. The default of Lehman Brothers in September 2008 and the demise of the investment bank business model have exposed the shaky foundations of the CDO market: liquidity drying in stress period and lack of acknowledgment of the counterparty risk in the CDS market. However, the initiatives that are currently discussed (standardization of that market, central clearing house) will, on the long term, expand the scope of that market and ultimately be beneficial for the development of those instruments.

End Notes

^a. Apart from rare guarantees offered by structuring desks or call optionality for the equity tranche of cash CDOs.

^b. In reality as pointed out in [6], the knowledge of the expected loss on the CDO tranche is sufficient to price it.

^c. The filtration $\mathcal{F}_t$ may embed more information than the cumulative losses up to that time.

^d. Taking into account even the timing of payment of losses, the two positions are the same.

References

- [1] Albanese, C., Chen, O., Dalessandro, A. & Vidler, A. (2005). *Dynamic Credit Correlation Modeling*, Working Paper, Imperial College.
- [2] Andersen, L. (2006). *Portfolio Losses in Factor Models: Term Structures and Intertemporal Loss Dependence*.
- [3] Andersen, L., Piterbarg, V. & Sidenius, J. (2005). *A New Framework for Dynamic Credit Portfolio Loss Modelling*. Working Paper, November.
- [4] Baheti, P., Mashal, R. & Naldi, M. (2006). *Step it Up or Start it Forward: Fast Pricing of Reset Tranches*, Lehman Brothers Quantitative Credit Research, Vol. 2006-Q1.
- [5] Duffie, D. & Gârleanu, N. (2001). Risk and the valuation of collateralized debt obligations, *Financial Analysts Journal* **57**, 41–59.

-
- [6] Hull, J. & White, A. (2006). Valuing credit derivatives using an implied copula approach, *Journal of Derivatives* **14**, 8–28. *for the Pricing of Portfolio Credit Derivatives*. Working Paper, ETZH.
- [7] Hull, J. & White, A. (2007). Forward and European options on CDO tranches, *Journal of Credit Risk* **3**, 63–73.
- [8] Hull, J. & White, A. (2008). Dynamic models of portfolio credit risk: a simplified approach, *Journal of Derivatives* **15**, 9–28.
- [9] Jäckel, P. (2008). *The Discrete Gamma Pool Model*. Working Paper, August.
- [10] Joshi, M. & Stacey, A. (2006). Intensity Gamma, *Risk* **19**, 78–83.
- [11] Schonbucher, P.J. (2006). *Portfolio Losses and the Term Structure of Loss Transition Rates: A New Methodology*

Related Articles

Collateralized Debt Obligations (CDO); Default Barrier Models; Forward-starting CDO Tranche; Intensity-based Credit Risk Models; Leveraged Super-senior Tranche.

OLIVIER TOUTAIN

Credit Default Swap Index Options

Portfolio credit default swaps (CDSs) referencing indices such as CDX and iTraxx are the most liquid instruments in today's credit market and options on these have become mainstream. A CDS index option (also called a *portfolio swaption*) is an option to enter into a portfolio swap as a protection buyer or a protection seller. A portfolio swap (also called a *CDS index swap*) is similar to a portfolio of single-name CDS all with the same coupon (for details, *see Credit Default Swap (CDS) Indices*).

Both portfolio swaps and swaptions are traded over the counter but are standardized. The conventions for how portfolio swaps are quoted and traded are important for properly valuing portfolio swaptions.

In this article, we outline the basic conventions and terminology for portfolio swaptions, explain the standard model used by most market participants, and briefly discuss other models and approaches.

Conventions and Terminology

A portfolio swaption is an option to enter into a portfolio swap as a protection buyer (payer swaption) or a protection seller (receiver swaption). The swaption is defined by the underlying portfolio swap, for example, 5-year CDX.IG.11, the expiration date, and the strike spread or strike price. For investment grade portfolios, it is a convention to specify a strike spread, whereas for high yield portfolios a strike price is usually specified. Trading is primarily in options on 5-year portfolio swaps. Option maturities are less than 1 year, with most liquidity in 1–3 months maturities. The standard option expiration dates are the 20th of each month.

The strike, whether it is specified as a spread or as a price, must be converted by a simple calculation to determine the cash amount to be exchanged between the swaption counterparties upon exercise. The calculation is easiest when a strike price is specified:

$$\begin{aligned}\text{Cash amount} &= \text{Notional} \cdot (\text{Strike price} - 100\%) \\ &\quad - \text{Accrued coupon}\end{aligned}\quad (1)$$

The cash amount is paid by the protection buyer. The accrued coupon enters the calculation because portfolio swaps, by convention trade with accrued coupon, similar to the way bonds trade with accrued interest. To simplify the exposition, we ignore accrued coupon in the remainder of the article.

When a strike spread is specified, the cash amount is calculated using the standard CDS valuation model, for example, as implemented in the Bloomberg CDSW screen:

$$\begin{aligned}\text{Cash amount} &= \text{Notional} \cdot \text{PV01} \cdot \\ &\quad (\text{Strike spread} - \text{Coupon})\end{aligned}\quad (2)$$

The coupon is the fixed premium rate for the underlying portfolio swap. When valuing a portfolio swaption, it is important to respect the exact market convention for calculating the PV01 such as the flat spread curve convention (*see Credit Default Swap (CDS) Indices*).

Another important market convention is that if the swaption is exercised, the option holder will buy or sell protection on all names in the portfolio including those that may have defaulted before option expiration.

The Standard Model

Now, suppose that V is the value at option expiration of owning protection on all names in the portfolio including those that have already defaulted. Option payoff at exercise is then

$$\begin{aligned}\text{Payer swaption payoff at exercise} \\ &= \max\{V - \text{Cash amount}, 0\} \\ \text{Receiver swaption payoff at exercise} \\ &= \max\{\text{Cash amount} - V, 0\}\end{aligned}\quad (3)$$

where the cash amount is calculated from the strike price as in equation (1) or from the strike spread using a CDS valuation model as in equation (2). The cash amount is not affected by defaults. In fact, if a strike price is specified, the cash amount is known with certainty before option expiration. If a strike spread is specified, the only uncertainty about the cash amount derives from uncertainty about the interest rate curve. However, when pricing portfolio swaptions, it is standard to assume that forward rates are realized.

To price a swaption, we must specify a stochastic model for V . In addition to assuming that risk-neutral valuation is proper [1, 3], the standard model is based on two minimal assumptions that are clarified further:

1. the spread of the underlying portfolio swap is lognormally distributed and
2. the model correctly prices a synthetic forward contract constructed by combining a long payer and a short receiver with the same strikes.

The standard model assumes that V is a function, $V(X)$, of a hypothetical spread:

$$X = E(X) \exp(-0.5\sigma^2 T + \sigma \text{Normal}(0, T)) \quad (4)$$

where $\text{Normal}(0, T)$ is random normal variable with mean 0 and variance T (the time, in years, to option expiration), and σ is the free parameter that we interpret as the spread volatility. $E(X)$ is the expected value of X . The function $V(X)$ is the one found in equation (2) when the cash amount is seen as a function of the strike spread.

The swaptions are priced by discounting their expected terminal payoff (risk-neutral valuation). To understand where $E(X)$ comes from, consider a payer and a receiver swaption both with a strike price of 100% or equivalently strike spreads equal to the coupon in the underlying portfolio swap. In this case, the cash amount in equation (1) or (2) is zero and the terminal payoff from a position that is long the payer and short the receiver is V . The value of this position is therefore

$$V_0 = D(T)E(V(X)) \quad (5)$$

where $D(T)$ is the discount factor to time T (option expiration).

The value of a position that pays V , that is, V_0 , can also be determined from the credit curve of the underlying portfolio (potentially using the credit curves of all the names in the portfolio) since it is simply the value of owning protection on all names in the portfolio but only having to pay premium from option expiration onward. Once we have a value for V_0 , $E(V(X))$ can be found as $V_0/D(T)$ and $E(X)$ can be implied from this value. We can then price the swaptions using σ as the only additional parameter.

It is recommended to solve the model numerically to get the most accurate pricing. However, by making a few simple approximations (such as simplifying the expression for the PV01 in equation (2)) it is possible to derive approximate closed-form solutions that look like Black formulas.

See [2] for details on the model outlined above.

Other Models and Approaches

The standard model is a simple approach to what could be a very complicated problem. Instead of trying to model the credit curves and default of each of the names in the portfolio, the approach in the standard model is to model the hypothetical spread on the aggregate portfolio that also includes defaulted names. Thereby the model has only one free parameter, the aggregate spread volatility, and the approach becomes similar to using Black–Scholes for S&P 500 options. This analogy to the equity world suggests paths to the next generation of models such as introducing stochastic volatility and jumps or creating a model that starts from the individual credits by modeling their default, spread volatility, and spread correlation.

References

- [1] Morini, M. & Brigo, D. (2007). *Arbitrage-free Pricing of Credit Index Options*, Working Paper, Bocconi University.
- [2] Pedersen, C. (2003). *Valuation of Portfolio Credit Default Swaptions*, *Lehman Brothers Quantitative Credit Research Quarterly*, 2003-Q4, pp. 71–81.
- [3] Rutkowski, M. & Armstrong, A. (2008). *Valuation of Credit Default Swaptions and Credit Default Index Swaptions*, Working Paper, University of New South Wales.

Related Articles

Credit Default Swaps; Credit Default Swap (CDS) Indices; Credit Default Swaption; Hazard Rate.

CLAUS M. PEDERSEN

Hazard Rate

Consider a *credit default swap* (CDS) (see **Credit Default Swaps**), where the premium payments are periodic and the terminal payment is a digital cash settlement of recovery rate $1 - \bar{\ell}$. For simplicity, we assume that the current time is normalized to $t = 0$, the risk-free rate r is constant throughout the maturity of the contract, and spreads are already given at standard interperiod rates (allowing us to ignore day-count fractions and division by period length). The cash flows of a CDS can be decomposed into the default leg and the premium leg. The default leg is a single lump sum compensation for the loss $\bar{\ell}$ on the face value of the reference asset made at the default time by the protection seller to the protection buyer, given that the default is before the expiration date T of the contract. The premium leg consists of the fees, called the *CDS spread*, paid by the protection buyer at dates t_m (assumed to be equidistant e.g., quarterly) until the default event or T , whichever is first. The spread S is given as a fraction of the unit notional.

A concise mathematical expression for both legs can be obtained *via* a point process representation. Suppose we have a filtered probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ satisfying the usual conditions. We model the default time as a random time τ in $[0, \infty]$ with an associated single jump point process

$$N_t = \mathbf{1}_{\{\tau \leq t\}} = \begin{cases} 1 & \text{if } \tau \leq t \\ 0 & \text{if } \tau > t \end{cases} \quad (1)$$

The default leg D_0 and premium leg $P_0(S)$ can now be expressed in terms of N as

$$D_0 = \mathbb{E}_0 \left[\bar{\ell} e^{-r\tau} N_\tau \right] = \mathbb{E}_0 \left[\int_0^T \bar{\ell} e^{-rs} dN_s \right] \quad (2)$$

$$P_0(S) = \mathbb{E}_0 \left[S \sum_{t_m} e^{-rt_m} (1 - N_{t_m}) \right] \quad (3)$$

where $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_t]$ is the conditional expectation with respect to time t , information $\mathcal{F}_t$, and the integral of equation (2) is defined in the Stieltjes sense. Finally, by an application of Fubini's theorem and integration by parts, equations (2) and (3) can be

expressed as

$$D_0 = \bar{\ell} e^{-rT} \mathbb{Q}_0[\tau \leq T] + \bar{\ell} r \int_0^T e^{-rs} \mathbb{Q}_0[\tau \leq s] ds \quad (4)$$

$$P_0(S) = S \sum_{t_m} e^{-rt_m} \mathbb{Q}_0[\tau > t_m] \quad (5)$$

The fair spread is the spread S^* for which $P_0(S^*) = D_0$, making the value of the contract at initiation 0. This simple expositional formulation shows that the modeling of survival probabilities under the pricing measure of the form $\mathbb{Q}_0[\tau > s]$ is the essence of CDS pricing. These quantities can be modeled in a unified way using the concept of *hazard rate*.

Hazard Rate and Default Intensity

Suppose that we have a filtered probability space $(\Omega, \mathcal{F}, \mathbb{Q})$ satisfying the usual conditions and that the default time of a firm is modeled by a random time τ , where $\mathbb{Q}[\tau = 0] = 0$ and $\mathbb{Q}[\tau > t] > 0$ for all $t \in \mathbb{R}_+$. We start under the assumption, which will be relaxed later, that the evolution of information only involves observations of whether or not default has occurred up to time t . In other words, we are dealing with the natural filtration $\mathcal{F}_t = \mathcal{N}_t = \sigma(N_s, s \leq t)$ of the right continuous, increasing process N , introduced earlier, completed to include the $\mathbb{Q}$ -negligible sets. Let $F(t) = \mathbb{Q}[\tau \leq t]$ be the cumulative distribution function of τ . Then, the *hazard function* of τ is defined by the increasing function $H : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ given as

$$H(t) = -\ln(1 - F(t)) \quad \forall t \in \mathbb{R}_+ \quad (6)$$

Suppose, furthermore, that F is absolutely continuous, admitting a density representation of $F(t) = \int_0^t f(s) ds$. The *hazard rate* of τ is defined by the nonnegative function $h : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ given as

$$h(t) = \frac{f(t)}{1 - F(t)} \quad (7)$$

under which we have

$$F(t) = 1 - e^{-H(t)} = 1 - e^{-\int_0^t h(s) ds} \quad \forall t \in \mathbb{R}_+ \quad (8)$$

Naturally, the component probabilities of equations (4) and (5) can be expressed in terms of the

hazard rate as

$$\begin{aligned} \mathbb{Q}[\tau > s | \mathcal{N}_t] \\ = \mathbf{1}_{\{\tau > t\}} e^{H(t) - H(s)} = \mathbf{1}_{\{\tau > t\}} e^{-\int_t^s h(u) du} \end{aligned} \quad (9)$$

$$\begin{aligned} \mathbb{Q}[t < \tau \leq s | \mathcal{N}_t] &= \mathbf{1}_{\{\tau > t\}} (1 - e^{H(t) - H(s)}) \\ &= \mathbf{1}_{\{\tau > t\}} \left(1 - e^{-\int_t^s h(u) du} \right) \end{aligned} \quad (10)$$

where $s \geq t$. Note that for a continuous h , $h(t)\Delta$ represents the first-order approximation of the probability of default between t and $t + \Delta$, given survival up to t . The term *hazard rate* stems from the fact that $h(t)$ can be thought of as the instantaneous ($\Delta \rightarrow 0$) rate of failure (in our case default arrival) at time t conditional on survival up to time t . Because of this conceptualization, the hazard rate is often referred to as the *forward default rate* in financial literature. While the term *default intensity* is also frequently used interchangeably with *hazard rate* [3], some authors [5] elect to distinguish between the two terms, where *intensity* is used to refer to the arrival rate conditioned on all observable information, and not only on survival. If survival is the only observable information, as in our current setting, the two terms are equivalent even under this distinction.

A useful alternate characterization of the hazard rate is possible using the martingale theory of point processes highlighted in [2]. As an increasing process, N has an obvious upward trend. The conditional probability of default by time $s \geq t$ is always greater than or equal to N_t itself, and hence N is a submartingale. It follows that N admits the Doob–Meyer decomposition [4] $N = M + A$, where M is an $\mathbb{F}$ -martingale and A is a right-continuous, $\mathbb{F}$ -predictable, increasing process starting from 0, both unique up to indistinguishability. A compensates for the upward trend such that $N - A$ is an $\mathbb{F}$ -martingale, and hence the popular terminology ($\mathbb{F}$ -)compensator (see **Compensators**). Compensators are interesting constructs in and of themselves, as their analytical properties correspond to probabilistic properties of the underlying random time. For instance, the almost sure continuity of sample paths of A is equivalent to the total inaccessibility of τ . Giesecke [6] outlines these properties and provides a direct compensator-based pricing application. As shown in [7], the connection between the compensator A and the hazard function H is that $A(t) = H(t \wedge \tau)$, if and

only if the cumulative distribution function $F(t)$ is continuous. Furthermore, if $F(t)$ is absolutely continuous we have $A(t) = \int_0^{t \wedge \tau} h(s) ds$, where h is the hazard rate. Therefore, under the continuity (absolute continuity) of $F(t)$, we can say that the hazard function H (hazard rate h) is the unique function such that $N(t) - H(t \wedge \tau)$ ($= N(t) - \int_0^{t \wedge \tau} h(s) ds$) is an $\mathbb{F}$ -martingale. However, the martingale/compensator characterization and the standard hazard function definition no longer coincide when $F(t)$ is not continuous.

In financial literature, we are also interested in cases where the current information filtration models not only survival but the observation of other processes as well. Let the total flow of information be modeled by $\mathbb{F} = \mathbb{N} \vee \mathbb{G}$, where $\mathbb{N}$ is once again the natural filtration of N (and all filtrations considered are right-continuous completions). Under certain conditions the previous concepts can be extended in a straightforward fashion. Let $F(t) = \mathbb{Q}[\tau \leq t | \mathcal{F}_t]$. First, we assume that τ is not a $\mathbb{G}$ -stopping time (while it is trivially an $\mathbb{F}$ -stopping time) such that the $\mathbb{G}$ -hazard process $H(t) = -\ln(1 - F(t))$ is well-defined. If H is absolutely continuous, admitting the $\mathbb{G}$ -progressive density representation $H(t) = \int_0^t h(s) ds$, then the process

$$M(t) := N(t) - H(t \wedge \tau) = N(t) - \int_0^{t \wedge \tau} h(s) ds \quad (11)$$

is an $\mathbb{F}$ -martingale, and the analogs of equations (9) and (10) are given by

$$\begin{aligned} \mathbb{Q}[\tau > s | \mathcal{F}_t] &= \mathbf{1}_{\{\tau > t\}} \mathbb{E} \left[e^{H(t) - H(s)} | \mathcal{F}_t \right] \\ &= \mathbf{1}_{\{\tau > t\}} \mathbb{E} \left[e^{-\int_t^s h(u) du} | \mathcal{F}_t \right] \end{aligned} \quad (12)$$

$$\begin{aligned} \mathbb{Q}[t < \tau \leq s | \mathcal{F}_t] &= \mathbf{1}_{\{\tau > t\}} \mathbb{E} \left[1 - e^{H(t) - H(s)} | \mathcal{F}_t \right] \\ &= \mathbf{1}_{\{\tau > t\}} \mathbb{E} \left[1 - e^{-\int_t^s h(u) du} | \mathcal{F}_t \right] \end{aligned} \quad (13)$$

In this setting, h is deemed the $\mathbb{G}$ -intensity or $\mathbb{G}$ -hazard rate. Even, in the case where τ is a $\mathbb{G}$ -stopping time ($\mathbb{D} \subset \mathbb{G}$) and thus the $\mathbb{G}$ -hazard function is not well defined, similar results can be obtained under certain conditions. Under certain restrictions on the

distributional properties of τ , we can still use point process martingale theory (see **Point Processes**) to find an increasing $\mathbb{F}$ -predictable process Λ for which the conditional survival probabilities are given by

$$\mathbb{Q}[\tau > s | \mathcal{F}_t] = \mathbf{1}_{\{\tau > t\}} \mathbb{E} \left[e^{\Lambda(t) - \Lambda(s)} | \mathcal{F}_t \right] \quad (14)$$

Routinely, if Λ is absolutely continuous then, $\Lambda(t) - \Lambda(s)$ in equation (14) can be replaced by its density representation $-\int_t^s \lambda(u) du$. The details of such conditions and results, as well as a general theory of hazard processes, are summarized in [1, 7].

Reduced-form Modeling and Other Issues

The importance of the concept of hazard rates (or intensities) lies in the fact that their direct modeling and parametrization is the prevalent industry practice in evaluating credit derivatives. Now that the CDS market has grown to one of great volume and liquidity, the realm of CDS spread modeling has become less of a pricing issue and more of a calibration one. *Reduced-form* modeling (see **Intensity-based Credit Risk Models**) refers to valuation methods in which one exogenously specifies the dynamics of an intensity model, much like we would for spot rates, and then calibrates the model parameters to fit the market spread data *via* a pricing formulation such as equations (4)–(5). A full-fledged model could incorporate features such as premium accrual, dependence of intensity with stochastic spot rates and the loss rate, and interaction/contagion effects with other names, which were ignored in our expositional formulation.

The assumptions, the underlying informational assumptions in particular, implied by the mere existence of a hazard rate are a nontrivial issue. Not all models admit an intensity process in their given information filtrations. For instance, in the classical first passage structural model under perfect information (see **Default Barrier Models**) the forward default rate (hazard rate with survival information only) exists, but the intensity process (hazard rate with all available information i.e. the firm value process) does not. Conceptually, the existence of a positive instantaneous default arrival rate implies a certain imperfection in the observable information, modeled either explicitly through a noisy filtration or implicitly through a totally inaccessible stopping time in the complete filtration. This underlines one of

the many issues surrounding the divergence of opinions and efforts for convergence in the reduced-form versus structural literature. Duffie and Singleton [5, 8] both provide a comprehensive overview of different credit models, while Giesecke [6] specifically outlines the different informational assumptions and their implications in intensity formulations.

References

- [1] Bielecki, T.R. & Rutkowski, M. (2001). *Credit Risk: Modeling, Valuation and Hedging*, Springer.
- [2] Bremaud, P. (1981). *Point Processes and Queues, Martingale Dynamics*, Springer-Verlag.
- [3] Brigo, D. & Mercurio, F. (2007). *Interest Rate Models - Theory and Practice, With Smile, Inflation and Credit*, 2nd Edition, Springer.
- [4] Dellacherie, C. & Meyer, P.A. (1982). *Probabilities and Potential*, North Holland, Amsterdam.
- [5] Duffie, D. & Singleton, K. (2003). *Credit Risk: Pricing, Measurement and Management*, Princeton University Press.
- [6] Giesecke, K. (2006). Default and information, *Journal of Economic Dynamics and Control* **30**, 2281–2303.
- [7] Jeanblanc, M. & Rutkowski, M. (2000). Modelling of default risk: an overview, in *Mathematical Finance: Theory and Practice*, Higher Education Press, Beijing, pp. 171–269.
- [8] Lando, D. (2004). *Credit Risk Modeling: Theory and Applications*, Princeton University Press.

Further Reading

- International Swaps and Derivatives Association (1997). *Confirmation of OTC Credit Swap Transaction Single Reference Entity Non-Sovereign*.
- International Swaps and Derivatives Association (2002). *2002 Master Agreement*.
- Tavakoli, J. (2001). *Credit Derivatives and Synthetic Structures, A Guide to Instruments and Structures*, 2nd Edition, John Wiley & Sons.

Related Articles

Compensators; Credit Default Swaps; Default Barrier Models; Duffie–Singleton Model; Intensity-based Credit Risk Models; Jarrow–Lando–Turnbull Model; Point Processes; Reduced Form Credit Risk Models.

JUNE HO KIM

Duffie–Singleton Model

The credit risk modeling approach of Duffie and Singleton [8, 9] falls into the class of *reduced-form* (see **Reduced Form Credit Risk Models**; **Intensity-based Credit Risk Models**) or intensity-based models in the sense that default is directly modeled as being triggered by a point process, as opposed to structural models (see **Structural Default Risk Models**) attempting to explain default through the dynamics of the firm’s capital structure, and the intensity of this process under a risk-neutral probability measure is related to an appropriately defined instantaneous credit spread. In its original construction, it is set out as an econometric model, that is, a model the parameters of which are estimated from the time series of market data, such as the weekly data of swap yields used in [8]. To this end, the model is driven by a set of state variables following a Markov process under the risk-neutral measure, and defaultable zero-coupon bond prices are exponentially affine functions of the state variables along the lines of the results derived by Duffie and Kan [6] for default-free models of the term structure of interest rates (see **Affine Models**). Duffie and Singleton [9] show that the model framework can be made specific in a way that also allows default intensities and default-free interest rates to be negatively correlated in a manner that is more consistent theoretically than in prior attempts in the literature.

A key assumption of Duffie–Singleton is the modeling of recovery in the event of default as an exogenously given fraction of the market value of the defaultable claim immediately prior to default. Under this assumption, the possibility of default on a claim can be priced by default-adjusting the interest rate with which the future cash flow (or payoff) from the claim is discounted. That is to say that today’s ($t = 0$) value V_0 of a claim with the (possibly random) payoff X at time $t = T$ can be calculated as the expectation under the spot risk-neutral measure Q ,

$$V_0 = E_0^Q \left[\exp \left\{ - \int_0^T R_t dt \right\} X \right] \quad (1)$$

where the discounting is given in terms of the default-adjusted short-rate process $R_t = r_t + h_t L_t$, with r_t the default-free continuously compounded short rate,

h_t the default hazard rate, and L_t the fraction of market value lost in the event of default. $\lambda_t = h_t L_t$ can be interpreted as a “risk-neutral mean-loss rate of the instrument due to default.” As a consequence, credit spread data alone (be it corporate bond yields, swap to treasury spreads, or credit default swap spreads) are insufficient to separate the “risk-neutral mean-loss rate” λ_t into its hazard rate h_t and loss fraction L_t .

The representation (1) lends the model considerable tractability, particularly for applications that do not require the separation of R_t into its components r_t , h_t , and L_t , since R_t could then be modeled directly as a function $\rho(Y_t)$ of a state variable process Y that is Markovian under Q . If the payoff of the claim is also Markovian in Y , say $X = g(Y_T)$, then the value of the claim at any time t (assuming that default has not occurred by time t) can be written as the conditional expectation

$$V_t = E^Q \left[\exp \left\{ - \int_t^T \rho(Y_s) ds \right\} g(Y_T) \middle| Y_t \right] \quad (2)$$

$\rho(Y_s)$ can be modeled analogously to any one of a number of tractable default-free interest rate term structure models. One possible choice of making the Markovian model specific is along the lines of a multifactor affine term structure model as studied by Dai and Singleton [3], in which r_t and λ_t are affine functions of the vector Y_t ,

$$r_t = \delta_0 + \sum_{i=1}^N \delta_i Y_t^{(i)} = \delta_0 + \delta_Y^\top Y_t \quad (3)$$

$$\lambda_t = \gamma_0 + \sum_{i=1}^N \gamma_i Y_t^{(i)} = \gamma_0 + \gamma_Y^\top Y_t \quad (4)$$

and Y_t follows an “affine diffusion”

$$dY_t = \mathcal{K}(\Theta - Y_t) dt + \Sigma \sqrt{S(t)} dW(t) \quad (5)$$

where W is an N -dimensional standard Brownian motion under Q , $\mathcal{K}$ and Σ are $N \times N$ matrices (which may, in general, be nondiagonal and asymmetric), and $S(t)$ is a diagonal matrix with the i th diagonal element given by

$$[S(t)]_{ii} = \alpha_i + \beta_i^\top Y_t \quad (6)$$

If certain admissibility conditions on the model parameters are satisfied [3], it follows from [6] that

default-free and defaultable zero-coupon bond prices are exponential affine functions of the state variables.

Duffie and Singleton [9] highlight that modeling Y as a vector of independent components following [2] “square-root diffusions” constrains the joint conditional distribution of r_t and λ_t in a manner inconsistent with empirical findings. In particular, the [3] conditions on admissible model parameters imply that such a model cannot produce negative correlation between the default-free interest rate and the default hazard rate. Duffie–Singleton instead propose to use a more flexible specification, which does not suffer from this disadvantage. In its three-factor form, it is given by

$$\alpha = \begin{pmatrix} 0 \\ 0 \\ \beta_3 \end{pmatrix} \quad \beta_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \beta_2 = \begin{pmatrix} 0 \\ \beta_{22} \\ 0 \end{pmatrix}$$

$$\beta_3 = \begin{pmatrix} \beta_{31} \\ \beta_{32} \\ 0 \end{pmatrix} \quad \delta_Y = \begin{pmatrix} \delta_1 \\ 1 \\ 1 \end{pmatrix} \quad \gamma_Y = \begin{pmatrix} \gamma \\ \gamma \\ 0 \end{pmatrix} \quad (7)$$

with all coefficients (including δ_0 and γ_0 in equations (3) and (4)) strictly positive. Furthermore,

$$\mathcal{K} = \begin{bmatrix} \kappa_{11} & \kappa_{12} & 0 \\ \kappa_{21} & \kappa_{22} & 0 \\ 0 & 0 & \kappa_{33} \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \sigma_{31} & \sigma_{32} & 1 \end{bmatrix} \quad (8)$$

with the off-diagonal elements of $\mathcal{K}$ being nonpositive. This specification ensures strictly positive credit spreads λ_t and can represent negative correlation between the increments of r and λ .

The “recovery-of-market-value” assumption at the core of the Duffie–Singleton framework is in line with market practice for defaultable derivative financial instruments such as swaps. For defaultable bonds, it is arguably more realistic to model the loss in the event of default as a fraction of the par value. However, Duffie and Singleton [9] provide evidence that par yield spreads implied by reduced-form models are relatively robust with respect to different recovery assumptions, and suggest that for bonds trading substantially away from par, pricing differences due to different recovery assumptions can be largely compensated by changes in the recovery parameters. The computational tractability gained through the “recovery-of-market-value” assumption may thus

justify accepting its slight inconsistency with legal and market practice.

The parallels of equation (1) to the valuation of contingent claims in default-free interest rate term structure models also extend to the methodology of Heath *et al.* [10] (HJM). Defining a term structure of “defaultable instantaneous forward rates” $\bar{f}(t, T)$ in terms of defaultable zero-coupon bond prices $\bar{B}(t, T)$ (i.e., the time t price of a bond maturing in T) by

$$\bar{B}(t, T) = \exp \left\{ - \int_t^T \bar{f}(t, u) du \right\} \quad (9)$$

the model can be written in terms of the dynamics of the $\bar{f}(t, T)$, the drift of which under the risk-neutral measure must obey the no-arbitrage restrictions, derived by Heath, Jarrow, and Morton (HJM) in the default-free case. Note that the $\bar{f}$ are “forward rates” only in the sense that equation (9) is analogous to the definition of instantaneous forward rates in the default-free case and their relationship to forward bond prices is less straightforward than for default-free forward rates. That is to say that typically for the forward price $\bar{F}(t, T_1, T_2) = \bar{B}(t, T_2)/\bar{B}(t, T_1)$ (where $B(t, T)$ is a default-free zero-coupon bond), one has

$$\bar{F}(t, T_1, T_2) \neq \frac{\bar{B}(t, T_2)}{\bar{B}(t, T_1)} = \exp \left\{ - \int_{T_1}^{T_2} \bar{f}(t, u) du \right\} \quad (10)$$

For the continuously compounded defaultable short rate $\bar{r}(t) = \bar{f}(t, t)$, the no-arbitrage restrictions imply

$$\bar{f}(t, t) = r_t + h_t L_t = R_t \quad (11)$$

which is equal to the default-adjusted short rate given in equation (1). In this sense, the risk-neutral mean-loss rate $h_t L_t$ is equal to the instantaneous credit spread $\bar{r}(t) - r_t$.

Cast in terms of HJM, the model is automatically calibrated to an initial term structure of defaultable discount factors $\bar{B}(t, T)$. This type of straightforward “cross-sectional” calibration makes the model useful not only for the econometric estimation followed by Duffie and Singleton [8] and others such as Duffee [4] and Collin-Dufresne and Solnik [1] but also for the relative pricing of credit derivatives.

The model can be extended in a number of directions, several of which are discussed in [9]. “Liquidity” effects can be modeled by defining a fractional

carrying cost ℓ of defaultable instruments, in which case the relevant discount rate $R_t = r_t + h_t L_t + \ell_t$ is adjusted for default and liquidity. The assumption of exogenous default intensity and recovery rate can be lifted, as in [5], by allowing intensities/recovery rates to differ for the counterparties in an over-the-counter (OTC) derivative transaction, with the intensity/recovery rate relevant for discounting determined by which counterparty is in the money. Jumps in the default-adjusted rate can be introduced along the lines of [6] while preserving the tractability of an affine term structure model. The model of single-obligor default considered by Duffie and Singleton [8, 9] can also be extended to the portfolio level using the copula function approach of Schönbucher and Schubert [11], since introducing default correlation through correlated diffusive dynamics of the default intensities h_t for different obligors is typically insufficient, resulting only in very mild correlation of defaults.

Historically, reduced-form models like Duffie–Singleton have been considered to be following a different paradigm than the more fundamental structural models where default is triggered when the value of the firm falls below a barrier taken to represent the firm’s liabilities. However, the two approaches have been reconciled by Duffie and Lando [7], who show that models based on a default intensity can be underpinned by a structural model in which bondholders are imperfectly informed about the firm’s value.

References

- [1] Collin-Dufresne, P. & Solnik, B. (2001). On the term structure of default premia in the swap and LIBOR markets, *Journal of Finance* **56**(3), 1095–1115.
- [2] Cox, J.C., Ingersoll, J.E. & Ross, S.A. (1985). A theory of the term structure of interest rates, *Econometrica* **53**(2), 385–407.
- [3] Dai, Q. & Singleton, K.J. (2000). Specification analysis of affine term structure models, *The Journal of Finance* **55**(5), 1943–1978.
- [4] Duffie, G. (1999). Estimating the price of default risk, *Review of Financial Studies* **12**(1), 197–226.
- [5] Duffie, D. & Huang, M. (1996). Swap rates and credit quality, *Journal of Finance* **51**(3), 921–949.
- [6] Duffie, J.D. & Kan, R. (1996). A yield factor model of interest rates, *Mathematical Finance* **6**(4), 379–406.
- [7] Duffie, D. & Lando, D. (2001). Term structures of credit spreads with incomplete accounting information, *Econometrica* **69**(3), 633–664.
- [8] Duffie, D. & Singleton, K.J. (1997). An econometric model of the term structure of interest-rate swap yields, *The Journal of Finance* **52**(4), 1287–1322.
- [9] Duffie, D. & Singleton, K. (1999). Modeling term structures of defaultable bonds, *Review of Financial Studies* **12**, 687–720.
- [10] Heath, D., Jarrow, R. & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation, *Econometrica* **60**(1), 77–105.
- [11] Schönbucher, P. & Schubert, D. (2001). *Copula Dependent Default Risk in Intensity Models*, University of Bonn. Working paper.

Related Articles

Affine Models; Constant Maturity Credit Default Swap; Intensity-based Credit Risk Models; Jarrow–Lando–Turnbull Model; Markov Processes; Multiname Reduced Form Models; Point Processes; Reduced Form Credit Risk Models.

ERIK SCHLÖGL & LUTZ SCHLÖGL

Jarrow–Lando–Turnbull Model

The credit-risk model of Jarrow, Lando, and Turnbull is based on a Markov chain with finite state space, modeled in discrete or continuous time. Economically, it relies on the appealing interpretation of using different rating classes, which are represented by the states of the Markov chain. Presumably, it is the first credit-risk model that incorporates rating information into the valuation of defaultable bonds and credit derivatives. An advantage of modeling the credit-rating process is that the resulting bond prices explicitly depend on the issuer's initial rating and possible rating transitions in the future. Moreover, the model allows to price derivatives whose payoffs depend on the credit rating of some reference bond, an application that is not straightforward in intensity-based models or structural-default models.

Technically, the model is formulated on a filtered probability space with a money-market account $B = \{B(t)\}_{0 \leq t \leq T}$ as numéraire. The state space of the underlying Markov chain is denoted by $S = \{1, \dots, K\}$, where state K represents default. The other states are identified with rating classes that are ordered according to increasing default risk, that is, state 1 represents the best rating. Transition probabilities from one state to another are specified *via* a probability matrix $\mathbf{Q}$ in discrete time and using a generator matrix $\mathbf{\Lambda}$ in continuous time. Multiple defaults are excluded by making the default state absorbing, which corresponds to specific choices of the last rows of $\mathbf{Q}$ and $\mathbf{\Lambda}$, respectively. The original model achieves a high level of tractability by imposing the following assumptions: existence of a unique equivalent martingale measure, independence of risk-free interest rates and credit migrations under the martingale measure, and a constant recovery R paid at the maturity of the defaulted bonds. It is further suggested in [8] that historical transition probabilities could be adjusted by some deterministic, time-dependent, proportional risk premium to derive the required transition matrix $\tilde{\mathbf{Q}}$ or generator matrix $\tilde{\mathbf{\Lambda}}$ under the martingale measure. Then, T -year survival probabilities are expressed in terms of this martingale measure, under which defaultable bonds, futures, and derivatives on risky bonds are priced by computing their expected discounted payoff. To be more precise, let us briefly

describe the discrete-time case. Denoting the matrix of risk premiums at time t by the $K \times K$ -dimensional diagonal matrix $\mathbf{\Pi}(t) = \text{diag}(\pi_1(t), \dots, \pi_{K-1}(t), 1)$, it is assumed that

$$\tilde{\mathbf{Q}}(t, t+1) - \mathbf{I} = \mathbf{\Pi}(t)(\mathbf{Q} - \mathbf{I}) \quad (1)$$

where $\mathbf{I}$ denotes the K -dimensional identity matrix and with assumptions ensuring that $\tilde{\mathbf{Q}}(t, t+1)$ is a probability matrix with absorbing state K . It is well known that the n -step transition matrix at time t under the martingale measure is given by

$$\tilde{\mathbf{Q}}(t, t+n) = \prod_{i=0}^{n-1} \tilde{\mathbf{Q}}(t+i, t+i+1), \quad \forall n \in \mathbb{N} \quad (2)$$

Let τ denote the random default time and $C(T)$ the random payoff at time T of a credit-risky claim. Then the value $C(t)$ at time t of this contingent claim is given by

$$C(t) = B(t) \cdot \tilde{\mathbb{E}}_t[C(T)/B(T)] \quad (3)$$

with $\tilde{\mathbb{E}}_t$ denoting the conditional expectation, with respect to the information at time t under the martingale measure. Under these assumptions, the price $P(t, T)$ of a default-free zero-coupon bond at time t , maturing at time T , is given by

$$P(t, T) = B(t) \cdot \tilde{\mathbb{E}}_t[1/B(T)] \quad (4)$$

The corresponding price $P_i^d(t, T)$ of a defaultable zero-coupon bond rated i at time t is given by

$$P_i^d(t, T) = P(t, T) \cdot \left(R + (1 - R) \cdot \tilde{Q}_i^i(\tau > T) \right) \quad (5)$$

with $\tilde{Q}_i(\tau > T) = (\tilde{Q}_i^1(\tau > T), \dots, \tilde{Q}_i^{K-1}(\tau > T), 0)$ denoting the time T survival probabilities for firms in the different rating classes at time t under the martingale measure. These survival probabilities are given by

$$\tilde{Q}_i^i(\tau > T) = 1 - \tilde{q}_{i,K}(t, T), \quad i = 1, \dots, K \quad (6)$$

where $\tilde{q}_{i,K}(t, T)$ denotes the respective element of $\tilde{\mathbf{Q}}(t, T)$. Further applications of the model include the construction of hedging strategies against rating changes and the pricing of options on risky debt.

Most results derived in the discrete model find their analog in the continuous version of the model. More complicated in a continuous framework is the derivation of martingale probabilities, specified by the generator matrix $\tilde{\Lambda}$. For the construction of this matrix, [8] presents an implicit calibration method that starts with a historical estimation of Λ . On the basis of the paradigm of choosing $\tilde{\Lambda}$ close to the historical Λ , a proportional risk premium is introduced, which is calibrated to observed bond prices.

The original references of the presented methodology are [9] and [8], and an excellent textbook summary is Chapter 12 of [4]. An introduction to Markov chains is given in [1] and [3]. Estimation procedures of the historical intensity matrix are studied in [10] and [7]. Considering generalizations of the model, let us mention [5] for stochastic recovery rates, [2, 12], and [13], for transition probabilities explained by state variables, and [11] for a different risk premium. Finally, a multifirm extension using a stochastic time change is presented in [6].

References

- [1] Anderson, W.J. (1991). *Continuous-Time Markov Chains. An Applications-Oriented Approach*, Springer Verlag, New York.
- [2] Arvanitis, A., Gregory, J. & Laurent, J.-P. (1999). Building models for credit spreads, *The Journal of Derivatives* **6**(3), 27–43.
- [3] Behrens, E. (2000). *Introduction to Markov Chains*, Vieweg Verlag, Braunschweig/Wiesbaden.
- [4] Bielecki, T.R. & Rutkowski, M. (2002). *Credit Risk: Modeling, Valuation and Hedging*, Springer Verlag, Berlin.
- [5] Das, S.R. & Tufano, P. (1996). Pricing credit-sensitive debt when interest rates, credit ratings and credit spreads are stochastic, *The Journal of Financial Engineering* **5**(2), 161–198.
- [6] Hurd, T.R. & Kuznetsov, A. (2007). Affine Markov chain model of multifirm credit migration, *The Journal of Credit Risk* **3**(1), 3–29.
- [7] Israel, R.B., Rosenthal, J.S. & Wei, J.Z. (2001). Finding generators for Markov chains via empirical transition matrices, with applications to credit risk, *The Journal of Mathematical Finance* **11**(2), 245–265.
- [8] Jarrow, R.A., Lando, D. & Turnbull, S.M. (1997). A Markov model for the term structure of credit risk spreads, *The Review of Financial Studies* **10**(2), 481–523.
- [9] Jarrow, R.A. & Turnbull, S.M. (1995). Pricing derivatives on financial securities subject to credit risk, *The Journal of Finance* **50**(1), 53–85.
- [10] Kavvathas, D. (2001). Estimating credit rating transition probabilities for corporate bonds, *AFA 2001 New Orleans Meetings*, New Orleans.
- [11] Kijima, M. & Komoribayashi, K. (1998). A Markov chain model for valuing credit risk derivatives, *The Journal of Derivatives* **6**(1), 97–108.
- [12] Thomas, L.C., Allen, D.E. & Morkel-Kingsbury, N. (2002). A hidden Markov chain model for the term structure of bond credit risk spreads, *International Review of Financial Analysis* **11**(3), 311–329.
- [13] Wei, J.Z. (2003). A multi-factor, credit migration model for sovereign and corporate debts, *The Journal of International Money and Finance* **22**(5), 709–735.

Related Articles

Credit Default Swaps; Duffie–Singleton Model; Hazard Rate; Multiname Reduced Form Models.

RUDI ZAGST & MATTHIAS A. SCHERER

Intensity-based Credit Risk Models

Reduced-form credit risk models have become standard tools for pricing credit derivatives and for providing a link between credit spreads and default probabilities. In structural models, following the Merton approach [1, 12], default is defined by a firm value hitting a certain barrier. In such an approach, the concept of credit spread is rather abstract since it is not modeled explicitly and therefore is not directly accessible and may also have dynamics that are not completely pleasing. Reduced-form models, however, concentrate on modeling the hazard rate or intensity of default, which is directly linked to the credit spread process. In contrast to a structural approach, the event of default in a reduced-form model comes about as a sudden unanticipated event (although the likelihood of this event may have been changing).

Deterministic Hazard Rates

Risk-neutral Default Probability

The basic idea around pricing default sensitive products is that of considering a risky zero-coupon bond of unit notional and maturity T . We write the payoff at maturity as

$$C(T, T) = \begin{cases} 1 & \text{default} \\ \delta & \text{no default} \end{cases} \quad (1)$$

where δ is an assumed recovery fraction paid immediately in the event of default. The price of a risky cash flow due at time T is then

$$C(t, T) = [S(t, T) + [1 - S(t, T)]\delta]B(t, T) \quad (2)$$

with $B(t, T)$ denoting the risk-free discount factor for time T as seen from time t , $S(t, T)$ is the risk-neutral survival (no default) probability (*see Hazard Rate*) in the interval $[t, T]$ or, equivalently, $1 - S(t, T)$ is the risk-neutral default probability. This style of approach was developed by Jarrow and Turnbull [8, 9].

Pricing a Credit Default Swap (CDS)

A credit default swap (CDS) (*see Credit Default Swaps*) has become a benchmark product for trading

credit risk and hence we base most of our analysis around CDS pricing. Standard assumptions used in pricing CDS include deterministic default probabilities, interest rates, and recovery values (or at least independence between these three quantities). In a CDS contract, the protection buyer will typically pay a fixed periodic premium, X_{CDS} , to the protection seller until the maturity date or the default (credit event) time (T). The present value of these premiums at time t can be written as

$$V_{\text{premium}}(t, T) = \sum_{i=1}^m S(t, t_i) B(t, t_i) \Delta_{i-1, i} X_{\text{CDS}} \quad (3)$$

where m is the number of premium payments and $\Delta_{i-1, i}$ represents the day count fraction.

The protection seller in a CDS contract will undertake in the event of a default to compensate the buyer for the loss of notional less some recovery value, δ . The value of the default component obtained by integrating over all possible default times is given by

$$V_{\text{default}}(t, T) = (1 - \delta) \int_t^T B(t, u) dS(t, u) \quad (4)$$

Note that due to the required negative slope of $S(t, u)$, this term will be negative; hence, the sum of equations (3) and (4) defines the value of a CDS from a protection provider's point of view.

Defining the Hazard Rate

In pricing a CDS, the main issue is to define $S(t, u)$ for all relevant times in the future, $t \leq u \leq T$. If we consider default to be a Poisson process driven by a constant intensity of default, then the survival probability is

$$S(t, u) = \exp[-h(u - t)] \quad (5)$$

where h is the intensity of default, often described as the hazard rate. We can interpret h as a forward instantaneous default probability; the probability of default in a small period dt conditional on no prior default is $h dt$. Default is a sudden unanticipated event (although it may, of course, have been partly anticipated due to a high value of h).

Link from Hazard Rate to Credit Spread

If we assume that CDS premiums are paid continuously,^a then the value of the premium

payments can be written as

$$V_{\text{premium}}(t, T) \approx X_{\text{CDS}} \int_t^T B(t, u) S(t, u) du \quad (6)$$

Under the assumption of a constant hazard rate of default, we can write $dS(t, u) = -hS(t, u) du$ and the default payment leg becomes

$$V_{\text{default}}(t, T) = -(1 - \delta)h \int_t^T B(t, u) S(t, u) du \quad (7)$$

The CDS spread will be such that the total value of these components is zero. Hence from $V_{\text{premium}}(t, T) + V_{\text{default}}(t, T) = 0$ we have the simple relationship

$$h \approx \frac{X_{\text{CDS}}}{(1 - \delta)} \quad (8)$$

The above close relationship between the hazard rate and CDS premium (credit spread) is important in that the underlying variable in our model is directly linked to credit spreads observed in the market. This is a key advantage over structural models whose underlying variables are rather abstract and hard to observe.

Simple Formulas

Suppose we define the risk-free discount factors *via* a constant continuously compounded interest rate $B(t, u) = \exp[-r(u - t)]$. We then have closed-form expressions for quantities such as

$$\begin{aligned} & V_{\text{premium}}(t, T) / X_{\text{CDS}} \\ & \approx \int_t^T \exp[-(r + h)(u - t)] du \\ & = \frac{1 - \exp[-(r + h)(T - t)]}{r + h} \end{aligned} \quad (9)$$

The above expression and equation (8) allow a quick calculation for the value of a CDS, or equivalently a risky annuity or DV01 for a particular credit.

Incorporating Term Structure

For a nonconstant intensity of default, the survival probability is given by

$$S(t, u) = \exp \left[- \int_t^u h(x) dx \right] \quad (10)$$

To allow for a term structure of credit (e.g., CDS premia at different maturities) and indeed a term structure of interest rates, we must choose some functional form for h . Such an approach is the credit equivalent of yield curve stripping, although due to the illiquidity of credit spreads much less refined, and was first suggested by Li [10]. The single-name CDS market is mainly based around 5-year instruments and other maturities will be rather illiquid. A standard approach is to choose a piecewise constant representation of the hazard rate to coincide with the maturity dates of the individual CDS quotes.

Extensions

Bonds and Basis Issues

Within a reduced-form framework, bonds can be priced in a similar way to CDS:

$$\begin{aligned} V_{\text{bond}}(t, T) = & \sum_{i=1}^m S(t, t_i) B(t, t_i) \Delta_{i-1, i} X_{\text{bond}} \\ & + S(t, T) B(t, T) - \delta \int_t^T B(t, u) dS(t, u) \end{aligned} \quad (11)$$

The first term above is similar to the default payment on a CDS but the assumption here is that the bond will be worth a fraction δ in default. The second and third terms represent the coupon and principal payments on the bond, respectively. It is therefore possible to price bonds *via* the CDS market (or *vice versa*) and indeed to calibrate a credit curve *via* bonds of different maturities from the same issuer. However, the treatment of bonds and CDS within the same modeling framework must be done with caution. Components such as funding, the CDS delivery option, delivery squeezes, and counterparty risk mean that CDS and bonds of the same issuer will trade with a basis representing nonequal risk-neutral default probabilities. In the context of the

formulas, the components creating such a basis would represent different recovery values as well as discount factors when pricing CDS and bonds of the same issuer.

Stochastic Default Intensity

The deterministic reduced-form approach can be extended to accommodate stochastic hazard rates and leads to the following expression for survival probabilities:-

$$S(t, u) = E^Q \left[\exp \left[- \int_t^u h(x) dx \right] \right] \quad (12)$$

This has led to various specifications for modeling a hazard rate process with parallels with interest-rate models for modeling products sensitive to credit spread volatility with examples to be found in [4, 5, 11]. Jarrow *et al.* [7] (see **Jarrow-Lando-Turnbull Model**) have extended such an approach to have a Markovian structure to model credit migration or discrete changes in credit quality that would lead to jump in the credit spread. Furthermore, credit hybrid models with hazard rates correlated to other market variables, such as interest rates, have been introduced. For example, see [13].

Portfolio Approaches

The first attempts at modeling portfolio credit products, such as basket default swaps and CDOs, involved multidimensional hazard rate models. However, it was soon realized that introducing the level of default correlation required to price such products realistically was far from trivial. This point is easily understood by considering that two perfectly correlated hazard rates will not produce perfectly correlated default events and more complex dynamics are required such as those considered by Duffie [3]. Most portfolio credit models have instead followed structural approaches (commonly referred to as *copula models* with the so-called Gaussian copula model becoming the market standard for pricing CDOs; see **Gaussian Copula Model**) for reasons of simplicity. Schonbucher and Schubert [14] have shown how to combine intensity and copula models. More recently, the search for

more sophisticated portfolio credit risk modeling approaches is largely based around reduced-form models as in [2] and [6] (see **Multiname Reduced Form Models**).

Conclusions

We have outlined the specification and usage of reduced-form models for modeling a default process and described the link between the underlying in such a model and market observed credit spreads. We have described the application of such models to vanilla credit derivative structures such as CDS and also more sophisticated structures such as credit spread options, credit hybrid instrument, and portfolio credit products.

End Notes

^aCDS premiums are typically paid quarterly in arrears but an accrued premium is paid in the event of default to compensate the protection seller for the period for which a premium has been paid. Hence the continuous premium assumption is only a mild approximation.

References

- [1] Black, F. & Cox, J. (1976). Valuing corporate securities: some effects of bond indenture provisions, *Journal of Finance* **31**, 351–367.
- [2] Chapovsky, D., Rennie, A. & Tavares, P. (2006). *Stochastic Intensity Modelling for Structured Credit Exotics*, working paper, Merrill Lynch.
- [3] Duffie, D. (1998). *First-to-Default Valuation*, Institut de Finance, University of Paris, Dauphine, and Graduate School of Business, Stanford University.
- [4] Duffie, D. (1999). Credit swap valuation, *Financial Analysts Journal* January/February, 73–87.
- [5] Duffie, D. & Singleton, K. (1999). Modeling term structures of defaultable bonds, *Review of Financial Studies* **12**(4), 687–720.
- [6] Inglis, S. & Lipton, A. (2007). Factor models for credit correlation, *Risk Magazine* **20**, 110–115.
- [7] Jarrow, R.A., Lando, D. & Turnbull, S.M. (1997). A Markov model for the term structure of credit spreads, *Review of Financial Studies* **10**, 481–523.
- [8] Jarrow, R.A. & Turnbull, S.M. (1992). Credit risk: drawing the analogy, *Risk Magazine* **5**(9), 63–70.
- [9] Jarrow, R.A. & Turnbull, S.M. (1995). Pricing derivatives with credit risk, *Journal of Finance* **50**, 53–85.

- [10] Li, D.X. (1998). *Constructing a Credit Curve*, *Credit Risk*, A RISK Special report (November 1998), pp. 40–44.
- [11] Longstaff, F. & Schwartz, E. (1995). Valuing risky debt: a new approach, *Journal of Finance* **50**, 789–820.
- [12] Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.
- [13] Schonbucher, P.A. (2002/3). Tree implementation of a credit spread model for credit derivatives, *Journal of Computational Finance* **6**(2), 1–38.
- [14] Schonbucher, P. & Schubert, D. (2001). *Copula Dependant Default Risk in Intensity Models*, working paper, Bonn University.

Related Articles

Credit Default Swaps; Duffie–Singleton Model; Hazard Rate; Multiname Reduced Form Models; Nested Simulation; Reduced Form Credit Risk Models.

JON GREGORY

Default Barrier Models

The modeling of default from an economic point of view is a great challenge due to the binary and low probability nature of such an event. Default barrier models provide an elegant solution to this challenge since they link the default event to the point at which some continuously evolving quantity hits a known barrier. In structural models of credit risk (*see Structural Default Risk Models*) the process and the barrier are interpreted in terms of capital structure of the firm as the value of the firm and its liabilities. More generally, one can view the process and the barrier as state variables that need not necessarily be observable.

Single-name Models

In the classic Merton framework [12], the value of a firm (asset value) is considered to be stochastic and default is modeled as the point where the firm is unable to pay its outstanding liabilities when they mature. The asset value is modeled as a geometric Brownian motion:

$$\frac{dV_t}{V_t} = \mu dt + \sigma dW \quad (1)$$

where μ and σ represent the drift and volatility of the asset, respectively, and dW is a standard Brownian motion. The original Merton model assumes that a firm has issued only a zero-coupon bond and will not therefore default prior to the maturity of this debt as illustrated in Figure 1. Denoting the maturity and face value of the debt by T and D respectively, the default condition can then be written as $V_T < D$. Through option pricing arguments, Merton then provides a link between corporate debt and equity *via* pricing formulae based on the value of the firm and its volatility (analogously to options being valued from spot prices and volatility). The problem of modeling default is transformed into that of assessing the future distribution of firm value and the barrier where default would occur. Such quantities can be estimated nontrivially from equity data and capital structure information. This is then the key contribution of Merton approach in that low-frequency binary events can be modeling *via* a continuous process and calibrated using high-frequency data.

Practical Extensions of the Merton Approach

The classic Merton approach has been extended by many authors such as Black and Cox [2] and Leland [10]. Commercially, it has been developed by KMV™ (now Moody's KMV) with the aim of predicting default *via* the assessment of 1-year default probability defined as EDF™ (expected default frequency). A more recent and related, although simpler, approach is CreditGrades™.

Moody's KMV Approach. This approach [8, 9] was inspired by the Merton approach to default modeling and aimed to lift many of the stylized assumptions and model the evolution and future default of a company in a realistic fashion. A key aspect of this is to account for the fact that a firm may default at any time but will not necessarily default immediately when they are bankruptcy insolvent (when $V_t < D$). Hence a challenge is to work out exactly where the default barrier is. KMV do this *via* considering both the short-term and long-term liabilities of the firm. Their approach can be broadly summarized in three stages:

- estimation of the market value and volatility of a firm's assets;
- calculation of the distance to default which is an index measure of default risk; and
- scaling of the distance to default to the actual probability of default using a default database.

The distance to default (DD) measure, representing a standardized distance from which a firm is above its default threshold, is defined by^a

$$DD = \frac{\ln(V/D) + (\mu - 0.5\sigma^2)T}{\sigma\sqrt{T}} \quad (2)$$

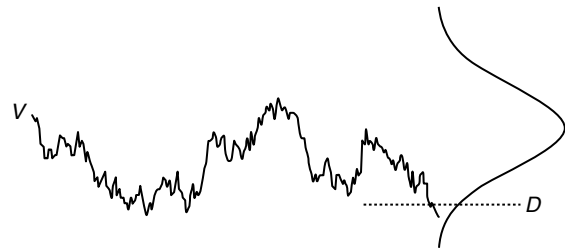


Figure 1 Illustration of the traditional Merton approach to modeling default based on the value of the firm being below the face value of debt at maturity

The default probability would then be given by $p_d = \Phi(-DD)$. A key element of the approach is to recognize the model risk inherent in this approach and rather to estimate the default probability empirically from many years of default history (and the calculated DD variables). We therefore ask ourselves the following question: for a firm with a DD of 4.0 (say), how often have firms with the same DD defaulted historically? The answer is likely to be considerably higher than the theoretical result of $\Phi(-4.0) = 0.003\%$. This mapping of DD to actual default probability could be thought of as an empirical correction for the non-Gaussian behavior of firm value.

CreditGrades Approach. The aim of CreditGrades is rather similar to that of KMV except that the modeling framework [3] is rather simpler, in particular without using empirical data in order to map to an eventual default probability. In the Credit Grades approach, the default barrier is given by

$$LD = \bar{L}De^{\lambda Z - \lambda^2/2} \quad (3)$$

where Z is a standard normal variable, D is the “debt per share”, $\bar{L}$ is an average recovery level, and λ creates an uncertainty in the default barrier. The level of the default barrier and the asset return are independent. Hence the main differences between the traditional Merton approach and CreditGrades is that the latter approach assumes that default can occur at any time when the asset process has dropped to a level of LD , whereas the Merton framework assumes $\bar{L} = 1$ and $\lambda = 0$ and no default prior to the maturity of the debt. CreditGrades recommends values of $\bar{L} = 0.5$ and $\lambda = 0.3$. A sensitivity analysis of these parameters should give the user a very clear understanding of the uncertainties inherent in estimating default probability.

Portfolio Models

While default barrier models have proved very useful for assessing single-name default probability and supporting trading strategies such as capital structure arbitrage, arguably an even more significant development has been their application in credit portfolio models. The basic strength of the default barrier approach is to provide the transformation necessary to

model default events *via* a multivariate normal distribution driven by asset correlations. The intuition of the approach makes it possible to add complexities such as credit migrations and stochastic recovery rates into the model.

Default Correlation

Consider modeling the joint default probability of two entities. Using the standard definition of a correlation coefficient, we can write the joint default probability as

$$p_{AB} = p_A p_B + \rho_{AB} \sqrt{p_A(1-p_A)p_B(1-p_B)} \quad (4)$$

where p_A and p_B are the individual default probabilities and ρ_{AB} is the default correlation. Assuming, without loss of generality, that $p_A \leq p_B$ and since the joint default probability can be no greater than the smaller of the individual default probabilities, we have

$$\begin{aligned} \rho_{AB} &= \frac{p_{AB} - p_A p_B}{\sqrt{p_A(1-p_A)p_B(1-p_B)}} \\ &\leq \frac{p_A - p_A p_B}{\sqrt{p_A(1-p_A)p_B(1-p_B)}} \\ &= \sqrt{\frac{p_A(1-p_B)}{p_B(1-p_A)}} \end{aligned} \quad (5)$$

This shows that the default correlation cannot be +1 (or indeed *via* a similar argument -1) unless the individual default probabilities are equal. There is therefore a maximum (and minimum) possible default correlation that changes with the underlying default probabilities. This suggests a need for more economic structure to model joint default probability.

Default Barrier Approach

Suppose that we write default as being driven by a standard Gaussian variable X_i being below a certain level $k = \Phi^{-1}(p)$. We can interpret X as being an asset return in the classic Merton sense, with k being a default barrier. Now joint default probability is readily defined *via* a bivariate Gaussian distribution:

$$p_{AB} = \Phi_2(\Phi^{-1}(p_A), \Phi^{-1}(p_B); \lambda_{AB}) \quad (6)$$

where Φ_2 is a cumulative bivariate cumulative distribution function and λ_{AB} is the “asset correlation”.

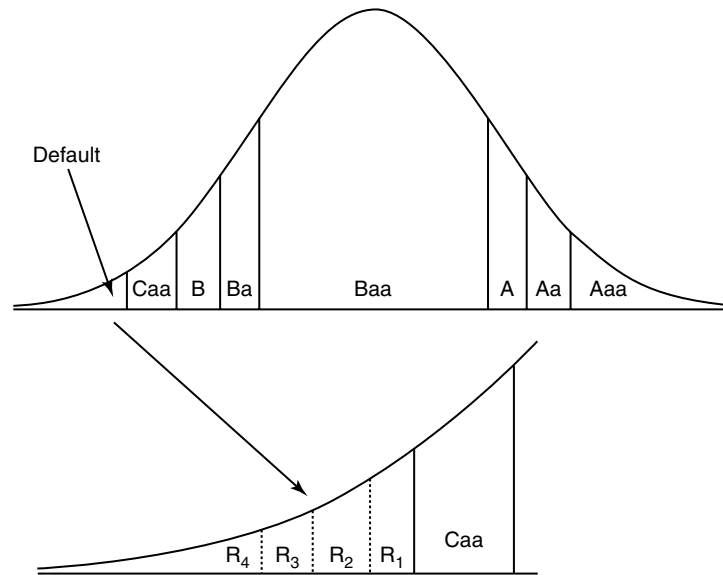


Figure 2 Illustration of the mapping of default and credit migrations thresholds as used in the CreditMetrics approach. The default region is also shown with additional thresholds corresponding to different recovery values with $R_1 < R_2 < R_3 < R_4$

Multiple names can be handled *via* a multivariate Gaussian distribution^b with Monte Carlo simulation or various factor-type approaches used for the calculation of multiple defaults and/or losses.

Although there is a clear link between this simple approach and the multidimensional Merton model, we have ignored the full path of the asset value process and linked default to just a single variable X_i . A more rigorous time-dependent approach can be found in [7], which is much more complex and time consuming to implement. In practice, the one-period approach is rather similar to the full approach for relative small default probabilities.

CreditMetrics

CreditMetrics [6], first published in 1997, is a credit portfolio model based on the multivariate normal default barrier approach. This framework assumes a default barrier as described above and also considers the mapping of credit migration probabilities onto the same normal variable. A downgrade can therefore be seen as a less extreme move not causing default. In addition to credit migrations, one can also superimpose different recovery rates onto the same mapping so that there is more than one default barrier with lower barriers representing more severe default

and therefore a lower recovery value; for example, see [1]. An illustration of the mapping is shown in Figure 2.

Regulatory Approaches

Basel 2. A key strength of the above framework is that defaults, credit migrations, and recovery rates can be modeled within a single intuitive framework with correlation parameters estimated from equity data. While other credit portfolio modeling frameworks have been proposed, the CreditMetrics style approach has been the most popular. Indeed, the Basel 2 formula [4] can be seen as arising from a simplified version of this approach with the following assumptions:

- no credit migration or stochastic recovery and
- infinitely large homogeneous portfolio.

Rating Agency Approaches to Structured Finance.

With the massive growth of the collateralized debt obligation (CDO) came a need for rating agencies (*see Structured Finance Rating Methodologies*) to model the risk inherent in a CDO structure with a view to assigning a rating to some or all of the tranches of the CDO capital structure. Rating a

tranche of a CDO is essentially the same problem as estimating capital on a credit portfolio and hence it may come as no surprise that the rating agencies models were based on default barrier approaches. The rating agencies models can be thought of as therefore heavily following the CreditMetrics approach. The credit crisis of 2007 brought very swift criticism of rating agency modeling approaches to rating all types of structure finance and CDO structures. This was related largely to poor assessment of the model parameters (specifically rather optimistic default probabilities and correlation assumptions) rather than a failure of the model itself.

CDO Pricing

A final and perhaps most exciting (although not for necessarily positive reasons) application of the default barrier approach is in the pricing of synthetic CDO structures. The market standard approach for pricing CDOs follows the work of Li [11] (*see Gaussian Copula Model*) who models time of default in a multivariate normal framework:

$$\begin{aligned} \Pr(T_A < 1, T_B < 1) \\ = \Phi_2(\Phi^{-1}(F_A(T_A)), \Phi^{-1}(F_B(T_B)); \gamma) \end{aligned} \quad (7)$$

where F_A and F_B are the distribution functions for the survival times T_B and T_B and γ is a correlation parameter. At first glance, although this uses the same multivariate distribution, or copula^c, this approach initially does not seem to be a default barrier model. However, as noted in [11], for a single period, the approaches are identical. Furthermore, as shown by Laurent and Gregory [5], the pricing of a synthetic CDO requires just the knowledge of loss distributions at each date up to the contractual maturity date (and not any further dynamical information). Hence, we can think of the Li approach as being again similar^d to the traditional framework of credit portfolio modeling, following CreditMetrics and ultimately inspired by the Merton approach to modeling default *via* the hitting time of a barrier. The recent strong criticism linking the model in [11] to the credit crisis [13] does not fairly consider the rather naïve calibration use of the model that has caused many of the problems in structured finance.

Conclusions

We have described the range of default barrier models used in default probability estimation, capital structure trading, credit portfolio management, regulatory capital calculations, and pricing and rating CDO products. The intuition that default can be modeled as a hitting of a barrier has been crucial to the rapid development of credit risk models. For credit portfolio risk in particular, the default barrier approach has been key to the development of models for many different purposes, driven from the same underlying structural framework. Given that some applications of the approach (most notably rating agency models and CDO pricing) have received large criticism, it is worth pointing out that one can only discredit the entire framework (including any multidimensional Merton approach) or realize that it is a misuse of the model rather than the model itself that lies at the heart of the problems.

End Notes

^aIn the proprietary Moody's KMV implementation, the default point is not the face value of debt but the current book value of their liabilities. This is often computed as short-term liabilities plus half long-term liabilities.

^bIt should be noted that alternatives to a Gaussian distribution (e.g., student-*t*) can and have been considered although the Gaussian approach has remained most common.

^cThis approach has become known as the *Gaussian copula model* which is perhaps confusing since the key point of the approach is the representation of the joint distribution of default times and not the choice of a Gaussian copula or multivariate distribution.

^dLi was at the time working at JP Morgan and so this is not surprising.

References

- [1] Arvanitis, A., Browne, C. Gregory, J. & Martin, R. (1998). A credit risk toolbox, *Risk* December, 50–55.
- [2] Black, F. & Cox, J. (1976). Valuing corporate securities: some effects of bond indenture provisions, *Journal of Finance* **31**, 351–367.
- [3] Finger, C., Finkelstein, V., Pan, G., Lardy, J.P. & Tiemey, J. (2002). *Credit-Grades Technical Document*, RiskMetrics Group.
- [4] Gordy, M. (2003). A risk-factor model foundation for ratings-based bank capital rules, *Journal of Financial Intermediation* **12**, 199–232.

-
- [5] Gregory, J. & Laurent, J.-P. (2005). Basket default swaps, CDO's and factor copulas, *Journal of Risk* **7**(4), 103–122.
 - [6] Gupton, G.M., Finger, C.C. & Bhatia, M. (1997). *CreditMetrics Technical Document*, Morgan Guaranty Trust Company, New York.
 - [7] Hull, J., Predescu, M. & White, A. (2005). *The Valuation of Correlation-Dependent Credit Derivatives Using a Structural Model*, working paper, available at SSRN: <http://ssrn.com/abstract=686481>
 - [8] Kealhofer, S. (2003). Quantifying default risk I: default prediction, *Financial Analysts Journal* **59**(1), 33–44.
 - [9] Kealhofer, S. & Kurbat, M. (2002). *The Default Prediction Power of the Merton Approach, Relative to Debt Ratings and Accounting Variables*, KMV LLC, Mimeo.
 - [10] Leland, H. (1994). Corporate debt value, bond covenants, and optimal capital structure, *Journal of Finance* **49**, 1213–1252.
 - [11] Li, D.X. (2000). On default correlation: a Copula approach, *Journal of Fixed Income* **9**, 43–54.
 - [12] Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.
 - [13] Wired Magazine: 17.03 (2009). *Recipe for Disaster: The Formula That Killed Wall Street*.

Related Articles

Credit Risk; Structural Default Risk Models.

JON GREGORY

Multiname Reduced Form Models

Currently, there are three established approaches for describing the default of a single credit: (i) reduced-form; (ii) structural; and (iii) hybrid. It has been an outstanding goal for many researchers to extend these approaches to baskets of several (potentially many) credits. In this article, we concentrate on the reduced-form approach and show how it works in single-name and multiname settings.

Single-name Intensity Models

For a single name, the main assumptions of the reduced-form model are as follows [8, 9, 12]. The name defaults at the first time a Cox process jumps from 0 to 1. The default intensity (hazard rate) $X(t)$ of this process is governed by a mean-reverting nonnegative jump-diffusion process

$$dX(t) = f(t, X(t)) dt + g(t, X(t)) dW(t) + J dN(t), \quad X(0) = X_0 \quad (1)$$

where $W(t)$ is a standard Wiener process, $N(t)$ is a Poisson process with intensity $\lambda(t)$, and J is a positive jump distribution; W, N, J are mutually independent. It is clear that we have to impose the following constraints:

$$f(t, 0) \geq 0, \quad f(t, \infty) < 0, \quad g(t, 0) = 0 \quad (2)$$

plus a number of other technical conditions to ensure that $X(t)$ stays nonnegative and is mean reverting.

For analytical convenience (rather than for stronger reasons), it is customary to assume that X is governed by the square-root stochastic differential equation (SDE):

$$dX(t) = \kappa(\theta(t) - X(t)) dt + \sigma\sqrt{X(t)} dW(t) + J dN(t), \quad X(0) = X_0 \quad (3)$$

with exponential (or hyperexponential) jump distribution [4]. However, for practical purposes it is more convenient to consider discrete jump distributions with jump values $J_m > 0$, $1 \leq m \leq M$, occurring with probabilities $\pi_m > 0$; such distributions are more flexible than parametric ones because they allow one to place jumps where they are needed.

In this framework, the survival probability of the name from time 0 to time T has the form

$$q(0, T) = \mathbb{E}_0 \left\{ e^{-\int_0^T X(t') dt'} \right\} = \mathbb{E}_0 \{ e^{-Y(T)} \} \quad (4)$$

where $Y(t)$ is governed by the following degenerate SDE:

$$dY(t) = X(t) dt, \quad Y(0) = 0 \quad (5)$$

More generally, the survival probability from time t to time T conditional on no default before time t has the form

$$\begin{aligned} q(t, T | X(t), Y(t)) &= \mathbb{I}_{(\tau > t)} \mathbb{E}_t \left\{ e^{-\int_t^T X(t') dt'} \middle| X(t), Y(t) \right\} \\ &= e^{Y(t)} \mathbb{I}_{(\tau > t)} \mathbb{E}_t \{ e^{-Y(T)} | X(t), Y(t) \} \end{aligned} \quad (6)$$

where τ is the default time and $\mathbb{I}_{(\tau > t)}$ is the corresponding indicator function. This expectation, and, more generally, expectations of the form $\mathbb{E}_t \{ e^{-\xi Y(T)} | X(t), Y(t) \}$, can be computed by solving the following augmented partial differential equation (PDE) (see [10], Chapter 13):

$$\mathcal{L}V(t, T, X, Y) + X V_Y(t, T, X, Y) = 0 \quad (7)$$

$$V(T, T, X, Y) = e^{-\xi Y} \quad (8)$$

where

$$\begin{aligned} \mathcal{L}V &\equiv V_t + \kappa(\theta(t) - X) V_X + \frac{1}{2} \sigma^2 X V_{XX} \\ &\quad + \lambda \sum_m \pi_m [V(X + J_m) - V(X)] \end{aligned} \quad (9)$$

Specifically, the following relation holds:

$$\mathbb{E}_t \{ e^{-\xi Y(T)} | X(t), Y(t) \} = V(t, T, X(t), Y(t)) \quad (10)$$

The corresponding solution can be written in the so-called affine form:

$$V(t, T, X, Y) = e^{a(t, T, \xi) + b(t, T, \xi)X - \xi Y} \quad (11)$$

where a, b are functions of time governed by the following system of ordinary differential equations (ODEs):

$$\begin{cases} \frac{da(t, T, \xi)}{dt} = -\kappa \theta(t) b(t, T, \xi) \\ \quad - \lambda \sum_m \pi_m [e^{J_m b(t, T, \xi)} - 1] \\ \frac{db(t, T, \xi)}{dt} = \xi + \kappa b(t, T, \xi) \\ \quad - \frac{1}{2} \sigma^2 b^2(t, T, \xi) \end{cases} \quad (12)$$

$$a(T, T, \xi) = 0, \quad b(T, T, \xi) = 0 \quad (13)$$

While in the presence of discrete jumps this system cannot be solved analytically, it is very easy to solve it numerically *via* the standard Runge–Kutta method. The survival probability $q(0, T)$ and default probability $p(0, T)$ have the form

$$\begin{aligned} q(0, T) &= e^{a(0, T, 1) + b(0, T, 1)X_0} \\ p(0, T) &= 1 - q(0, T) = 1 - e^{a(0, T, 1) + b(0, T, 1)X_0} \end{aligned} \quad (14)$$

Assuming for simplicity that the short interest rate $r(t)$ is deterministic and the protection payments are made continuously, we can write the value U of a credit default swap (CDS) paying an up-front amount v and a coupon s in exchange for receiving $1 - R$

(where R is the default recovery) on default as follows:

$$U = -v + V(0, X_0) \quad (15)$$

Here, $V(t, X)$ solves the following pricing problem:

$$\mathcal{L}V(t, X) - (r + X)V(t, X) = s - (1 - R)X \quad (16)$$

$$V(T, X) = 0 \quad (17)$$

where $\mathcal{L}$ is given by expression (9). Using Duhamel's principle, we obtain the following expression for V :

$$\begin{aligned} V(t, X) &= -s \int_t^T D(t, t') e^{a(t, t', 1) + b(t, t', 1)X} dt' \\ &\quad - (1 - R) \int_t^T D(t, t') d[e^{a(t, t', 1) + b(t, t', 1)X}] \end{aligned} \quad (18)$$

where

$$D(t, t') = e^{-\int_t^{t'} r(t'') dt''} \quad (19)$$

is the discount factor between two times t and t' . Accordingly,

$$\begin{aligned} U &= -v - s \int_0^T D(0, t') (1 - p(0, t')) dt' \\ &\quad + (1 - R) \int_0^T D(0, t') dp(0, t') \end{aligned} \quad (20)$$

For a given up-front payment v , we can represent the corresponding par spread $\hat{s}$ (i.e., the spread that makes the value of the corresponding CDS zero) as follows:

$$\hat{s}(T) = \frac{-v + (1 - R) \int_0^T D(0, t') dp(0, t')}{\int_0^T D(0, t') (1 - p(0, t')) dt'} \quad (21)$$

It is clear that the numerator represents the payout in the case of default, while the denominator represents the risky DV_{01} . Conversely, for a given

spread we can represent the par up-front payment in the form

$$\begin{aligned} \hat{v} = & -s(T) \int_0^T D(0, t') (1 - p(0, t')) dt' \\ & + (1 - R) \int_0^T D(0, t') dp(0, t') \end{aligned} \quad (22)$$

In these formulas, we implicitly assume that the corresponding CDS is fully collateralized, so that in the event of default $1 - R$ is readily available. Shortly, we will evaluate CDS spreads in the presence of the counterparty risk.

In general, there is not enough market information to calibrate the diffusion and jump parts. So, typically, they are viewed as given constants, and the mean-reversion level $\theta(t)$ is calibrated in such a way that the whole par spread curve is matched.

Multiname Intensity Models

The Two-name Case

It is very tempting to extend the above framework to cover several correlated names. For example, consider two credits, A, B and assume for simplicity that their default intensities coincide,

$$X_A(t) = X_B(t) = X(t) \quad (23)$$

and both names have the same recovery $R_A = R_B = R$. For a given maturity T , the default event correlation ρ is defined as follows:

It is clear that

$$\begin{aligned} p_A(0, T) &= p_B(0, T) = p(0, T) \\ &= 1 - e^{a(0, T, 1) + b(0, T, 1)X_0} \end{aligned} \quad (27)$$

Simple calculation yields

$$\begin{aligned} p_{AB}(0, T) &= \mathbb{E}_0 \left\{ e^{-\int_0^T (X_A(t') + X_B(t')) dt'} \right\} \\ &\quad + p_A(0, T) + p_B(0, T) - 1 \\ &= \mathbb{E}_0 \left\{ e^{-2 \int_0^T X(t') dt'} \right\} + 2p(0, T) - 1 \end{aligned} \quad (28)$$

so that

$$\begin{aligned} \rho(0, T) &= \frac{\mathbb{E}_0 \left\{ e^{-2 \int_0^T X(t') dt'} \right\} - (1 - p(0, T))^2}{p(0, T)(1 - p(0, T))} \\ &= \frac{e^{a(0, T, 2) + b(0, T, 2)X_0} - e^{2a(0, T, 1) + 2b(0, T, 1)X_0}}{(1 - e^{a(0, T, 1) + b(0, T, 1)X_0}) e^{a(0, T, 1) + b(0, T, 1)X_0}} \end{aligned} \quad (29)$$

It turns out that in the absence of jumps, the corresponding event correlation is very low [12]. However, if large positive jumps are added (while overall survival probability is preserved), then correlation can increase all the way to one. Assuming that

$$\rho(0, T) = \frac{P(\tau_A \leq T, \tau_B \leq T) - P(\tau_A \leq T)P(\tau_B \leq T)}{\sqrt{P(\tau_A \leq T)(1 - P(\tau_A \leq T))P(\tau_B \leq T)(1 - P(\tau_B \leq T))}} \quad (24)$$

$$\rho(0, T) = \frac{p_{AB}(0, T) - p_A(0, T)p_B(0, T)}{\sqrt{p_A(0, T)(1 - p_A(0, T))p_B(0, T)(1 - p_B(0, T))}} \quad (25)$$

where τ_A, τ_B are the default times, and

$$\begin{aligned} p_A(0, T) &= P(\tau_A \leq T), \quad p_B(0, T) = P(\tau_B \leq T) \\ p_{AB}(0, T) &= P(\tau_A \leq T, \tau_B \leq T) \end{aligned} \quad (26)$$

$T = 5y, \kappa = 0.5, \sigma = 7\%$, and $J = 5.0$, we illustrate this observation in Figure 1.

In the two-name portfolio, we can define two types of CDSs which depend on the correlation: (i) the first-to-default (FTD) swap; (ii) the

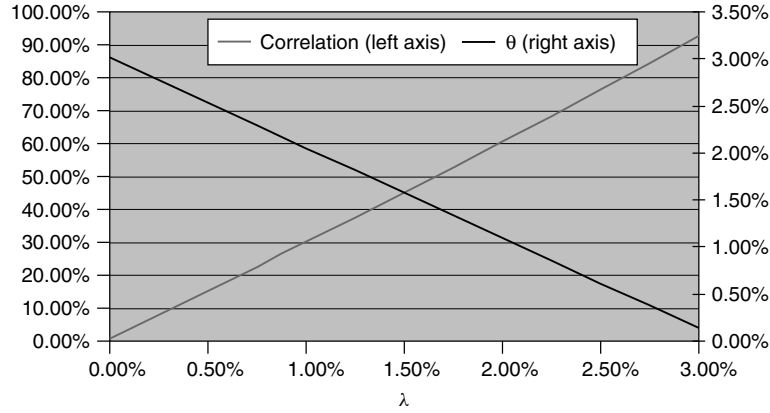


Figure 1 Correlation ρ and mean-reversion level $\theta = X_0$ as functions of jump intensity λ . Other parameters are as follows: $T = 5y$, $\kappa = 0.5$, $\sigma = 7\%$, and $J = 5.0$

second-to-default (STD) swap. The corresponding par spreads (assuming that there are no up-front payments) are

where V is the value of a fully collateralized CDS on name B with spread s , and $V_+ = \max\{V, 0\}$, $V_- = \min\{V, 0\}$. It is clear that the discount rate

$$\hat{s}_1(T) = \frac{(1-R) \int_0^T D(0, t') d \left[1 - e^{a(0, t') + b(0, t') X_0} \right]}{\int_0^T D(0, t') e^{a(0, t') + b(0, t') X_0} dt'} \quad (30)$$

$$\hat{s}_2(T) = \frac{(1-R) \int_0^T D(0, t') d \left[1 - \left(2e^{a(t', 1) + b(t', 1) X_0} - e^{a(t', 2) + b(t', 2) X_0} \right) \right]}{\int_0^T D(0, t') \left(2e^{a(t', 1) + b(t', 1) X_0} - e^{a(t', 2) + b(t', 2) X_0} \right) dt'} \quad (31)$$

It is clear that the relative values of $\hat{s}_1, \hat{s}_2$ very strongly depend on whether or not jumps are present in the model (see Figure 2).

However, an even more important application of the above model is the evaluation of counterparty effects on fair CDS spreads. Let us assume that name A has written a CDS on reference name B . It is clear that the pricing problem for the value of the uncollateralized CDS $\tilde{V}$ can be written as follows:

$$\begin{aligned} \mathcal{L}\tilde{V}(t, X) - (r + 2X) \tilde{V}(t, X) \\ = s - (1-R)X - (RV_+(t, X) + V_-(t, X))X \end{aligned} \quad (32)$$

is increased from $r + X$, in equation (16), to $r + 2X$, in equation (32), since there are two cases when the uncollateralized CDS can be terminated due to default: when the reference name B defaults and when the issuer A defaults. The terms on the right represent a continuous stream of coupon payments, the amount received if B defaults before A , and the amount received (or paid) in case when A defaults before B . Although equation (32) is no longer analytically solvable, it can be solved numerically *via*, say, an appropriate modification of the classical Crank–Nicholson method. It turns out that in the presence of jumps the value of the fair par spread goes down dramatically.

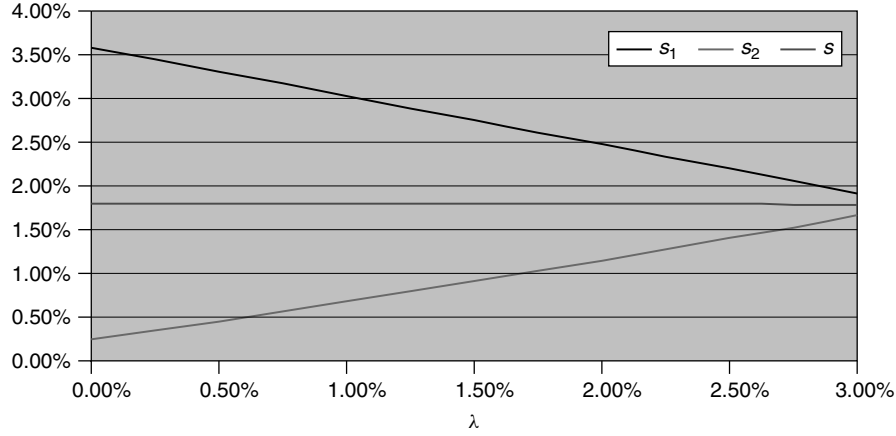


Figure 2 FTD spread $\hat{s}_1$, STD spread $\hat{s}_2$, and single-name CDS spread $\hat{s}$ as functions of jump intensity λ . Other parameters are the same as in Figure 1. It is clear that jumps are necessary to have $\hat{s}_1$ and $\hat{s}_2$ of similar magnitudes

The Multiname Case

The above modeling framework has been expanded in various directions and used as a basis for several coherent intensity-based models for credit baskets; see [2, 3, 6, 7, 11].

To start, we briefly summarize the affine jump-diffusion model of Duffie–Garleanu [3] and Mortensen [11]. Consider a basket of N names with equal unit notionals and equal recoveries R . Let us assume that the corresponding default intensities can be decomposed as follows:

$$X_i(t) = \beta_i X_c(t) + \tilde{X}_i(t) \quad (33)$$

where X_c is the common intensity driven by the following SDE:

$$\begin{aligned} dX_c(t) &= \kappa_c (\theta_c - X_c(t)) dt + \sigma_c \sqrt{X_c(t)} dW_c(t) \\ &\quad + J_c dN_c(t) \\ X_c(0) &= X_{c0} \end{aligned} \quad (34)$$

while $\tilde{X}_i$ are idiosyncratic intensities driven by similar SDEs:

$$\begin{aligned} d\tilde{X}_i(t) &= \kappa_i (\theta_i - \tilde{X}_i(t)) dt + \sigma_i \sqrt{\tilde{X}_i(t)} dW_i(t) \\ &\quad + \tilde{J}_i dN_i(t) \\ \tilde{X}_i(0) &= \tilde{X}_{i0} \end{aligned} \quad (35)$$

Here, $1 \leq i \leq N$. The processes $\tilde{X}_i(t)$, $X_c(t)$ are assumed to be independent. In this formulation, β_i are similar to the β_i appearing in the capital asset pricing model (CAPM). We note that θ_c, θ_i are assumed to be constant. In the original Duffie–Garleanu formulation, it was assumed that all $\beta_i = 1$. However, this assumption is very restrictive since it limits the magnitude of the common factor by the size of the lowest spread X_i , so that, in general, high correlation cannot be achieved. It was lifted in the subsequent paper by Mortensen. Of course, to preserve analyticity, one needs to impose very rigid conditions on the coefficients of the corresponding SDEs, since, in general, the sum of two affine processes is not an affine process. Specifically, the following should hold:

$$\kappa_i = \kappa_c = \kappa, \quad \sigma_i = \sqrt{\beta_i} \sigma_c, \quad \lambda_i = \lambda, \quad J_{im} = \beta_i J_{cm} \quad (36)$$

Even when the above constraints are satisfied, there are too many free parameters in the model. A reduction in their number is achieved by imposing the following constraints:

$$\frac{\beta_i \theta_c}{\beta_i \theta_c + \theta_i} = \frac{\lambda_c}{\lambda_c + \lambda} = \frac{X_c(0)}{X_c(0) + X_{ave}(0)} = \omega \quad (37)$$

where ω is a correlation-like parameter representing the systematic share of intensities, and $X_{ave}(0)$ is the average of $X_i(0)$. When ω is low, the dynamics of intensities is predominantly idiosyncratic, and it is systemic when ω is close to one.

Provided that equation (36) is true, the affine ansatz still holds, so that survival probabilities of individual names can be written in the form

$$\begin{aligned}
q_i(t, T | X_i(t)) &= \mathbb{I}_{(\tau_i > t)} \mathbb{E}_t \left\{ e^{-\int_t^T X_i(t') dt'} \middle| X_i(t) \right\} \\
&= \mathbb{I}_{(\tau_i > t)} \mathbb{E}_t \left\{ e^{-\beta_i [Y_c(T) - Y_c(t)]} \middle| X_c(t) \right\} \\
&\quad \times \mathbb{E}_t \left\{ e^{-[\tilde{Y}_i(T) - \tilde{Y}_i(t)]} \middle| \tilde{X}_i(t) \right\} \\
&= \mathbb{I}_{(\tau_i > t)} e^{a_c(t, T, \beta_i) + b_c(t, T, \beta_i) X_c(t) + a_i(t, T, 1) + b_i(t, T, 1) \tilde{X}_i(t)}
\end{aligned} \tag{38}$$

Moreover, conditioning the dynamics of spreads on the common factor $Y_c(T)$, we can write idiosyncratic survival probabilities as follows:

$$\begin{aligned}
q_i(t, T | \tilde{X}_i(t), Y_c(T)) &= \mathbb{I}_{(\tau_i > t)} e^{-\beta_i [Y_c(T) - Y_c(t)] + a_i(t, T, 1) + b_i(t, T, 1) \tilde{X}_i(t)}
\end{aligned} \tag{39}$$

$$\begin{aligned}
q_i(0, T | \tilde{X}_{i0}, Y_c(T)) &= e^{-\beta_i Y_c(T) + a_i(0, T, 1) + b_i(0, T, 1) \tilde{X}_{i0}}
\end{aligned} \tag{40}$$

First, we perform the calibration of the model parameters to fit 1y and 5y CDS spreads for individual names. Once this calibration is performed, we can apply the usual recursion and calculate the conditional probability of loss of exactly n names, $0 \leq n \leq N$, in the corresponding portfolio, or, equivalently, of the loss of size $(1 - R)n$, which we denote as $p(0, T, n | Y)$.

For a tranche of the portfolio which covers losses from the attachment point α to the detachment point δ , $0 \leq \alpha < \delta \leq 1$, the relative tranche loss is defined as follows:

$$\Lambda_{\alpha, \delta}(L) = \frac{\max\{\min\{L, \delta N\} - \alpha N, 0\}}{(\delta - \alpha) N} \tag{41}$$

Its conditional expectation has the form

$$p_{\alpha, \delta}(0, T | Y) = \sum_{n=0}^N \Lambda_{\alpha, \delta}((1 - R)n) p(0, T, n | Y) \tag{42}$$

In order to find the unconditional expectation, we have to integrate $p_{\alpha, \delta}(0, T | Y)$ with respect to the distribution $f(Y)$ of the common factor Y . The latter distribution can be found *via* the inverse Laplace transform of the function

$$\phi(\xi) = \int_0^\infty e^{-\xi Y} f(Y) dY = e^{a_c(0, T, \xi) + b_c(0, T, \xi) X_{c0}} \tag{43}$$

by numerically calculating the Bromwich integral in the complex plane

$$\begin{aligned}
f(Y) &= \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} e^{\xi Y} \phi(\xi) d\xi \\
&= \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} e^{\xi Y + a_c(0, T, \xi) + b_c(0, T, \xi) X_{c0}} d\xi
\end{aligned} \tag{44}$$

Both standard and more recent methods allow one to calculate the inverse transform without too much difficulty; see, for example, [1]. Finally, we calculate the unconditional expectation of the tranche loss by performing integration over the common factor:

$$p_{\alpha, \delta}(0, T) = \int_0^\infty p_{\alpha, \delta}(0, T | Y) f(Y) dY \tag{45}$$

Knowing this expectation, we can represent par spread and par up-front for the tranche in question by slightly generalizing formulas (21) and (22). In other words,

$$s_{\alpha, \delta}(T) = \frac{-v + \int_0^T D(0, t') dp_{\alpha, \delta}(0, t')}{\int_0^T D(0, t') (1 - p_{\alpha, \delta}(0, t')) dt'} \tag{46}$$

$$\begin{aligned}
v &= -s_{\alpha, \delta}(T) \int_0^T D(0, t') (1 - p_{\alpha, \delta}(0, t')) dt' \\
&\quad + \int_0^T D(0, t') dp_{\alpha, \delta}(0, t')
\end{aligned} \tag{47}$$

Equity tranches with $\alpha = 0$, $\delta < 1$ (and, in some cases, other junior tranches) are traded with a fixed spread, say $s = 5\%$, and an up-front determined by formula (47); more senior tranches are traded with zero up-front and spread determined by formula (46).

Treatment of super-senior tranches with $\delta = 1$ has to be slightly modified, but we do not discuss the corresponding details for the sake of brevity.

The affine jump-diffusion model allows one to price tranches of standard on-the-run indices, such as CDX and iTraxx with reasonable (but not spectacular) accuracy, and can be further used to price bespoke tranches; however, one can argue that the presence of the stochastic idiosyncratic components makes it unnecessarily complex. In any case, the very rigid relationships between the model parameters suggest that the choice of these components is fairly limited and rather artificial.

Two models without stochastic idiosyncratic components were independently proposed in the literature. The first one, due to Chapovsky *et al.* [2], assumes purely deterministic idiosyncratic components, and represents q_i as follows:

$$q_i(0, T | Y_c(T)) = e^{-\beta_i(T)Y_c(T) + \xi_i(T)} \quad (48)$$

where, X_c, Y_c are driven by SDEs (1) and (5), while $\xi_i(T)$ is calibrated to the survival probabilities of individual names. The second one, due to Inglis–Lipton [6], models conditional survival probabilities directly, and postulates that $q_i(0, T | Y_c)$ can be represented in the logit form

$$q_i(0, T | Y_c(T)) = \mathbb{E}_i \left\{ \frac{1}{1 + e^{Y_c(T) + \chi_i(T)}} \right\} \quad (49)$$

We now describe the Inglis–Lipton model in some detail. To calibrate the model to individual CDS spreads, we need to solve the following pricing problem:

$$\hat{\mathcal{L}}V(t, X, Y) + XV_Y(t, X, Y) = 0 \quad (50)$$

$$V(T, X, Y) = \frac{1}{1 + e^Y} \quad (51)$$

where

$$\begin{aligned} \hat{\mathcal{L}}V \equiv & V_t + f(t, X) V_X + \frac{1}{2} g^2(t, X) V_{XX} \\ & + \lambda \sum_m \pi_m [V(X + J_m) - V(X)] \end{aligned} \quad (52)$$

and determine $\chi_i(T)$ from the following algebraic equation (rather than a PDE):

$$V(0, 0, \chi_i(T)) = q_i(0, T), \quad 1 \leq i \leq N \quad (53)$$

As before, we can easily calculate the probability of loss of exactly n names, $0 \leq n \leq N$, $p(0, T, n | Y)$, conditional on Y . We can then solve the pricing equation (50) with the terminal condition

$$V_{\alpha, \delta}(T, X, Y) = p_{\alpha, \delta}(0, T | Y) \quad (54)$$

and find the expected losses for an individual tranche at time 0:

$$p_{\alpha, \delta}(0, T) = V_{\alpha, \delta}(0, X_0, 0) \quad (55)$$

Here, $p_{\alpha, \delta}(0, T | Y)$, $p_{\alpha, \delta}(0, T)$ have the same meaning as in equations (42) and (45). In order to price senior tranches rare but large jumps are necessary. Since, as a rule, we need to analyze several tranches with different attachments, detachments, and maturities at once, it is more convenient to solve the forward version of equation (50) and find $p_{\alpha, \delta}(0, T)$ by integration. Thus, we are in a paradoxical situation when it is more efficient to perform calibration to individual names backward and calibration to tranches forward, rather than the other way round.

When derivatives explicitly depending on the number of defaults, such as leveraged super-senior (LSS) tranches, are considered, the X, Y dynamics requires augmentation with the dynamics of the number of defaulted names n . Since we are dealing with a “pure birth” process, we can use the well-known results due to Feller [5] and others and obtain the following expression for the one-step transition probability:

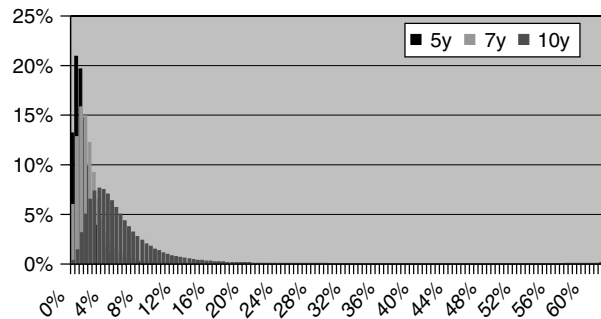
$$\begin{aligned} h(t, X, Y, n) &= \frac{-\sum_{n'=0}^n [p_t(t, T, n' | Y) + X p_Y(t, T, n' | Y)]}{p(t, T, n | Y)} \\ &= \frac{\sum_{n'=n+1}^N [p_t(t, T, n' | Y) + X p_Y(t, T, n' | Y)]}{p(t, T, n | Y)} \end{aligned} \quad (56)$$

The corresponding backward Kolmogoroff equation has the following form:

$$\begin{aligned} \hat{\mathcal{L}}V(t, X, Y, n) + XV_Y(t, X, Y, n) + h(t, X, Y, n) \\ \times [V(t, X, Y, n+1) - V(t, X, Y, n)] = 0 \end{aligned} \quad (57)$$

Table 1 Market quotes and full dynamic model calibration results. We quote par up-front payments with 5% spread for equity tranches, and par spreads for all other tranches^a

α	δ	5y	5y	7y	7y	10y	10y
		Market	Model	Market	Model	Market	Model
0%	3%	21.75%	21.76%	29.00%	28.89%	36.88%	36.94%
3%	6%	150.5	149.8	210.5	215.6	377.0	379.5
6%	9%	72.5	73.7	108.0	100.7	158.0	159.2
9%	12%	52.5	51.3	72.0	72.3	104.5	98.8
12%	22%	32.5	32.6	46.0	47.6	63.5	64.3
0%	100%	49.0	46.7	56.0	53.6	65.0	63.4

^aadapted from [7]**Figure 3** Loss distributions for 5y, 7y, 10y implied by the calibrated dynamic model (adapted from [7])

If need occurs, a multifactor extension of the above model can be considered.

Table 1 shows the quality of calibration achievable in the above framework for the on-the-run iTraxx index on November 9, 2007. We show the corresponding loss distributions in Figure 3.

This model can naturally be used to price bespoke baskets (as long as an appropriate standard basket is determined). It does not suffer from any of the drawbacks of the standard mapping approaches used for this purpose.

We note in passing that Inglis–Lipton [6] describe a static version of their model which is perfectly adequate for the purposes of pricing standard and bespoke tranches, even under the current extreme market conditions.

Conclusion

In general, multiname intensity models have many attractive features. They are naturally connected to single-name intensity models. In order to account for the observed tranche spreads in the market, they have to postulate periods of very high intensities which

gradually mean-revert to moderate and low levels. Mean-reversion of the default intensities serves as a useful mechanism which allows one to price tranches with different maturities in a coherent fashion. Of course, due to the presence of large jumps, it is very difficult to provide convincing hedging mechanisms in such models. However, since we assume that jumps are discrete, it is possible in principle to hedge a given bespoke tranche with a *portfolio* of standard tranches. This is a topic of active research and experimentation at the moment, and we hope to present the outcome of this research in the near future.

Acknowledgments

I am grateful to my colleagues S. Inglis, J. Manzano, A. Rennie, A. Sepp, and D. Shelton for illuminating discussions of the subject matter.

References

- [1] Abate, J. & Whitt, W. (1995). Numerical inversion of Laplace transforms of probability distributions, *ORSA Journal on Computing* 7(1), 36–43.

-
- [2] Chapovsky, A., Rennie, A. & Tavares, P. (2001). Stochastic intensity modeling for structured credit exotics, *The International Journal of Theoretical and Applied Finance* **10**, 633–652.
- [3] Duffie, D. & Garleanu, N. (2001). Risk and valuation of collateralized debt obligations, *Financial Analysis Journal* **57**, 41–59.
- [4] Duffie, D., Pan, J. & Singleton, K. (2000). Transform analysis and asset pricing for affine jump diffusions, *Econometrica* **68**, 1343–1376.
- [5] Feller, W. (1970). *An Introduction to Probability Theory and its Applications*, Wiley, New York, Vol. 1.
- [6] Inglis, S. & Lipton, A. (2007). Factor models for credit correlation, *Risk Magazine* **20**(12), 110–115.
- [7] Inglis, S., Lipton, A., Savescu, I. & Sepp, A. (2008). Dynamic credit models, *Statistics and its Interface* **1**, 211–227.
- [8] Jarrow, R. & Turnbull, S. (1995). Pricing options on financial securities subject to credit risk, *Journal of Finance* **50**, 53–85.
- [9] Lando, D. (1998). On Cox processes and credit risky securities, *Review of Derivatives Research* **2**, 99–120.
- [10] Lipton, A. (2001). *Mathematical Methods for Foreign Exchange*, World Scientific, Singapore.
- [11] Mortensen, A. (2006). Semi-analytical valuation of basket credit derivatives in intensity-based models, *The Journal of Derivatives* **13**(4), 8–26.
- [12] Schonbucher, P. (2003). *Credit Derivatives Pricing Models*, Wiley, Chichester.

ALEXANDER LIPTON

Default Time Copulas

Copulas are used in mathematical statistics to describe *multivariate distributions* in a way that separates the marginal distributions from the codependence structure. More precisely, any multivariate distribution can be “decomposed” into its *marginal distributions* and a multivariate distribution with uniform marginals. Suppose $X_1, \dots, X_n$ are real-valued stochastic variables with marginal distributions

$$f_i(x) = P(X_i \leq x), \quad i = 1, \dots, n \quad (1)$$

where the right-hand side denotes the probability that X_i takes a value less than or equal to x . Suppose further that C is a distribution function on the n -dimensional unit hypercube with uniform marginals.^a Then we can define a joint distribution of $(X_1, \dots, X_n)$ by

$$P(X_1 \leq x_1, \dots, X_n \leq x_n) = C(f_1(x_1), \dots, f_n(x_n)) \quad (2)$$

We say that C is the *copula function* of the joint distribution. Clearly, the copula function for a given distribution is unique. Existence, that is, the actual existence of a copula function for any joint distribution, is established by *Sklar’s Theorem* [3]. Given the definition of a copula, it is clear that a default time copula is a copula for the joint distribution of default times. Here, as in other applications in finance, the main advantage of using a copula formulation is that the marginal distributions are implied from the market, independent of information about mutual dependencies between default times. Specifically, the distribution of the time of default of a single firm can be implied^b from the par spread of the credit default swap (CDS) contracts on the debt of the firm. This distribution is represented by the “default curve”:

$$p_i(t) = P(\tau_i \leq t) \quad (3)$$

where τ_i is the stochastic default time of the i th firm.^c Once we have determined the marginal distributions of the default rates of single firms in this way, we may model mutual dependencies between these default times by choosing a suitable copula function

and writing the joint distribution of default times as in equation (2). From a practical point of view it is a great advantage that, by construction, the marginal distributions are unchanged under a change of copula. This allows us to preserve the calibration to market CDS quotes while adjusting the codependence structure.

Factor Copulas

In practice, copula functions are rarely specified directly for the default times. Instead, we introduce stochastic “default trigger variables” X_i such that we can identify events

$$\{X_i \leq h_i(t)\} \equiv \{\tau_i \leq t\} \quad (4)$$

for suitable nondecreasing functions $h_i : \mathbb{R}_+ \rightarrow \mathbb{R}$ such that

$$P(X_i \leq h_i(t)) = p_i(t) \quad (5)$$

We may regard the trigger variables as just a convenient mathematical device, but we may also follow Merton [2] and view X_i as the (return of the) value of the assets of the i th firm. With this interpretation, we may further interpret $h_i(T)$ for some fixed time horizon T as the face value of the firm’s debt maturing at T . In this picture, default coincides with insolvency.

One advantage of using default trigger variables rather than default times is that the codependency of firm values is more susceptible to economic reasoning. For example, we can think of asset values as being driven by a common factor representing general economic conditions. Then we would use a decomposition such as

$$X_i = f_i(Z) + \epsilon_i \quad (6)$$

where Z is the common factor, ϵ_i are idiosyncratic components independent of each other and of Z , and f_i are suitable “loading functions”. Note that, conditional on a given factor value, the trigger variables and, therefore, the default times, will be independent. The (unconditional) joint distribution is determined, for given distributions of Z and the ϵ_i s, by the loading functions f_i .

A default time copula specified by default triggers with the decomposition in equation (6) is called a *factor copula*. Most, if not all, copula models used in derivatives pricing are factor copulas.

Pricing with Copula Models

The generic application of default time copulas is in the pricing of CDO tranches, that is, tranches of a portfolio of debt instruments referencing a (large) number of issuers. Such a tranche is a special case of a security whose future cash flows is a function of the default times of the issuers. The present value of such a security is given by an expectation over the joint default time distribution, which, in the general case, has to be evaluated by Monte Carlo, that is, by random sampling from the distribution. However, as we shall now discuss, for certain types of securities, the expectation can be calculated by a much faster method if a factor copula is used.

Loss Distributions

Although it is true that a CDO tranche depends on the joint default time distribution, it does so in a rather special way since, in fact, it only depends on the total *loss* in the portfolio; in particular, it does not depend on the identity of the defaulted names, or on the order in which they default. More precisely, we can compute the value of a tranche if we know the distribution of the cumulated portfolio loss out to any time up to tranche maturity.^d As we shall now see, the computation of such loss distributions is particularly simple in a factor model.

We shall first show how to compute the distribution of portfolio loss to some fixed horizon t conditional on some given factor value z . To lighten the notation, we suppress the parameters z and t . Let p_i be the conditional probability that the i th issuer defaults and assume that the loss in default is given by some constant^e u . Further define

$$P_l^{(n)} = P(L^{(n)} = lu) \quad (7)$$

where $L^{(n)}$ is the default loss from the first n issuers (in some arbitrary order).

Then we have the following recursion relation (see [1])

$$P_l^{(n+1)} = (1 - p_{n+1})P_l^{(n)} + p_{n+1}P_{l-1}^{(n)} \quad (8)$$

which allows us to build the loss distribution for any portfolio from the trivial case of the empty portfolio

$$P_l^{(0)} = \delta_{l,0} \quad (9)$$

From the conditional loss distributions, we obtain the unconditional loss distribution by integration^f over z .

We remark that using equation (8) amounts to explicitly doing the convolution of the independent conditional loss distribution for each issuer in order to obtain the distribution of the portfolio loss. This convolution could also be done by Fourier techniques although this involves a somewhat greater computational burden. Note that by suitably inverting the convolution, one may compute the sensitivities of the tranche value to the parameters, for example, default probability, of each issuer. These are very important quantities in financial risk management.

Concluding Remarks

Models based on default time copulas are in widespread use for pricing and risk managing portfolio credit derivatives such as CDO tranches. The important special case of factor copulas combines the dual advantages of providing a clear economical interpretation of default time codependence and of allowing computationally efficient implementations.

The main practical limitation of copula models is that they are not *dynamic* models in the sense that they do not allow any conditioning on the future state of the world. This means that copula models cannot be reliably used, for example, in the pricing of options on tranches since here we have to be able to determine the distribution of the value of the underlying tranche conditioned on the state at option expiration time. To address such problems, we need a model that specifies the stochastic dynamics of a sufficient set of state variables. For example, we could specify the joint dynamics of all default intensities. Any such model would, of course, produce a joint default time distribution which would be describable by a copula and marginals. But this is not a one-to-one relationship since different dynamic models can produce the same copula. In this sense, the copula approach is more efficient for securities that depend only on the joint distribution of default times.

End Notes

^aThis simply means that $C : [0, 1]^n \rightarrow [0, 1]$ is nondecreasing in each argument, $C(0, \dots, 0) = 0$, $C(1, \dots, 1) = 1$, and that, for any i and any $y_i \in [0, 1]$,

$$\int_0^1 dy_1 \dots \int_0^1 dy_{i-1} \int_0^1 dy_{i+1} \dots \int_0^1 dy_n C(y_1, \dots, y_n) = y_i.$$

^bGiven suitable assumptions about *recovery* in default.

^cNote that this distribution is the so-called risk-neutral distribution, which differs from the real-world, or physical, distribution unless there is no *risk premium* associated with the risk of default.

^dIn practice, this is approximated by a finite set of times.

^eThis assumption is just for notational convenience; the extension to issuer specific, and possibly random, loss amounts is straightforward.

^fIf z has real dimension ≤ 3 , a quadrature scheme can be used, otherwise Monte Carlo integration is more efficient.

References

- [1] Andersen, L., Sidenius, J. & Basu, S. (2003). All your hedges in one basket, *RISK* November, 67–72.

- [2] Merton, R. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.

- [3] Sklar, A. (1959). *Fonctions de Répartition à n Dimensions et Leurs Marges*, Publications de l'Institut de Statistique de L'Université de Paris, Paris, Vol. 8, pp. 229–231.

Related Articles

Copulas: Estimation; Copulas in Econometrics; Copulas in Insurance; Exposure to Default and Loss Given Default; Gaussian Copula Model; Random Factor Loading Model (for Portfolio Credit); Recovery Rate.

JAKOB SIDENIUS

Gaussian Copula Model

Li [5] has introduced a copula function approach to credit portfolio modeling. In this approach, the author first introduces a random variable to denote the survival time for each credit and characterizes its properties using a density function or a hazard rate (see **Hazard Rate**). This allows us to move away from a one-period framework so that we could incorporate the term structure of default probabilities for each name in the portfolio. Then, the author introduces copula functions (see **Copulas: Estimation**) to combine information from all individual credits and further assumes a correlation structure among all credits. Mathematically, copula functions allow us to construct a joint distribution of survival times with given marginal distributions as specified by individual credit curves. This two-stage approach to forming a joint distribution of survival times has advantages. First, it incorporates all information on each individual credit. Second, we have more choices of different copula functions to form a good joint distribution to serve our purpose than if we assume a joint distribution of survival times from the start. While the normal copula function was used in [5] for illustration due to the simplicity of its economic interpretation of the correlation parameters and the relative ease of computation of its distribution function, the framework does allow use of other copula functions. We also discuss an efficient “one-step” simulation algorithm of survival times in the copula framework by exploring the mathematical property of copula functions in contrast to the period by period simulation as suggested earlier by others.

Default Information of a Single Name

To price any basket credit derivative structure, we first need to build a credit curve for each single credit in the portfolio, and then we need to have a default correlation model so that we can link all individual credits in the portfolio.

A credit curve for a company is a series of default probabilities to future dates. Traditionally, we use rating agency’s historical default experience to derive this information. From a relative value trading perspective, however, we rely more on market information from traded assets such as risky bond prices,

asset swap spreads, or, nowadays, directly the single-name term structure of default swap spreads to derive market implied default probabilities. These probabilities are usually called *risk neutral default probabilities*, which, in general, are much higher than the historical default probabilities for the rating class to which this company belongs. Mathematically, we use the distribution function of survival time to describe these probabilities. If we denote τ as an individual credit’s survival time which measures the length of time from today to the time of default, we use $F(t)$ as the distribution function defined as follows:

$$F(t) = \Pr[\tau \leq t] = 1 - S(t) \quad (1)$$

where $S(t)$ is called the *survival probability* up to time t . The marginal probabilities of defaults such as the ones over one-year periods, or hazard rates in continuous term, are usually called a *credit curve*. In general, for single-name default swap pricing, only a credit curve is needed in the same way as an interest rate curve is needed to price an interest rate swap.

Correlating Defaults through Copula Functions

Central to the valuation of the credit derivatives based on a credit portfolio is the default correlation. To put it in simple terms, default correlation measures the impact of one credit default on other credits. Intuitively, one would think of default correlation as being driven by some common macroeconomic factors. These factors tend to tie all industries into the common economic cycle, a sector-specific effect or a company-specific effect. From this angle, it is generally believed that default correlation is positive even between companies in different sectors. Within the same sector, we would expect companies to have an even higher default correlation since they have more commonalities. For example, overcapacity in the telecommunication industry after the internet/telecom bubble resulted in the default of numerous telecommunication and telephone companies. However, the sheer lack of default data means those assumptions are difficult to verify with any degree of certainty. Then we have to resort to an economic model to solve this problem.

From a mathematical point of view, we know the marginal distribution of survival time of each credit in the portfolio and we need to find a joint survival

time distribution function such that the marginal distributions are the same as the credit curves of individual credits. This problem cannot be solved uniquely. There exist a number of ways to construct a joint distribution with known marginals. Copula functions, used in multivariate statistics, provide a convenient way to specify any joint distribution with given marginal distributions.

A copula function (*see* **Copulas: Estimation**) is simply a specification of how to use the univariate marginal distributions to form a multivariate distribution. For example, if we have N -correlated uniform random variables $U_1, U_2, \dots, U_N$, then

$$\begin{aligned} C(u_1, u_2, \dots, u_N) \\ = \Pr\{U_1 < u_1, U_2 < u_2, \dots, U_N < u_N\} \quad (2) \end{aligned}$$

is the joint distribution function, which gives the probability that all of the uniforms are in the specified N -dimensional space cube. Using this joint distribution function C and N marginal distribution functions $F_i(t_i)$, which describe N credit curves, we form another function as follows: $C[F_1(t_1), F_2(t_2), \dots, F_N(t_N)]$. It can be shown that this function is a distribution function for the N -dimensional random vector of survival times where, as desired, the marginal distributions are $F_1(t_1), F_2(t_2), \dots, F_N(t_N)$; see [5]. So a copula function is nothing more than a joint distribution of uniform random variables from which we can build a joint distribution with a set of given marginals.

Then we need to solve two problems. First, which copula function should we use? Second, how do we calibrate the parameters in a copula function? Suppose we study a credit portfolio of two credits over a given period. The marginal default probabilities are given by the two credit curves constructed using market information or historical information. From an economic perspective, a company defaults when its asset falls below its liability. However, in the relative value trading environment, we know the default probability from the credit curve constructed using market information such as default swap spreads, asset swap spreads, or risky bond prices. Assume that there exists a standardized “asset return” X and a critical value x , and when $X \leq x$ the company would default, that is,

$$\begin{aligned} \Pr[X_1 \leq x_1] &= \Phi(x_1) = q_1 \\ \Pr[X_2 \leq x_2] &= \Phi(x_2) = q_2 \quad (3) \end{aligned}$$

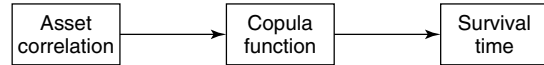
where Φ is the cumulative univariate standard normal distribution. We use Φ_n to denote the n -dimension cumulative normal distribution function. If we assume that the asset returns follow a bivariate normal distribution $\Phi_2(x, y, \rho)$ with correlation coefficient ρ , the joint default probability is given by

$$\begin{aligned} \Pr[X_1 \leq x_1, X_2 \leq x_2] \\ = \Pr[X_1 \leq \Phi^{-1}(q_1), X_2 \leq \Phi^{-1}(q_2)] \\ = \Phi_2[\Phi^{-1}(q_1), \Phi^{-1}(q_2), \rho] \quad (4) \end{aligned}$$

This expression suggests that we can use a Gaussian copula function with asset return correlations as parameters.

The above argument need not be associated with a normal copula. Any other copula function would be still able to give us a joint survival time distribution while preserving the individual credit curves. We have to use extra conditions in order to choose an appropriate copula function. When we compare two copula functions, we need to control the marginal distribution-free correlation parameter such as the rank correlation.

This approach gives a very flexible framework based on which we can value many basket structures. It can be expressed in the following graph:



We also present an efficient simulation algorithm here to implement our framework. To simulate correlated survival times, we introduce another sequence of random variables $X_1, X_2, \dots, X_n$ such that

$$X_i = \Phi^{-1}(F(\tau_i)) \quad (5)$$

where $\Phi^{-1}(\cdot)$ is a one-dimensional standard normal inverse function. $X_1, X_2, \dots, X_n$ follow a joint normal distribution with a given correlation matrix Σ . From this equation, we see that there is a one-to-one mapping between X_i and τ_i . Any problem associated with τ_i could be transformed into a problem associated with X_i , which follows a joint normal distribution. Then we could make use of an efficient calculation method of multivariate normal distribution function.

The correlation parameters, in the framework of our credit portfolio model, can be roughly interpreted

as the asset return correlation. However, in most practical uses of the current model, we either set the correlation matrix using one constant number or two numbers as the inter- and intraindustrial correlation for trading models. We could either use an economic model to asset correlation or we can calibrate the parameters using traded instruments involving correlation such as first-to-default or collateralized debt obligation (CDO) tranches.

The commonly used one or two correlation parameters are strongly associated with factor models for asset returns. For example, the one correlation parameter $\rho \geq 0$ corresponds to a one-factor asset return model where each asset return can be expressed as follows:

$$X_i = \sqrt{\rho} \cdot X_m + \sqrt{1 - \rho} \cdot \varepsilon_i \quad (6)$$

where X_m represents the common factor return and ε_i is the idiosyncratic risk associated with credit asset i . Vasicek [7] and Finger [3] use this one-factor copula for portfolio loss calculation. For a detailed discussion on this one-factor copula model, the reader is referred to these two references.

If we use two parameters, the interindustry correlation ρ_o and the intraindustrial correlation ρ_I , then for each credit of industry group $k = 1, 2, \dots, K$, we can express the asset return as follows [6]:

$$X_i = \sqrt{\rho_I - \rho_o} \cdot X_k + \sqrt{\rho_o} \cdot X_m + \varepsilon_i \quad (7)$$

Using these factor models, we can substantially reduce the dimensionality of the model. The number of independent factors then does not depend on the size of the portfolio. For example, for a portfolio whose credits belong to 10 industries, we just need to use 11 independent factors, one factor for each industry and one common factor for all credits. We could substantially improve the efficiency of our simulation or analytical approach once we exploit the property of the factor models embedded in the correlation structure. Some other orthogonal transformations such as the ones obtained by applying principal component analysis could also be used to reduce the dimension.

Loss Distribution

For a given credit portfolio, the first information investors would like to know is its loss distribution over a given time horizon in the future such as

1 year or 5 years. This would give the investor some idea about the possible default loss of his investment in the next few years. The information we need to use in our framework is as follows: the credit curve of each credit that characterizes the default property over the time horizon, the recovery assumption, and the asset correlation structures. Many useful risk measurements, such as the expected loss, the unexpected loss, or the standard deviation of loss, the maximum loss, Value-at-Risk (VaR) or the conditional shortfall, could be obtained easily once the total loss distribution is calculated.

Here we study the property of the loss distribution using a numerical example. The base case used is as given in Table 1.

Figure 1 shows the excess loss distribution where the x -axis is the loss amount and y -axis is the probability of loss more than a given amount in the x -axis. All excess loss functions would start from 1 and gradually go to zero. If we include the zero loss in the probability calculation, then the probability of having nonnegative losses is always 1. We purposely exclude the zero loss in the calculation so that we can see the probability of having zero loss in the graph explicitly. Let us define the excess loss more precisely. Suppose that L represents the total loss of the portfolio, which is a random variable, since we do not know for sure what value it takes. For a given set of loss amounts $l_0, l_1, \dots, l_n$, we can calculate the probability of excess loss $p_0, p_1, \dots, p_n$ as follows:

$$p_i = S(l_i) = \Pr[L > l_i] \quad (8)$$

The excess loss distribution essentially depicts (l_i, p_i) . The reason we use excess loss distribution instead of loss distribution, which is defined as $F(l_i) = 1 - S(l_i)$, is mainly due to the fact that many interesting properties of the loss distribution can be viewed more explicitly from the excess loss distribution graph than from the ordinary loss distribution

Table 1 Assumptions on a Credit Portfolio

Number of Assets	100
Credit spread	200 bps
Correlation	50%
Maturity	5 years
Recovery	30%

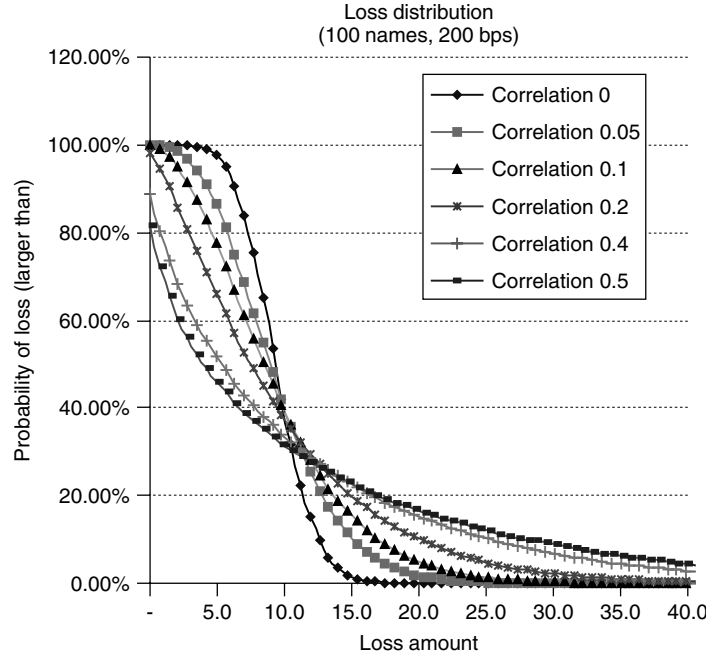


Figure 1 Excess Loss Distribution

graph. For example, the expected loss using the density function $f(l)$ of the loss distribution can be calculated as follows:

$$\mu_L = E(L) = \int_0^\infty l \cdot f(l) dl = \int_0^\infty S(l) dl \quad (9)$$

which is just the area below the excess loss distribution line. Some other quantity such as the expected loss of tranch securities (loss with a deductible and a ceiling) could also be more simply expressed if we use excess loss function. We discuss this point in the next section when we discuss about CDO pricing.

Figure 1 shows the impact of correlation on the total excess loss distribution. From the graph we see that the probability of having zero loss increases from almost 0 to about 20% when correlation changes from 0 to 50%. The default probability over 5 years for each name is $1 - e^{-5 \cdot 0.2 / (1 - 30\%)} = 13.31\%$, and the probability of having no default of a portfolio with 100 independent names is $(1 - 13.31\%)^{100}$, practically 0. However, when the correlation is high, default occurs more in bulk, which makes the probability of having zero loss go up to 20%. When correlation is high, more loss would be pushed to

the right, which makes the excess loss distribution tail much fatter since the expected total loss, the area below the excess loss function line, does not change along with the change in correlation. This can be shown using a credit VaR, which is defined as the excess loss, that probability of loss larger than this value is less than a given percentage such as 1%. The 1% credit VaR for various correlation values are given in Table 2.

In practice, it is very important to quickly obtain an accurate total excess loss distribution. There are a variety of methods that have been used for the total loss distribution. Here, we present the details on the recursive method in a one-factor Gaussian copula model and briefly summarize the conditional normal approximation approach.

Table 2 Correlation vs C-VaR

Correlation (%)	C-VaR
0	14.7
10	25.2
20	32.9
50	53.9
75	67.2%

We consider a credit portfolio consisting of n underlying credits whose notional amounts are N_i and fixed recovery rates are R_i , $i = 1, 2, \dots, n$. We consider the aggregate loss from today to time t as a fixed sum of random variables X_i :

$$L_n(t) = \sum_{i=1}^n l_i(t) = \sum_{i=1}^n (1 - R_i) \cdot N_i \cdot I_{(\tau_i < t)} \quad (10)$$

where τ_i is the survival time for the i th credit in the credit portfolio and I is the indicator function, which is 1 in the case $\tau_i \leq t$ and 0 otherwise. The distribution function of survival time c is denoted as $F_i(t) = \Pr[\tau_i \leq t]$. The specification of the survival time distribution $F_i(t)$ is usually called a *credit curve*, which can be derived from market credit default swap spreads.

From the above equation, we can calculate the total loss distribution as

$$\begin{aligned} F_{L(t)}(x) &= \Pr[L_n(t) \leq x] = \Pr\left[\sum_{i=1}^n X_i \leq x\right] \\ &= \int \Pr\left[\sum_{i=1}^n X_i \leq x | F\right] \cdot dF \end{aligned} \quad (11)$$

Conditional on the common factor F , all X_i are independent, and then we just need to calculate the convolution of n independent random variables, c . As discussed in the last section, we know that X_i are independent conditional on the common factor X_M in the one-factor model. Each X_i can take only two discrete values with constant recovery rate assumption as follows: the loss would be 0 if default does not occur or $B_i = (1 - R_i)N_i$ if default occurs.

$$f(x|F) = \begin{cases} 0, & 1 - q_i(t|F) \\ B_i, & q_i(t|F) \end{cases} \quad (12)$$

where $q_i(t|F)$ is the conditional default probability for credit i before time t .

The density of the conditional total loss distribution can be calculated recursively over the partial sum $L_j = L_{j-1} + X_j$. We then have the following recursive formula:

$$f_{L_j}(x|F) = \begin{cases} p_j \cdot f_{L_{j-1}}(x|F), & x < B_j \\ p_j \cdot f_{L_{j-1}}(x|F) + q_j \cdot f_{L_{j-1}}(x - B_j|F), & x \geq B_j \end{cases} \quad (13)$$

This has been described in [4] and also in [1]. The unconditional total loss distribution is obtained by simply integrating the conditional loss distribution over the common factor F . In the simple case of one-factor Gaussian copula model, we use a Gaussian quadrature for the integration over the one common factor. In the one-parameter case, the conditional default probability can be calculated directly as follows:

$$\begin{aligned} q_i(t|X_M) &= \Pr[\tau_i < t | X_M] \\ &= \Pr[F_i^{-1}(N(X_i)) < t | X_M] \\ &= \Pr[X_i < N^{-1}(F(t)) | X_M] \\ &= N\left(\frac{N^{-1}(q_i(t)) - \rho \cdot X_M}{\sqrt{1 - \rho}}\right) \end{aligned} \quad (14)$$

where $q_i(t)$ is the unconditional default probability of credit i before time t .

Another approach uses the conditional normal approximation. Conditional on the common factors, all credits are independent. On the basis of the law of large numbers, the total conditional loss distribution can be approximated by a normal distribution. The mean and variance of this normal distribution can be simply calculated similarly as we do in the above one-factor case. More details are given as follows.

Conditioning on the common factor X_M , we can compute the mean and variance of the total loss variable, $L|X_M$,

$$\begin{aligned} M_v &= \sum_{i=1}^n N_i \cdot (1 - R_i) \cdot q_i(t|X_M) \\ \sigma_v^2 &= \sum_{i=1}^n N_i^2 \cdot (1 - R_i)^2 \cdot q_i(t|X_M)(1 - q_i(t|X_M)) \end{aligned} \quad (15)$$

The conditional normal approach uses normal distributions to approximate the conditional loss distribution. The normal distribution has the same mean and variance as computed above. In general, other distributions, such as inverse normal or Student- t , can be used. The normal distribution is chosen because of the central limit theorem, which states that the

sum of independent distributions (but not identical distribution) approaches a normal distribution as the number of the independent distributions increases. In this case, the independent distributions are the distributions of the indicator functions of $N_i \cdot (1 - R_i) \cdot I_{[\tau_i < t]}$, which are independent when conditioned on the common factor X_M .

Given the conditional normal approach, the conditional expected loss for a tranche with attachment and detachment K_L^T and K_U^T can be easily computed in closed form as follows:

$$\begin{aligned} E(L^T(t)|X_M) = & (M_v - K_L^T) \Phi\left(\frac{M_v - K_L^T}{\sigma_v}\right) \\ & + \sigma_v \cdot \phi\left(\frac{M_v - K_L^T}{\sigma_v}\right) \\ & - (M_v - K_U^T) \Phi\left(\frac{M_v - K_U^T}{\sigma_v}\right) \\ & - \sigma_v \cdot \phi\left(\frac{M_v - K_U^T}{\sigma_v}\right) \end{aligned} \quad (16)$$

where ϕ is the one-dimensional normal density function.

With the calculated conditional expected loss, the unconditional expected loss is obtained simply by integrating over the common factor X_M .

$$E(L^T(t)) = \int_{-\infty}^{+\infty} E(L^T(t)|y) \cdot \phi(y) dy \quad (17)$$

However, by choosing normal distribution in its approximation, the approach has its limitations. First, a normal variable can have a negative value with non-zero probability. As we know, the loss in a portfolio should never be negative. However, this limitation only affects the equity tranche (the most junior tranche) and can be mitigated through the method described below. Second, as a loss is a summation of discrete loss variables, when a portfolio consists of only a few underlying names, then approximating the loss by a continuous variable (such as a normal variable) might not be a good approximation. This limitation also applies to some extreme case when the loss is dominated by only a few underlying names. In general, the conditional normal approach is a very good approximation when the number of names in a portfolio is larger than 30. Most of the CDO portfolios have the number of names larger than 30.

To mitigate the negative loss problem for an equity tranche one can use the following method, which preserves the expected loss of a CDO portfolio. An equity tranche $[0, K_U^T]$ with detachment point K_U^T has payoff as follows:

$$L^T(t) = L(t) - \max[L(t) - K_U^T, 0] \quad (18)$$

The conditional expected loss for the equity tranche is

$$\begin{aligned} E(L^T(t)|X_M) = & M_v - (M_v - K_U^T) \cdot \Phi\left(\frac{M_v - K_U^T}{\sigma_v}\right) \\ & - \sigma_v \cdot \phi\left(\frac{M_v - K_U^T}{\sigma_v}\right) \end{aligned} \quad (19)$$

This has been proven to work well for index equity tranche of size more than 2%. Alternatively, we can also use inverse Gaussian distribution to approximate for the equity tranche since inverse Gaussian distribution takes only a positive value.

Risk Measurement and Hedging

Once a model and a mapping algorithm are chosen we can price all credit portfolio trades, and produce a series of risk measures based on the model. These risk measures are then used to form a hedging strategy for the trading book. The commonly used risk measures are as follows:

- **Credit spread delta:** This is defined as the sensitivity of the mark-to-market value of a position to the instantaneous movement of the spread of a single entity, with all other parameters remaining constant. It is calculated through perturbation of the individual credit curves. Individual spread delta is reported as the change in value of the trade for a 1 basis point (1 bp) move in the indicated spread. Individual spread delta can be calculated as parallel moves in the individual curve (in which each spread on a particular curve is moved by 1 bp in a parallel fashion) or as bucketed moves in the curve. Individual spread delta is calculated trade-by-trade, and aggregated on the basis of issuer name, industry, or portfolio level. When we change the spread, we recalibrate the credit curve or instantaneous marginal default probability. Global spread delta is defined as the

change in the portfolio value change when all the underlying reference credit curves move by 1 bp. Global spread is calculated by bumping all spread curves of the underlying reference credits simultaneously in a parallel way or in buckets. Sometimes, we also study the sensitivities of our trade or book with respect to a large spread movement. Another common practice is to adjust the individual spread movement with respect to the index spreadsheet. The reason is that not all individual spread moves by the same amount when index moves. A statistical beta based on regression analysis is usually used.

- **Single-name spread gamma:** This is defined as the sensitivity of individual spread delta to a 1-bp move in a particular reference credit. As such, it represents the second-order price sensitivity with respect to a change in the spreads of the reference credit. Individual spread gamma is calculated by bumping one credit curve a time while all other credit curves remain the same for portfolio transactions. Global gamma is defined as the change in the global spread delta (which is defined as the portfolio value change when all the underlying reference credit curves move by 1 bp) of a portfolio for a 1-bp move in all reference credit spreads simultaneously. Global spread gamma is calculated by bumping all spread curves of the underlying reference credits simultaneously in a parallel way. Sometimes, we simply use a large spread movement as a measure of gamma risk by bumping the current spread by 50%. Similar to spread delta risk, we can also use bucket gamma risks which are more computationally challenging.
- **Jump-to-default risk:** We measure it by simply assuming one-name defaults right away or at a specific time in the future. We can also study the group jump-to-default risk.
- **Time-decay risk:** This measures the risk that as time passes, or maturity shortens, the value of portfolio transactions changes. For portfolio credit default swap, its survival time curve is most likely not flat, which makes the time decay an important risk factor.
- **Correlation risk:** Since we use a base correlation curve, we could measure the risk in terms of parallel change or bucket correlation change. In practice, we very often see a correlation curve twist, which reflects market changing perception

about different tranche risks. We can measure this risk by creating a sensitivity report for the whole book with respect to each base correlation point.

In practice, we tend to minimize spread and gamma risks, control jump-to-default risk, and also make correlation risk flat. We would also like to have positive carry: we receive more cash inflow than outflow. The hedging instruments we use are single name and index credit default swaps and index tranches. Very often, broker dealers tend to incur residual risks by using hedge ratios higher than the model-based amount to maintain a positive carry, but this strategy does not work well all the time, especially during turmoil, when there are unexpected defaults or jumps in spreads. We can use index tranche to hedge the base correlation risk. Sometimes, we can also use the index plus complementary tranches to hedge the correlation risk. In conclusion, for any hedging strategy, there will be a residual risk. Traders very often use their own view toward the market to selectively keep some residual risk.

In conclusion, the Gaussian copula function approach along with a base correlation method provides a simple and flexible framework to price basket credit derivatives. We further studied the framework and gained some more insights of it, especially from the conditional perspective of its correlation structure. This shows that the Gaussian copula function implies a too strong correlation structure. The reason for this is that we describe each credit using only two states: default or survival. This simple way of binary description creates too strong a conditional default property. It is also associated with the simple way of specifying the correlation structure using only one parameter or pairwise constant correlation, in practice, even though the original framework allows a completely flexible correlation matrix specification. Another possible reason is that this framework still misses certain fundamental driving factors such as volatilities of individual names in the framework.

We briefly discuss the risk measurement and risk management issues using the Gaussian copula function method. From the pricing formula, we can obtain all necessary risk measures, such as spread DV01, jump-to-default risk, and gamma risk of individual spreads or the general index. We can also obtain the sensitivities of the portfolio of portfolio transactions with respect to each point of base correlation curve.

The hedging instruments we could use include single name and index CDS, and index tranche, or even options on single name and index. However, we need to bear in mind that hedging strategies lead to residual risks.

There is an urgent need to come up with alternative models. Many alternative methods and enhancements of the current ones have been suggested (*see* **Random Factor Loading Model (for Portfolio Credit); Reduced Form Credit Risk Models; Structural Default Risk Models; Jarrow–Lando–Turnbull Model; Duffie–Singleton Model; Multiname Reduced Form Models; Intensity Gamma Model**). However, there is no market consensus on the pricing models for basic CDOs and *i*th-to-default transactions. The ultimate judge of a model should be its hedging performance. Extensive empirical studies on the performance of models from the hedging perspective need to be done comparing new models with the Gaussian copula plus the base correlation approach [2].

The current copula framework gained its popularity due to its simplicity. However, there is little economic justification for its framework. Its popularity in credit portfolio trading might match that of Black–Scholes formula for option valuation, but it lacks the theoretical background of the Black–Scholes formula. We essentially have a credit portfolio model without a solid credit portfolio theory. More theoretical studies are needed to make further advancement in this area besides various extensions and improvements, which attempt to fit the current market data.

References

- [1] Andersen, L., Sidenius, J. & Basu, S. (2003). All your hedges in one basket, *Risk* November, 67–72.
- [2] Cont, R. & Kan, R. (2008). *Dynamic Hedging of Portfolio Credit Derivatives*. Columbia University Financial Engineering Report, 2008, <http://ssrn.com/abstract=1349847>
- [3] Finger, C. (1999). Conditional approaches for Creditmetrics portfolio distributions, *CreditMetrics Monitor* April, 14–33.
- [4] Klugman, S.A., Panjer, H.H. & Willmot, G.E. (1998). *Loss Distribution: From Data to Decisions*, John Wiley & Sons, Inc.
- [5] Li, D.X. (2000). On default correlation: a copula function approach, *Journal of Fixed Income* March, 41–50.
- [6] Li, D.X. & Skarabot, J. (2004). Pricing and hedging synthetic CDOs, a chapter in the book, in *Credit Derivatives: A Definitive Guide*, J. Gregory, ed., Risk Publications.
- [7] Vasicek, O. (2004). Probability of loss on loan portfolio (KMV Working Paper, 1987), in *Derivatives Pricing: The Classic Collection*, P. Carr, ed., Risk Books.

Further Reading

- Li, D. & Liang, M. (2005). *A Mixture Copula Function Approach to CDO and CDO Squared Pricing*.

Related Articles

Base Correlation; CDO Tranches; Impact on Economic Capital; Collateralized Debt Obligations (CDO); Default Time Copulas; Duffie–Singleton Model; Intensity Gamma Model; Jarrow–Lando–Turnbull Model; Multiname Reduced Form Models; Random Factor Loading Model (for Portfolio Credit); Reduced Form Credit Risk Models; Structural Default Risk Models.

DAVID XIANGLIN LI

Base Correlation

In the Gaussian copula model (*see Gaussian Copula Model*) for pricing collateralized debt obligation (CDO) tranches, the most important input is the correlation parameter defining the degree of dependence among the defaults in the portfolio. Given single-name default probabilities implied by credit default swap spreads and recovery assumptions, the Gaussian copula model thus establishes a correspondence between this correlation parameter and the spreads of CDO tranches. Using an analogy with the notion of Black–Scholes implied volatility of an equity option, one can define a notion of implied correlation for CDO tranches. Different notions of “implied correlation” are possible, but the one which ended up being adopted by the market has been the notion of *base correlation*, explained later.

In early applications of the Gaussian copula approach for valuation of CDO tranches, market participants were using different correlations for different tranches, leading to a correlation skew, similar to the implied volatility skew observed in equity and index options. Typically, the market tends to use higher correlation for senior tranches. This way of associating different correlations for different tranches is called a *compound correlation* method. However, this compound correlation approach of pricing each tranche using a different flat correlation is not self consistent. For example, the total loss of all tranches from a whole capital structure CDO might not add up to the total loss of the underlying portfolio. Another problem is that the implied compound correlation might not be unique since the correspondence between the correlation parameter in the Gaussian copula model and the tranche spread is not one-to-one for mezzanine tranches.

A more recent market development is to treat a CDO tranche as a call or put spread on the total loss of the underlying portfolio. For example, long a 3–7% tranche protection on the standard North America credit default swap spread index (CDX) is equivalent to long a call option on the total loss with a strike price of 3% of the total notional amount of the underlying portfolio and short a call with a strike price of 7% of the total notional amount of the underlying portfolio. Then we just need to price all equity tranches with different detachment levels or simply different equity tranche size since

the attachment for the equity tranche is always equal to zero. Correspondingly, the implied correlation could be given just for equity tranche with different tranche sizes or detachment levels. This way of quoting correlation and pricing CDOs by associating a correlation curve with different equity tranche sizes is called the *base correlation method* first introduced by J. P. Morgan.

To illustrate the base correlation method, we use simple synthetic CDO structures in which the investor of a given tranche provides default protection against the total loss (L) of a portfolio over a certain range $[B, B + \Delta]$ and over a time period of $[0, t]$. The loss function for the protection seller is as follows:

$$L(B, \Delta, t) = \begin{cases} 0, & L(t) \leq B \\ L(t) - B, & B < L(t) \leq B + \Delta \\ \Delta, & L(t) > B + \Delta \end{cases} \quad (1)$$

This is similar to an insurance contract with a deductible of B and a ceiling of $B + \Delta$. We can look at this payoff function from an option perspective. Figure 1 depicts the payoff function of a 5-tranche CDO against the total loss distribution. Buying a super senior tranche is equivalent to selling a call option on the total loss of the portfolio with a strike of the attachment point of the super senior tranche. Buying an equity tranche is equivalent to selling a put option with the strike price equal to the equity size. The investor of any middle tranche is equivalent to having an option strategy of *spread* in which he/she is long a call option with strike price of $B + \Delta$ and short a call with strike price of B . Looking at a tranche from this angle gives us alternative views to the properties of CDO tranches.

From valuation perspective, we just need to value both premium leg and loss protection leg. We assume that the accumulative tranche loss up to time t is $L^T(t)$, with the attachment level B and the detachment level $B + \Delta$, and the tranche size is $\Delta = K_U - K_L$. If we ignore the interest rate impact, the payoff of the tranche $[B, B + \Delta]$ can be expressed as follows:

$$L(B, \Delta, t) = \max(L(t) - B, 0) - \max(L(t) - (B + \Delta), 0) \quad (2)$$

The expected tranche loss is the area of the excess loss distribution bounded by the subordination levels

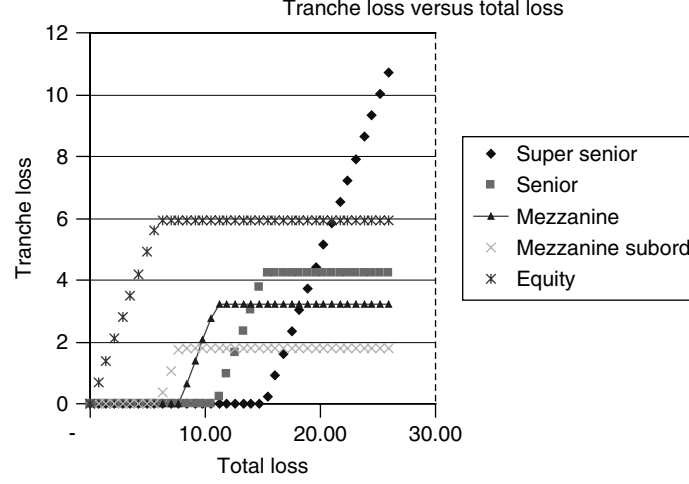


Figure 1 Tranche loss vs total loss

B and $B + \Delta$ if we ignore the interest rate. This can be taken as the difference of two equity tranche $[0, B + \Delta]$ and $[0, B]$. We can value them by simply attaching a correlation number to the tranche size or detachment point expressed by a base correlation. That is,

$$\begin{aligned} E[L(B, \Delta, t)] &= E[\max(L(t) - B, 0) | \rho(B)] \\ &\quad - E[\max(L(t) - (B - \Delta), 0) | \rho(B + \Delta)] \end{aligned} \quad (3)$$

However, the base correlation method does not guarantee absence of arbitrage. To make any credit portfolio model arbitrage free, we have the following necessary conditions around tranche loss function:

$$\begin{aligned} L(B, \Delta, t) &< L(B', \Delta, t) \quad \text{if } B < B' \quad \text{and} \\ L(B, \Delta, t_1) &< L(B, \Delta, t_2) \quad \text{if } t_1 < t_2 \end{aligned} \quad (4)$$

That is, for two tranches with everything else equal except its attachment point, the one with the higher attachment should have lower expected loss; for two tranches with everything else equal except the time interval over which we study the loss, the loss should become bigger as we increase the length of time over which we study the loss.

We have observed that for two tranches with the same tranche size, the one with more support or higher attachment level could have a break-even spread higher than the one with a lower attachment

level in the base correlation framework. In addition, the base correlation method along with the Gaussian copula model does not tell us too much about how to price nonindex portfolio tranches, neither more complicated structures such as CDO squared or CDO trades with both long and short credits as collaterals. We still need to use various loss mapping algorithms to derive a base correlation curve for the bespoke portfolio from the base correlation curve from an index tranche market.

The current market benchmark method is to use Gaussian copula function along with a base correlation method. In this framework, there is a one-to-one relationship between the tranche spread and implied base correlation curve. From a given set of index tranche spreads, we can use calibration to obtain a base correlation curve. Table 1 provides the market quote for CDX 9 on September 29, 2008 when the CDX 9 is 180 bps.

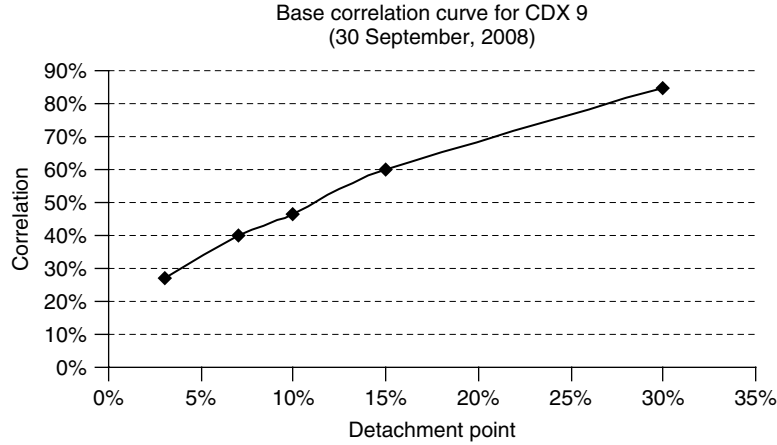
The base correlation graph is given in Figure 2.

Usually, the base correlation curve is an up sloping curve. Recently, owing to the high uncertainties in the financial market, we very often have trouble to calibrate a base correlation curve. We use a random recovery assumption to calibrate the base correlation curve. More importantly, we should incorporate the volatilities of the spreads into the framework, which has yet been formulated.

However, this base correlation approach does not tell us how to price bespoke portfolio tranches using a base correlation curve obtained from the index

Table 1 CDX 9 market quote and base correlation curve as of 29 September 2008

Swap Level 180.00					
Attachment (%)	Detachment (%)	Bid	Offer	Mid	Base correlation (%)
0.00	3.00	73 5/8	74 1/8	73 7/8	27.12
3.00	7.00	1097.00	1105.00	1101	40.16
7.00	10.00	492.00	497.00	494 1/2	46.53
10.00	15.00	199.00	204.00	201 1/2	59.77
15.00	30.00	92.00	96.00	94	84.78

**Figure 2** Base correlation curve as of September 29, 2008

tranche market. Practitioners in the market employ various mapping algorithms on the basis of the bespoke portfolio loss characteristics relative to the index portfolio loss to derive a bespoke portfolio base correlation curve from the index tranche market's base correlation curve. We provide a summary here and some general comments.

In the original copula function framework, we intend to study a portfolio problem in two stages. First, we use a credit curve to describe a single name default property. Second, we use a copula function to describe the default correlation problem. The correlation structures and correlation parameters are independent from the magnitude of single name default probabilities. That is why, we say that the correlation parameter is more rank-type correlation, which provides an order of default. In practice we often use a single correlation parameter. We have to emphasize that it is not the original framework that allows a single correlation parameter, but simply a convenient choice made by practitioners so that it is easy

to communicate with each other about the correlation.

Under these constraints, it is quite difficult to compare one credit portfolio against another credit portfolio with respect to the correlation. We simply try to link the correlation with the total loss distribution characteristics. Here are three most commonly used base correlation mapping methods.

1. Normalized strike (NS)

In this approach, we simply link the correlation on the basis of the expected loss of the portfolio. The intuitive idea is as follows: a mezzanine tranche of a bespoke credit portfolio with higher spreads should behave like an equity tranche of the index. Then we can use the following mapping approach

$$\rho_{\text{bespoke}}(K) = \rho_{\text{index}} \left[\frac{EL_{\text{index}}}{EL_{\text{bespoke}}} \cdot K \right] \quad (5)$$

This is easy to implement. However, it does not fully consider the dispersion of a credit portfolio relative

to the index. A few credits with a higher spreads could distort the scaling substantially. In addition, comparing two portfolios in terms of their expected values is a very simplistic comparison.

2. Normalized loss ratio

This approach uses the equity tranche loss, instead of the portfolio expected loss only, for the base correlation mapping. The formula is as follows:

$$\begin{aligned} & \frac{EL_{\text{bespoke}}[0, K] \text{ with } \rho_{\text{bespoke}}(K)}{EL_{\text{bespoke}}} \\ &= \frac{EL_{\text{index}}[0, K] \text{ with } \rho_{\text{index}}(K)}{EL_{\text{index}}} \end{aligned} \quad (6)$$

This is still easy to implement. It makes the use of equity tranches loss with different sizes in addition to the portfolio expected loss. From the comparison of two portfolios in terms of loss distribution perspective, it provides more detailed comparison than the normalized strike method, but still does not capture the loss variations within a tranche. Sometimes, it is impossible to find a solution for this method.

3. Percentile-to-percentile loss distribution mapping

$$\begin{aligned} & \Pr(L_{\text{bespoke}} > K_{\text{bespoke}}; \rho_{\text{index}}(K)) \\ &= \Pr(L_{\text{index}} > K; \rho_{\text{index}}(K)) \end{aligned} \quad (7)$$

This approach compares the loss distribution of bespoke portfolio against that of an index portfolio directly by the percentile-to-percentile comparison. We simply find a corresponding strike for the bespoke portfolio whose excess loss probability at the strike

(to be solved) is equal to that of the index portfolio while both excess probabilities are calculated at the same index base correlation at the index strike level. From loss distribution comparison perspective, it provides the most comprehensive comparison. Both average spread and the dispersion of spread should be captured by this method.

Although all these mapping algorithms based on loss distribution are intuitive, none of them are fully theoretically justifiable. Without fully using individual name information such as asset return volatilities and correlations, it is difficult to compare one portfolio against another portfolio as we do in equity portfolio.

Further Reading

- Li, D.X. (2000). On default correlation: a copula function approach, *Journal of Fixed Income* **9**, 41–50.
 McGinty, L., Eric B., Rishad, A. & Martin, W. (2004). Introducing base correlation, *Credit Derivative Strategy*, J.P. Morgan.

Related Articles

Collateralized Debt Obligations (CDO); Default Time Copulas; Modeling Correlation of Structured Instruments in a Portfolio Setting; Gaussian Copula Model; Random Factor Loading Model (for Portfolio Credit).

DAVID XIANGLIN LI

Random Factor Loading Model (for Portfolio Credit)

Consider a portfolio of N risky assets, all assumed (for simplicity) to generate a \$1 loss at the time of default. Let τ_i denote the random default time for asset i , such that the total portfolio loss $L(T)$ on the horizon $[0, T]$ is

$$L(T) = \sum_{i=1}^N 1_{\tau_i \leq T} \quad (1)$$

From credit default swap (CDS) or bond markets, we can normally extract risk-neutral survival probabilities

$$Q_i(T) = \Pr(\tau_i > T), \quad i = 1, \dots, N \quad (2)$$

for all T ; this information locks risk-neutral expected portfolio losses at

$$E(L(T)) = \sum_{i=1}^N E(1_{\tau_i \leq T}) = \sum_{i=1}^N (1 - Q_i(T)) \quad (3)$$

To be able to construct the entire distribution of $L(T)$ —and not just its first moment—we need additional information about the default codependencies among the N assets.

The default codependence model that we consider here is in the class of *factor models*, in the sense that codependence is induced solely by a scalar^a random variable Z —the so-called *systematic factor*—that affects all assets through a factor “loading” function. Conditional on Z , all N default times τ_i are assumed to be independent of each other.

In practice, specification of factor loading is done using *conditional survival time distributions* $q_i : \mathbb{R}_+ \times \mathbb{R} \rightarrow [0, 1]$, defined by

$$q_i(t, z) \equiv \Pr(\tau_i > t | Z = z) \\ i = 1, \dots, N, \quad t \geq 0 \quad (4)$$

Complete prescription of a factor model requires specification of (i) all N functions q_i and (ii) the probability distribution of the systematic factor Z . As should be obvious, the q_i ’s cannot be prescribed

arbitrarily, as they are subject to strong consistency and regularity conditions. For instance, we know from basic probability

$$\Pr(\tau_i > T) = \int q_i(T, z) \Pr(Z \in dz) = Q_i(T) \quad (5)$$

which, for any given distribution of Z , provides an important constraint on q_i . Other regularity conditions—including those associated with the fact that L must be never decreasing in T —are reviewed in [1].

We emphasize the importance of the assumption of conditional independence, which allows for the application of efficient numerical methods to construct the (discrete) distribution of $L(T)$ in equation (1). Andersen *et al.* [4] give one such algorithm and discuss in detail its application in price and sensitivity computations for collateralized debt obligations (CDOs).

Random Factor Loading Models

A standard recipe for specifying the functions q_i ’s in a financially meaningful way is to assume that $\Pr(\tau_i > T) = \Pr(X_i > H_i(T))$ for a deterministic default barrier $H_i(T)$ and a *default driver* X_i of the form

$$X_i = \beta_i Z + e_i, \quad i = 1, \dots, N \quad (6)$$

where β_i is a firm-specific constant, Z is a one-dimensional systematic factor, and e_i is a residual variable idiosyncratic to firm i and independent of Z and e_j , $j \neq i$. As Z is often loosely considered a proxy for the state of the “market”, equation (6) has some qualitative similarity with the CAPM setup, with X_i loosely representing the asset returns on firm i . The Gaussian copula model falls in the class of equation (6) as do many Levy-type copula models.

The RFL (random factor loading) class starts from equation (6), but alters the dependence of X_i on Z from strictly linear to a generic functional relationship. Specifically, one writes

$$X_i = A_i(Z) + e_i, \quad i = 1, \dots, N \quad (7)$$

where A_i is a possibly firm-specific deterministic function. For reasons of tractability, it is most common to assume that Z and e_i are Gaussian, and to (arbitrarily) normalize such that $E(X_i) = 0$. In this case, one has

$$X_i = A_i(Z) + \epsilon_i + m_i, \quad i = 1, \dots, N \quad (8)$$

where Z and all residuals ϵ_i are independent standard Gaussian variables (i.e., distributed as $\mathcal{N}(0, 1)$), and the constant m_i is set to $m_i = -E(A_i(Z))$. Going forward, equation (8) shall be our working definition of a one-factor RFL model.

By moving away from strictly linear “loading” on the systematic variable, RFL models can incorporate a number of empirical observations about default codependence dynamics. Most importantly, we have the ability to increase the loading for low values of Z (= a “bad” market outcome) as a way of modeling the well-established fact that equity price correlations tend to increase in a market downturn. This, in turn, tends to fatten the upper tail of the distribution of $L(T)$, an effect that is consistent with the market for synthetic CDOs.

Some Analytical Results

For simplicity, let us now drop^b the subscript i on A_i and m_i . For completely arbitrary specifications of $A(z)$, there is obviously no hope that the distribution of X_i in equation (8) has a closed-form representation. If $A(z)$ is taken to be piecewise linear, however, such a result exists, as shown in [3]. To list it, let us define thresholds $\theta_0 < \theta_1 < \dots < \theta_{K-1}$ and then write

$$A(z) = (\alpha_0 z + \beta_0) 1_{z \leq \theta_0} + \sum_{k=1}^{K-1} (\alpha_k z + \beta_k) 1_{z \in (\theta_{k-1}, \theta_k]} + (\alpha_K z + \beta_K) 1_{z > \theta_{K-1}} \quad (9)$$

where the slopes $\{\alpha_k\}_{k=0}^K$ and intercepts $\{\beta_k\}_{k=0}^K$ are given constants. Let $\Phi(x)$ be the Gaussian conditional default functions (CDFs), and let $\Phi_2(x, y; \rho)$ be the bivariate Gaussian CDF at the correlation level ρ . Define, for $k = 1, \dots, K-1$,

$$\begin{aligned} \psi(k, x) = & \Phi_2 \left(\frac{x - \beta_k - m}{\sqrt{1 + \alpha_k^2}}, \theta_k; \frac{\alpha_k}{\sqrt{1 + \alpha_k^2}} \right) \\ & - \Phi_2 \left(\frac{x - \beta_k - m}{\sqrt{1 + \alpha_k^2}}, \theta_{k-1}; \frac{\alpha_k}{\sqrt{1 + \alpha_k^2}} \right) \end{aligned} \quad (10)$$

and

$$\psi(0, x) = \Phi_2 \left(\frac{x - m - \beta_0}{\sqrt{1 + \alpha_0^2}}, \theta_0; \frac{\alpha_0}{\sqrt{1 + \alpha_0^2}} \right) \quad (11)$$

$$\psi(K, x) = \Phi_2 \left(\frac{x - m - \beta_K}{\sqrt{1 + \alpha_K^2}}, -\theta_{K-1}; \frac{-\alpha_K}{\sqrt{1 + \alpha_K^2}} \right) \quad (12)$$

Then

$$\Pr(X_i \leq x) = \sum_{k=0}^K \psi(k, x) \quad (13)$$

We can use equation (13) to ensure that the model is in calibration with market-observed default probabilities, by insisting that the default barrier function $H_i(T)$ is set (by numerical root search) such that

$$\sum_{k=0}^K \psi(k, H_i(T)) = 1 - Q_i(T), \quad i = 1, \dots, N \quad (14)$$

We also notice that (with φ being the Gaussian density)

$$\begin{aligned} E(X_i) = & m - \alpha_0 \varphi(\theta_0) \theta_0 + \beta_0 \Phi(\theta_0) \\ & + \sum_{k=1}^{K-1} \alpha_k (\varphi(\theta_{k-1}) - \varphi(\theta_k)) \\ & + \sum_{k=1}^{K-1} \beta_k (\Phi(\theta_k) - \Phi(\theta_{k-1})) \\ & + \alpha_K \varphi(\theta_{K-1}) + \beta_K \Phi(-\theta_{K-1}) \end{aligned} \quad (15)$$

which can be used to set m such that $E(X_i) = 0$.

CDO Calibration

The free parameters of the RFL model are those involved in setting the function $A(z)$. Assuming that A is piecewise linear on K different intervals, we evidently have a total of $3K + 2$ parameters in

the model: K interval break-points $\theta_0, \theta_1, \dots, \theta_{K-1}$; $K + 1$ slopes $\alpha_0, \dots, \alpha_K$; and $K + 1$ intercepts $\beta_0, \dots, \beta_K$. In general, this number of parameters is too high, so to avoid overfitting one normally locks some of these parameters manually and calibrates the rest against observed CDO prices. A few common parameter strategies are listed below.

Classic RFL

In this approach, which was developed in [3], we set all $K + 1$ intercepts (β_k) to zero, leaving $2K + 1$ free parameters for the calibration. In typical applications, sufficient calibration accuracy is often reached with $K = 2$ break-points, for a total of five free calibration parameters. We note that when intercepts are forced to zero, the function A is of the form

$$A(z) = a(z) \cdot z \quad (16)$$

where a is piecewise flat. Comparison with equation (6) shows that, loosely, the function $\sqrt{a(z)}$ can be interpreted as a “random correlation” function. Consistent with earlier discussion, the calibrated $a(z)$ will always be decreasing in z , that is, when the economy variable Z is low (= bad economy) correlations increase, and *vice versa*.

Discrete RFL

In this style, we set all $K + 1$ slopes to zero, yielding

$$A(z) = b(z) \quad (17)$$

where b is piecewise flat at levels $b_0, b_1, \dots, b_K$. Evidently then, the distribution of $A(Z)$ is here simply a discrete distribution^c taking on $K + 1$ different values with $K + 1$ different probabilities. As the RFL model is normalized to work with the term $A(Z) - E(A(Z))$, we can add an arbitrary constant to the function $b(z)$ without altering the model; equivalently, we are free to lock one of the b_k values to a fixed constant (e.g., zero) without losing any generality. As a consequence, the effective number of parameters here is $2K$.

Fixed-slope RFL

Our setup here is similar to that in discrete RFL, but now we allow for a nonzero constant slope α to be used for all line segments: $\alpha_k = \alpha, k = 0, \dots, K$. We include this parameter in the set of free parameters to be optimized on, so now the problem dimension is $2K + 1$.

Comments and Extensions

Unlike a number of other factor models, the RFL model extends the basic Gaussian copula model by altering the CDF q_i , rather than the density of the systematic factor Z . On the other hand, we can rewrite the RFL model as

$$X_i = Y + \epsilon_i, \quad Y = A(Z) + m \quad (18)$$

So, if we elect to treat the variable Y , rather than Z , as our systematic factor, the RFL model can, in fact, also be interpreted as extending the Gaussian copula model through a change of systematic factor density. Indeed, by a suitable choice of A , *all* factor models with Gaussian residuals can be cast as an RFL model of the type in equation (8). Andersen and Piterbarg [2] discuss this in more detail, using the models in [5, 6, 8] as examples.

In [1], the RFL model is extended to allow for Poisson- or mixture-style jumps in both residuals and in the systematic factor; the model in [7] is a special case of such a jump-extended RFL model. Andersen [1] also discusses methods to introduce a dynamic element into RFL and other factor models, by letting the density of Z depend on time.

End Notes

^aExtensions to vector-valued Z is straightforward; see, for example, [3].

^bTo keep free parameters at a manageable level, it is, in fact, common to use a single A for an entire portfolio. Firm-specific A functions may, however, be of use when mixing portfolios or as part of a bespoke mapping rule.

^cSo rather than optimizing on the θ_k parameters, we can work directly with the discrete probabilities $\Phi(\theta_k) - \Phi(\theta_{k-1})$.

References

- [1] Andersen, L. (2006/2007). Portfolio losses in factor models: term structures and intertemporal loss dependence, *Journal of Credit Risk* **2**(4), 3–31.
- [2] Andersen, L. & Piterbarg, V.L. (2008). *The Definitive Guide to CDOs – Market, Application, Valuation, and Hedging*, Risk Books.
- [3] Andersen, L. & Sidenius, J. (2004/2005). Extensions of the Gaussian Copula: random recovery and random factor loadings, *Journal of Credit Risk* **1**(1), 29–70.
- [4] Andersen, L., Sidenius, J. & Basu, S. (2003). All your hedges in one basket, *Risk* **16**, 67–72.

- [5] Guegan, D. & Houdain, J. (2005). *Collateralized Debt Obligations Pricing and Factor Models: A New Methodology Using Normal Inverse Gaussian Distributions*. Working Paper.
- [6] Inglis, S. & Lipton, A. (2007). *Factor Models for Credit Correlation*. Working Paper, Merrill Lynch.
- [7] Willemann, S. (2005). *Fitting the CDO Correlation Skew: A Tractable Structural Jump Model*. Working Paper, Aarhus Business School.
- [8] Xu, G. (2006). *Extending Gaussian Copula with Jumps to match Correlation Smile*. Working Paper, Wachovia Securities, defaultrisk.com.

Related Articles

Base Correlation; Collateralized Debt Obligations (CDO); Copulas: Estimation; Credit Portfolio Simulation; Default Barrier Models; Gaussian Copula Model; Local Correlation Model; Multi-name Reduced Form Models.

LEIF B.G. ANDERSEN

Local Correlation Model

The local correlation model is a credit portfolio loss model that generalizes the Gaussian copula model (see **Gaussian Copula Model**) and allows to account for the base correlation smile (see **Base Correlation**).

The local correlation model is essentially similar to an exotic copula model; its originality lies mainly in its economic interpretation. The model consists of a slight modification in the analytical specification of the Gaussian copula (see **Gaussian Copula Model**):

$$A_i = \sqrt{\rho(X)}X + \sqrt{1 - \rho(X)}\varepsilon_i \quad (1)$$

Each reference entity in the portfolio is represented by its asset value A_i , driven by a specific factor ε_i , and an economic factor X , common to all names in the economy. In the Gaussian framework, the correlation parameter ρ is constant and therefore A_i is Gaussian. On the contrary, in the local correlation model, $\rho(X)$ is a function of the economy factor, making assets non-Gaussian. As a result, $\rho(X)$ is not a measure of the actual asset correlation in the traditional sense. In the local correlation model, default correlation does not depend on the average spread of the portfolio, its dispersion, rating, or industrial sectors, but rather on the global state of the economy. In other words, the local correlation function is universal and can be used regardless of the composition of the reference portfolio. We can, therefore, calibrate our local correlation function to standard Collateralized Debt Obligation (CDO) tranches and use it to price bespoke CDOs. The model's applications are discussed later.

Large Pool Framework

Conditional on a state of the economy, the local correlation model behaves just like any other Gaussian copula model. The conditional cumulative loss is given as

$$\begin{aligned} L(K|X) \\ = P \left[(1 - RR) \times \frac{1}{n} \sum_{i=1}^n 1_{\{A_i \leq G_i^{-1}(p_i)\}} \leq K | X \right] \end{aligned} \quad (2)$$

In equation (2), RR is a fixed recovery rate assumption, n is the number of obligors in the portfolio, A_i is the i th asset value as defined in the previous section, p_i is the probability of default for the i th obligor, and G_i is A_i 's cumulative distribution function. P is the risk-neutral probability. In an infinitely diversified portfolio, asset distributions and default probabilities do not depend on the obligor, and equation (2) becomes

$$\begin{aligned} L(K|X) &= P \left[1_{\{A \leq G^{-1}(p)\}} \leq \frac{K}{1 - RR} | X = x \right] \\ &= 1 \left\{ N \left(\frac{G^{-1}(p) + x\sqrt{\rho(x)}}{\sqrt{1 - \rho(x)}} \right) \leq \frac{K}{1 - RR} \right\} \end{aligned} \quad (3)$$

Finally, equation (3) yields the expression for unconditional cumulative loss:

$$L(K) = P \left[N \left(\frac{G^{-1}(p) + X\sqrt{\rho(X)}}{\sqrt{1 - \rho(X)}} \right) \leq \frac{K}{1 - RR} \right] \quad (4)$$

Assuming that cumulative loss distribution L and asset distribution G are known, we can use equation (4) to obtain the local correlation function.

Implied Loss Distribution

Let us first derive an expression for cumulative loss distribution. The only information we have about a portfolio's cumulative loss comes from CDO market quotes. For example, five single-tranche CDOs are quoted based on the iTraxx Main portfolio—the European credit benchmark—and the CDX IG portfolio—the US credit benchmark. These prices form the “base correlation skew”—five constant correlations obtained using the Gaussian copula framework (see **Gaussian Copula Model; Base Correlation**). We write the naïve cumulative loss obtained using constant base correlations as $L(K, \rho_K^{\text{Base}})$. The naïve cumulative loss coincides with the actual cumulative loss $L(K)$ on the five market points. $L(K)$ can then be continuously interpolated using the following formula:

$$\begin{aligned} L(K) &= L(K, \rho_K^{\text{Base}}) + \frac{\partial L}{\partial K}(K, \rho_K^{\text{Base}}) \\ &\quad \times \int_0^K \frac{\partial L}{\partial \rho}(k, \rho_K^{\text{Base}}) dk \end{aligned} \quad (5)$$

Results and Interpretation

One last assumption is made before we derive the final result: we need to assume that the asset distribution G is a Gaussian distribution, that is, $G \equiv N$. We drop this assumption later on. Equations (4) and (5) can finally be combined to obtain an analytical expression for local correlation $\rho(X)$.

Let us now discuss the economic interpretation of equation (4). We define the idiosyncratic threshold as

$$\tilde{\varepsilon}(X) = \frac{G^{-1}(p) + X\sqrt{\rho(X)}}{\sqrt{1 - \rho(X)}} \quad (6)$$

Equation (4) can be rewritten according to the following equation:

$$L(K) = N(x_K), \quad x_K = \tilde{\varepsilon}^{-1} \left(N^{-1} \left(\frac{K}{1 - RR} \right) \right) \quad (7)$$

This means that for each level of loss K in the portfolio, there is an equivalent^a state of the economy x_K . Thus, for every state of the economy, there is a single corresponding loss level for any diversified portfolio. We can, therefore, interpret the local correlation $\rho(X)$ in a given state of the economy as the equivalent constant correlation to be used for a tiny tranche of size dK , centered at strike $K = L^{-1} \cdot N(X)$.

The results so far are summarized as follows. The large pool assumption allows us to relate the local correlation function to the cumulative loss distribution. Furthermore, assuming that assets are normally distributed, we can generate a mapping of each state of the economy into a loss level for any given portfolio. The local correlation can, therefore, be interpreted as the correlation of a tiny tranche centered at the corresponding loss level in a given state of the economy.

Relaxing the Gaussian Assumption

We made two crucial assumptions in the previous sections: we assumed that the portfolio could be considered infinitely diversified—the large pool assumption—and we also assumed that despite the local correlation specification (1), assets could be considered to be Gaussian.

If we drop the Gaussian assumption, the law of assets becomes

$$G_i^\rho(z) = P[A_i \leq z] \\ = \int N \left(\frac{z + x\sqrt{\rho(x)}}{\sqrt{1 - \rho(x)}} \right) \varphi(x) dx \quad (8)$$

Equations (2) and (8) can be used simultaneously to solve both the local correlation function $\rho(x)$ and asset distribution G at the same time, through a fixed-point algorithm. We initialize the algorithm by setting $G \equiv N$ and solve equation (2) to get the corresponding $\rho(x)$ function. The result is substituted into equation (8), which yields a new version of G . We iterate this process until the local correlation function is stable.

Relaxing the Large Pool Assumption

We now relax the large pool assumption. Considering equation (2), individual default probabilities are now used in their general form:

$$g_{i|X} = N \left(\frac{G^{-1}(p_i) + X\sqrt{\rho(X)}}{\sqrt{1 - \rho(X)}} \right) \quad (9)$$

Assuming that the function $\rho(X)$ is known, we can compute the loss distribution $L(K)$ via equation (2), using Andersen's combinatorial algorithm as presented in [2]. We, therefore, need to provide the model with a functional form for local correlation such as a parametric representation. In [1] (see **Random Factor Loading Model (for Portfolio Credit)**), Andersen *et al.* use a piecewise constant correlation function in their random factor loading model. Continuous functions such as piecewise linear functions or cubic splines can also be used. Such a parameterization suggests that the model needs to be calibrated to market prices through high-dimension optimization. Such an optimization is beyond the scope of this article.

Application to Exotic CDO Valuation

The problem of mapping the correlation of bespoke portfolios against standard ones has been a hot topic ever since standard CDO tranches started to trade. The question is, “How do we obtain the Gaussian correlation assumption to use with a nonstandard

portfolio and nonstandard subordination from standard CDO quotes?” Amongst other authors, Turc and Very [4] have described several ways of choosing the right equivalent correlation for nonstandard CDO pricing [3]. The probability-matching approach turns out to be the most consistent technique. It suggests using the index correlation corresponding to an equivalent strike in probabilistic terms. In other words, the index strike and bespoke strike are equivalent if they have the same probability of being reached. In the local correlation model, remember that the function $\rho(X)$ is considered as a universal constant, independent of the portfolio. Thus, equation (7) yields

$$L_{\text{index}}(K_{\text{index}}) = N(X) = L_{\text{bespoke}}(K_{\text{bespoke}}) \quad (10)$$

Equation (10) shows that the local correlation model is equivalent to the probability-matching approach. It is, therefore, consistent with one of the most popular market practices for bespoke CDO pricing.

Acknowledgments

We would like to acknowledge the contribution of Philippe Very, of Natixis, for the development of the local correlation model.

End Notes

^aEquivalence in terms of probability.

References

- [1] Andersen, L. & Sidenius, J. (2004). *Extensions to the Gaussian Copula: Random Recovery and Random Factor Loadings*.
- [2] Andersen, L., Sidenius, J. & Basu, S. (2003). All your hedges in one basket, *Risk* November, 67–72.
- [3] Jeffery, C. (2006). Credit model meltdown, *Risk Magazine* 19(11), 21–25.
- [4] Turc, J. & Very, P. (2008). Pricing CDOs with a smile: the local correlation model, in *Frontiers in Quantitative Finance: Volatility and Credit Risk Modeling*, R. Cont, ed., Wiley, Chapter 9.

Further Reading

Burtschell, X., Gregory, J. & Laurent, J.-P. (2005, 2008). *A Comparative Analysis of CDO Pricing Models*.

Related Articles

Base Correlation; CDO Tranches: Impact on Economic Capital; Collateralized Debt Obligations (CDO); Gaussian Copula Model; Modeling Correlation of Structured Instruments in a Portfolio Setting; Random Factor Loading Model (for Portfolio Credit).

JULIEN TURC & BENJAMIN HERZOG

Intensity Gamma Model

The intensity gamma model is a model for pricing portfolio credit derivatives developed in [2]. Its innovation was to use stochastic time change (*see Time Change*) to achieve clustering of defaults and therefore correlation between them. This was in contrast to the popular models available at the time, which relied on copulas, such as the Gaussian copula model (*see Gaussian Copula Model*). The model is designed to infer the joint distribution of the default times of some set of names from the marginal distributions. Some additional market information about the level of correlation is required for calibration; this is typically obtained from liquid correlation products.

Once a joint distribution for the default times has been obtained (in some suitable pricing measure), products whose cash flows depend on those defaults can be priced. The classic examples include collateralized debt obligations (CDOs) (*see Collateralized Debt Obligations (CDO)*) and n th-to-default baskets (*see Basket Default Swaps*). However, it is no more difficult to handle other products in which the cash flows depend on the exact defaults occurring in quite arbitrary ways, such as CDO-squared (*see CDO Square*). This approach is more robust and more general than techniques founded on mapping methodologies such as base correlation (*see Base Correlation*). The intensity gamma model, like other reduced-form models (*see Reduced Form Credit Risk Models*), is an arbitrage-free model that matches the market and allows pricing of derivatives on arbitrary functions of defaults, rather than attempting to interpolate prices of nontraded tranches.

On the other hand, the intensity gamma model is not designed to capture the volatility of credit spreads or stochastic correlation. The credit spreads of individual names are deterministic functions of time, within the model, up until their default. This is a significant limitation and certainly prevents the use of the model for handling certain products such as options on CDOs.

Model Definition

The core idea of the model is that defaults are driven by a business time process, X_t , which is an increasing

adapted process with $X_0 = 0$. This process intuitively represents the amount of bad news that has arrived by time t , and as bad news arrives, each name has a chance of defaulting in response to the bad news, with names behaving *conditionally* independently, that is, for a given amount of bad news, each name then defaults independently with a certain probability. The fact that all names are driven by the same business time process introduces correlation into the defaults.

To describe more precisely, we start by considering a homogeneous default process in which individual default rates are constant across the time period of interest. The business time process (X_t) will be a subordinator, that is, an increasing Lévy process (*see Lévy Processes*), and in practice it is normally taken to be a constant drift at rate a plus either a gamma process or the sum of two independent gamma processes [1]. A gamma process, (Γ_t) , is described by two parameters, which we denote by γ and λ , and then Γ_t has a gamma($\gamma t, \lambda$)-distribution, that is to say a density function f given as

$$f(x) = \frac{\lambda^{\gamma t}}{\Gamma(\gamma t)} x^{\gamma t-1} e^{-\lambda x} \quad (1)$$

Suppose that we have some set of reference names $\{A_i : i \in I\}$. In the most basic form of the intensity gamma model, each name A_i defaults at some constant rate c_i with respect to the passage of business time as given by the process (X_t) . That is to say, conditional on the path $(X_t)_{t \leq T}$, the name A_i will have survived (i.e., not defaulted) to time T with probability

$$e^{-c_i X_T} \quad (2)$$

with the default events being independent after conditioning on (X_t) .

Calibration of the Model to Individual Default Probabilities

We assume that single-name survival probabilities have been inferred from CDS spreads or bond prices using some simple recovery rate assumption, as is common practice in CDO pricing. The model survival

probability is given by integrating the conditional survival probability (2) against the law of X_T , which amounts to taking a Laplace transform of this law. Suppose for now that this law (and its parameters) has somehow been determined. In the case when it is a single gamma distribution with parameters $(\gamma T, \lambda)$, the required Laplace transform is simple and well known to be

$$\frac{1}{(1 + c_i/\lambda)^{\gamma T}} \quad (3)$$

When X_T is a sum of independent gamma processes, then the survival probability is the product of the corresponding probabilities for each gamma process (3); similarly, a drift of rate a introduces a further multiplicative term into the survival probability of $e^{-ac_i T}$. Thus, there is a straightforward analytic expression for the survival probability of an individual name, and we can use this to quickly solve for c_i to fit the market-implied survival probability for the time horizon of interest.

More generally, we may wish to calibrate to market-implied survival probabilities for a series of time horizons $0 = t_0 < t_1 < t_2 < \dots < t_n = T$. This can be done by generalizing the definition to allow a default rate (with respect to business time) $c_i(t)$, which is a function of calendar time t , which is constant over each subinterval of calendar time of the form $[t_j, t_{j+1}]$.

Pricing Correlation Products

In the following section, we discuss the issue of how to choose the parameters of our business time process X_t and describe now how we price a correlation product, such as a CDO, once these parameters have been specified. From the preceding section, we rapidly calibrate the default rates of all relevant individual names to the single-name credit market at some fixed sequence of time horizons, $0 = t_0 < t_1 < \dots < t_n = T$. The correlation product is then priced by Monte Carlo. For each Monte Carlo path, we must first draw the random business time process, $(X_t)_{t \leq T}$, where T is the last relevant possible default date for the product; this can be done with a finite-dimensional random draw

and results in an (arbitrarily precise) approximation to the business time path by one which contains a finite number of jumps imposed on a path with a constant drift: details can be found in [2] adapted from a method in [1]. Having determined the random path (X_t) , one can then generate the defaults by drawing an independent uniformly distributed random variable U_i for each name A_i and saying that A_i has defaulted by time S if and only if

$$\exp\left(-\int_0^S c_i(t) dX_t\right) < U_i \quad (4)$$

Calibrating to the Correlation Market

First, note that doubling the business time process (corresponding to doubling the λ parameters and the drift parameter a) would be canceled out exactly by halving all the default intensities c_i ; thus, the business time process effectively has a redundant parameter and we may assume, without loss of generality, that $a = 1$. Therefore, if there are two independent gamma processes, there are four free parameters controlling the process and hence the default correlation.

With four free parameters, it is possible to obtain a variety of shapes of the correlation graph, and one can therefore calibrate to multiple tranches simultaneously rather than having to use different correlations for each tranche. This ability to match the correlation smile was a major motivation for the model's introduction. For any given choice of business process parameters, calibrating to the single-name market is instant and pricing CDO tranches by Monte Carlo is fairly quick. A multidimensional route finder can then be used to calibrate the business process parameters to as many independent tranche prices as there are parameters. In practice, one will choose the quoted tranche prices from some major index such as iTraxx or CDX. The index will typically be chosen to have similar properties in terms of maturity, region, diversity, and credit quality as the bespoke correlation product that we are ultimately aiming to price.

For further details on the model, we refer the reader to the original paper [2] or to the recent book [3].

References

- [1] Cont, R. & Tankov, P. (2003). *Financial Modelling with Jump Processes*, Chapman and Hall.
- [2] Joshi, M.S. & Stacey, A.M. (2006). Intensity gamma: a new approach to pricing portfolio credit derivatives, *Risk Magazine* July 78–83.
- [3] O’Kane, D. (2008). *Modelling Single-name and Multi-name Credit Derivatives*, Wiley.

Related Articles

Basket Default Swaps; Collateralized Debt Obligations (CDO); Lévy Processes; Intensity-based Credit Risk Models; Reduced Form Credit Risk Models.

MARK JOSHI & ALAN STACEY

Modeling Correlation of Structured Instruments in a Portfolio Setting

Credit events are correlated. Corporate or retail loan portfolios can exhibit wide swings in losses as economic factors common to the underlying entities drive defaults or deterioration in credit quality. Though challenging, the modeling and data issues related to single-name credit instruments are well understood [9, 10]. However, the correlation of structured instruments remains much more complicated and less understood. Traditional approaches to modeling economic capital, credit-VaR (Value-at-Risk), or structured instruments whose underlying collateral is composed of structured instruments treat structured instruments as a single-name credit instrument (i.e., a loan-equivalent).^a Though tractable, the loan-equivalent approach requires appropriate parameterization to achieve a reasonable description of the cross-correlation between the structured instrument and the rest of the portfolio. We address this challenge by calibrating the loan-equivalent correlation parameters to the dynamics observed in a granular model of the structured instrument.

In the granular model, the underlying reference entities associated with the collateral pool are used to simulate collateral losses, which are translated to structured instrument loss using its subordination level. For ease of exposition, we assume pass-through waterfalls throughout the article; the waterfall structure and subordination level are completely determined by attachment and detachment points. The structured instrument is said to be in distress if it incurs a loss. Simulated losses are used to calculate a probability of distress and loss given distress for each structured instrument and a joint probability of distress (JPD) for the structured instrument and other instruments in the portfolio. The JPD and the individual probabilities of distress are then used to back out an implied “asset return” correlation associated with the structures’ loan-equivalent reference entities (henceforth, *the loan-equivalent correlation*). By taking the probability of distress as the loan-equivalent probability of default (PD), the loss given distress as the loan-equivalent loss given

default, and the deal maturity as the loan-equivalent maturity the parameterization of the loan-equivalent is complete.

In addition to the benefits of using loan-equivalents in a portfolio setting, the simplicity associated with loan-equivalent correlations serves as a useful summary statistic for understanding portfolio-referent risk characteristics. For example, when the collateral pool is parameterized under a wide range of values, the resulting correlation of the structure instrument with the rest of the portfolio is far higher than that of the underlying collateral pool or for other classes of single-name instruments, such as corporate exposures. This is because the idiosyncratic shocks in a collateral pool offset one another, and the systematic portion is left to “run the show”. Moreover, loan-equivalent correlations between two subordinated tranches can be substantially lower than their senior counterparts. The higher correlation exhibited between two credit instruments in the tail region of the loss distribution (the loss region of the collateral pool where a senior tranche enters distress) drives the difference. This is because a senior tranche distress is most likely driven by systematic shock, which makes it more likely that it will be accompanied by a distress in the other senior tranche.

Correlation among CDOs can also stem from the overlap in reference entities associated with their collateral pools. Although it is intuitive that loan-equivalent correlations increase with the degree of overlap, the effect is stronger for more junior tranches. This finding follows from the junior tranches’ susceptibility to idiosyncratic noise. With a higher degree of overlap, the tranches share more of the idiosyncratic shocks.

Recent literature on structured instruments’ correlation primarily addresses correlations between collateral pool reference entities and their effect on tranche pricing. For example, Agca and Islam [1] examines the impact of correlation increase among underlying assets on the value of a CDO equity tranche, and shows that CDO equity can be short on correlation. The normal copula model, studied in detail in [4], has become the industry standard for pricing structured instruments and often results in a market price-implied correlation smile. Several papers address this phenomenon; Moosbrucker [5] explains it using variance-gamma distributions. This article, however, focuses on the

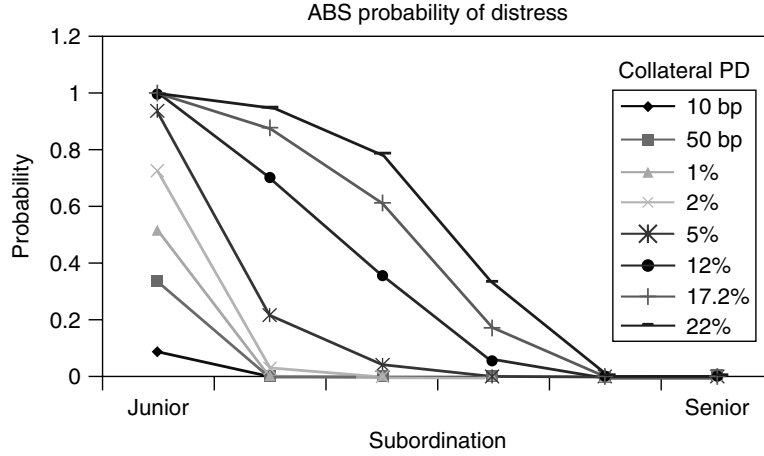


Figure 1 ABS probability of distress changing with subordination for collateral parameters: $R^2 = 10\%$, $LGD = 70\%$, and collateral PD ranging from 10 basis points to 22%

correlation structure between a structured instrument and other instruments in the portfolio. This structured instrument correlation is a function of the correlation of the underlying collateral instruments among themselves and with other instruments in the portfolio, as well as of the deal's waterfall structure and the instrument's subordination level. Coval *et al.* [2] show that, similar to economic catastrophe bonds, structured instruments are highly correlated with the market, but offer far less compensation than economic catastrophe bonds. The authors suggest in [3] that this lower compensation occurs because rating agencies base their structured instrument ratings on distress probability and expected loss, ignoring the high correlation with the market.

Modeling the Correlation of Two ABSs

Consider an asset-backed security (ABS) with a homogeneous collateral pool (e.g., similar risk California auto-loans). In a homogeneous pool, we can model the systematic portion of reference entity i 's asset return process (r_i) using a single factor (Z) with a common factor loading (R):

$$r_i = R \cdot Z + \sqrt{1 - R^2} \cdot \varepsilon_i \quad (1)$$

Henceforth, we use the square form of the factor loading, R^2 , because of its interpretation as the proportion of asset return variation explained by the common factor Z . The single factor Z and the idiosyncratic portion ε_i are standard Brownian motion processes and, therefore, so is the asset return. A default occurs when the asset return process drops below a default threshold (DT), which happens with probability PD. These PDs and other parameters for the underlying instruments in the collateral pool are publicly available.^b

In analyzing ABS characteristics, we made simplifying assumptions. We assumed a single period, and the analysis abstracted from subtleties such as reinvestment, collateralization, or wrapping. Collateral pools were composed of instruments with Loss Given Default (LGD) = 70% and reference entities with $R^2 = 10\%$. Each deal consisted of seven ABSs at different subordination levels. Deal maturity was set at 1 year. All deals were simulated 100 000 times in Moody's KMV's RiskFrontierTM to obtain the ABSs' probabilities of distress and losses given distress, as well as to account for collateral pool correlations properly. See [8] for the methodology used in RiskFrontierTM.

Figure 1 demonstrates how the ABS 1-year distress probability changes with the subordination level for collateral pools; PD values range from as low as 10 basis points to as high as 22%, covering a

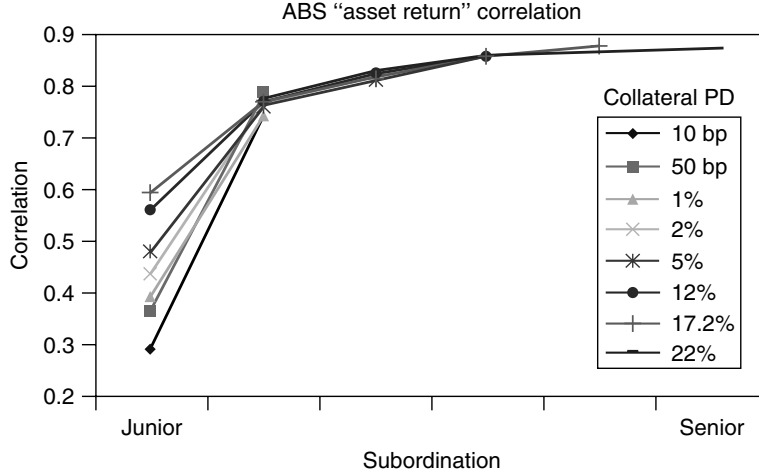


Figure 2 ABS loan-equivalent reference entity’s “asset return” correlation for collateral parameters: $R^2 = 10\%$, $LGD = 70\%$, and collateral PD ranging from 10 basis points to 22%

variety of instruments in varying economic environments—from prime loans in boom times to subprime mortgages during downturns. As expected, the ABS probability of distress is decreasing in the subordination level. Moreover, it appears that the distress probability curve starts out concave, switches, and then becomes convex. This convexity switch occurs at the maximum point of the bell-shaped collateral loss distribution function. Higher collateral PD pushes this maximum point to higher loss points.

To better understand the correlation structure of these ABSs, we compute the loan-equivalent reference entity’s R^2 for each deal. To this end, the simulation holds two similar copies of each ABS. Collateral pools for the two copies contain statistically similar instruments (but not the same instruments, i.e., different idiosyncratic shocks). The empirical JPD (JPD_{sim}) of the similar ABS pairs is calculated from the simulation results. Using a normal copula, we choose the loan-equivalent reference entity’s “asset return” correlation ρ_{12}^A , which results in the empirical JPD :

$$JPD(PD_1, PD_2, \rho_{12}^A) = JPD_{sim} \quad (2)$$

where $JPD(PD_1, PD_2, \rho_{12}^A)$ is the joint, 1-year distress probability of the two ABSs in a normal copula with correlation ρ_{12}^A and 1-year default probabilities

PD_1, PD_2 . To be clear, the “asset return” correlation ρ_{12}^A differs from the ABSs’ distress correlation ρ_{12}^D . Of course, the two are related:

$$\rho_{12}^D = \frac{JPD(PD_1, PD_2, \rho_{12}^A) - PD_1 \cdot PD_2}{\sqrt{PD_1 \cdot (1 - PD_1) \cdot PD_2 \cdot (1 - PD_2)}} \quad (3)$$

Generally, ABS “asset return” correlations are much higher than those of the underlying collateral. In our example, the asset return correlation for any pair of auto-loans, one from each ABS, is 10%. However, the “asset return” correlation for ABS loan-equivalents is 29% for the junior note and low collateral PD and significantly increases with subordination (Figure 2). This shows that the average collateral correlation is a highly biased estimate for the loan-equivalent correlation, since most of the idiosyncratic noise in the collateral pools washes out as the idiosyncratic shocks in the pool offset one another, leaving the common systematic factor to drive dynamics.

Results in Figures 1 and 2 are based on collateral pools of, for example, homogeneous auto-loans that share exposure to the same systematic factor (e.g., all in California). Therefore, the initial correlation between pairs of auto-loans was high to begin with. To examine the significance of pool diversification, consider an example with 50 state factors

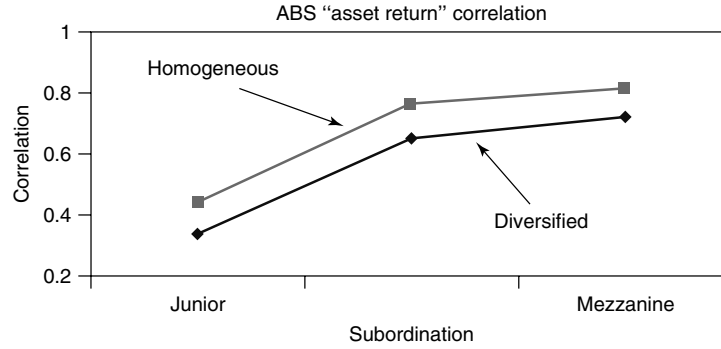


Figure 3 ABS loan-equivalent reference entity's "asset return" correlation—diversified pool (50 factors) *versus* homogeneous pool (common factor) with collateral pool parameters: $PD = 2\%$, $R^2 = 10\%$, $LGD = 70\%$

$(Z_1, \dots, Z_{50})$, where the auto-loans are equally distributed between the states. Each auto loan i has an associated state factor $k(i)$, such that $r_i = R \cdot Z_{k(i)} + \sqrt{1 - R^2} \cdot \varepsilon_i$. Using Moody's KMV's GCorr Retail™, we set the correlation between any two state factors to be 65% ($\rho(Z_k, Z_l) = 0.65$ for any $1 \leq k < l \leq 50$).^c Figure 3 compares the ABS "asset return" correlations between the diversified pool (50 factors) and the homogeneous pool (common factor) for pools with collateral $R^2 = 10\%$ corresponding with a range of retail loans including student loans and consumer loans [8]. Collateral PD was set to 2%.

Figure 3 demonstrates that the loan-equivalent reference entity associated with a diversified pool is less correlated than that of a homogeneous pool. This finding is not surprising, given that state factors are not perfectly correlated, allowing for diversification. This exercise is particularly important when modeling a CDO of ABSs whose collateral pools are focused on different geographic locations.

Collateralized Debt Obligation (CDO) Correlations

Even though CDO deals are similar to ABS deals, several important dynamics specific to CDOs impact correlation structure. First, corporate entities have lower default probabilities and higher R^2 than typical reference entities associated with ABS collateral instruments. Second, names commonly overlap

in collateral pools of different CDO deals. Thompson and Rajendra [7] show that some names appear in more than 50% of deals in Fitch's Synthetic CDO Index (Ford Motor Co -56.52% , and General Motors Corp -52.40%). Third, credit portfolios commonly have overlapping names across the single-name portion of the portfolio and the CDO collateral pool.

This section is divided into two subsections. The first analyzes the correlation between a CDO and a single-named instrument whose reference entity does not overlap with the CDO collateral pool. The second analyzes the correlation between two CDOs with varying degrees of collateral overlap.

Correlation between a Collateralized Debt Obligation (CDO) and a Single-name Instrument

Morokoff [6] calculated the loan-equivalent reference entity R^2 for a pass-through CDO tranche from the joint probability of tranche distress and single-name instrument default (outside the CDO's collateral pool), known henceforth as the joint default-distress probability. He considered a collateral pool of N -homogeneous instruments in a single factor environment. The single-name instrument was assumed similar to the instruments in the homogeneous collateral pool. Morokoff's methodology relies on using the independence of defaults, conditional on the realization of systematic risk factors, to compute the conditional joint default probability and then the expectation is taken over the systematic risk factor:

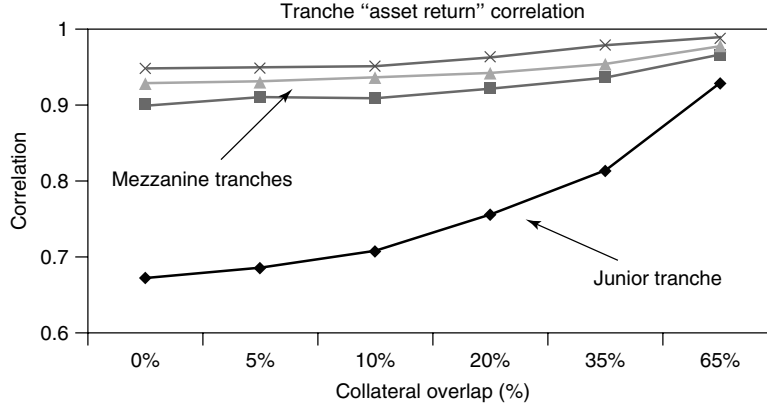


Figure 4 The effect of collateral overlap on tranche loan-equivalent reference entity's "asset return" correlation with collateral pool parameters: $PD = 2\%$, $R^2 = 25\%$, $LGD = 70\%$

$$\begin{aligned}
 P(\text{Tranche distress and single default}) &= E_Z(P(\text{Tranche distress and single default} | Z = z)) \\
 &= E_Z(P(\# \text{ defaults in pool} \geq \hat{a} | Z = z) \cdot p(z)) \\
 &= E_Z\left(\left(1 - \sum_{k=0}^{\hat{a}-1} P(\# \text{ defaults in pool} = k | Z = z)\right) \cdot p(z)\right) \\
 &= E_Z\left(\left(1 - \sum_{k=0}^{\hat{a}-1} \binom{N}{k} p(z)^k (1 - p(z))^{N-k}\right) \cdot p(z)\right) \\
 &= p - \sum_{k=0}^{\hat{a}-1} \binom{N}{k} E_z(p(z)^{k+1} (1 - p(z))^{N-k})
 \end{aligned} \tag{4}$$

where $\hat{a}$ is the required number of defaults in the pool to cause a distress to the tranche, $p(z) = N \left(\frac{N^{-1}(p) - R \cdot z}{\sqrt{1 - R^2}} \right)$ represents the PD conditional on the common factor $Z = z$, and $p = E_Z[p(Z)]$.

Similar to our analysis of ABS correlations, one can infer an "asset return" correlation ρ from the default probability, distress probability, and joint default-distress probability using a normal copula: $JPD = N(N^{-1}(\text{tranche distress prob}), N^{-1}(p), \rho)$.

Finally, $R^2_{\text{Loan equivalent}}$ can be computed as $\frac{\rho^2}{R^2}$.

Note that these calculations present an alternative to the simulations in the ABS section. However, the computation becomes infeasible for large pools and is, therefore, more suitable for CDOs. The normal approximation to the binomial distribution

can alleviate the cumbersome calculations, but for some values of the systematic factors z , the conditional default probability $p(z)$ is too small for the normal approximation to work.

Correlation of Collateralized Debt Obligations (CDOs) with Overlapping Names

Intuitively, the degree of overlap associated with two CDO tranches increases their correlation. To better understand the quantitative impact of overlap, we used simulation methods similar to those in the section Modeling the Correlation of Two ABSs. The relevant difference is that a certain percentage of the collateral pool is assumed to overlap when analyzing the correlations across the deals. Figure 4

presents the tranche loan-equivalent asset return correlations computed using the methodology in the section Modeling the Correlation of Two ABSs. The analysis was conducted on a high-yield deal with typical high-yield CDO properties ($PD = 2\%$ and $R^2 = 25\%$). Along the x -axis, the collateral overlap varies from 0 to 65%. Each curve represents a subordination level for the tranche. Interestingly, the loan-equivalent asset correlation increase with overlap is more substantial for junior tranches. This follows from the susceptibility of junior tranches to idiosyncratic shocks. After all, the junior tranche can easily experience distress even when the systematic shock is high; all it takes is a single reference entity to realize a particularly low idiosyncratic shock. On the other hand, senior tranches are extremely unlikely to experience distress when the systematic shock is high. Therefore, the degree of overlap, which only affects idiosyncratic shocks, has a much greater effect on junior tranches.

Acknowledgments

We would like to thank Zhenya Hu for help with the simulations and Mikael Nyberg for his suggestions.

End Notes

^a. Exceptions include Moody's KMV RiskFrontier™, which models the terms of the subordinated note along with the correlation structure of the underlying collateral pool as it relates to the other instruments in the portfolio. Please see [5] for additional details.

^b. For example, PD , and LGD estimates are available through Moody's Economy.com, and R^2 estimates are available through Moody's KMV.

^c. Moody's KMV's GCorr Retail™ provides pairwise correlations for retail counterparties defined by MSA in the

UnitedStates (e.g., San Francisco or New York City) and product type (e.g., auto loan or student loan, etc.). The correlations are estimated using delinquency rates data from Equifax and Moody's Economy.com. For more details on Moody's KMV's GCorr Retail™ please see [8].

References

- [1] Agca, S. & Islam, S. (2007). *Can CDO Equity Be Short on Correlation?* working paper.
- [2] Coval, J.D., Jurek, J.W. & Stafford, E. (2008). *Economic Catastrophe Bonds*, HBS Finance working paper No. 07–102, available at: <http://ssrn.com/abstract=995249>
- [3] Coval, J.D., Jurek, J.W. & Stafford, E. (2008). *Re-Examining the Role of Rating Agencies: Lessons from Structured Finance*. working paper.
- [4] Gregory, J. & Laurent, J.P. (2004). In the core of correlation, *Risk* October, 87–91.
- [5] Moosbrucker, T. (2006). Explaining the correlation smile using variance gamma distributions, *The Journal of Fixed Income* **16**(1), 71–87.
- [6] Morokoff, W. (2006). *Modeling ABS, RMBS, CMBS: 'Loan Equivalent' Approach in Portfolio Manager*. unpublished work done at Moody's KMV.
- [7] Thompson, A. & Rajendra, G. (2006). *Global CDO market 2005 review*. Deutsche Bank whitepaper.
- [8] Wang, J., Zhang, J. & Levy, A. (2008). *Modeling Retail Correlations in Credit Portfolios*. Moody's KMV research whitepaper
- [9] Zeng, B. & Zhang, J. (2001). *An Empirical Assessment of Asset Correlation Models*, Moody's KMV whitepaper, available at: http://www.moodyskmv.com/research/files/wp/emp_assesment.pdf
- [10] Zhang, J., Zhu, F. & Lee, J. (2008). *Asset Correlation, Realized Default Correlation, and Portfolio Credit Risk*, available at: http://www.moodyskmv.com/research/files/wp/Asset_Correlation_and_Portfolio_Risk.pdf

TOMER YAHALOM, AMNON LEVY &
ANDREW S. KAPLIN

Special-purpose Vehicle (SPV)

A special-purpose vehicle (SPV) is a legal entity created by a firm for a specific business objective, usually in the context of **securitization**. SPVs are often used to finance a project or issue securities without putting at risk the original holder of the securitized portfolio.

We focus here on quasi-operating companies, which are SPVs operating in primarily one type of business: interest rate and foreign exchange derivatives, credit derivatives, and buy and hold assets or others. Derivative product companies (DPCs), structured investment vehicles (SIVs), and credit derivative product companies (CDPCs) are the most well-known operating companies.

Derivative Product Companies (DPCs)

(DPCs) are intermediaries between financial institutions (known as their *parents* or *sponsors*) and their third-party counterparties [2, 5]. DPCs intermediate swaps between the sponsor and third parties under the approved International Swaps and Derivatives Association (ISDA) Master Agreement. Enhanced subsidiaries differ from other derivative-product subsidiaries, as their credit ratings do not depend on their parent's guarantee. A DPC may engage in over-the-counter interest rate, currency, and equity swaps and options as well as certain exchange-traded futures and options depending on its individual structure. A DPC is capitalized at a level appropriate for the scope of its business activities and desired rating. In most cases, DPCs have been set up to overcome credit sensitivity in the derivative-product markets. There are two types of DPCs: continuation and termination structures. The continuation structures are designed to honor their contracts to full maturity even when a wind-down event occurs, while termination structures are designed to honor their contracts to full maturity, or should certain events occur, to terminate and cash settle all their contracts prior to their final maturity. DPCs are typically AAA rated and are often referred to be the *AAA face of the sponsor*. They are market risk neutral by mirroring their trades with the third parties with the parent

or the sponsor. They are exposed to the credit risk of third parties. The structure is equipped with exit strategies and resources so that upon certain wind-down scenarios, the vehicle is expected to meet its derivative obligations with AAA certainty.

The market for DPCs developed in early 1990. Every bank seeking to be eligible as an AAA counterparty in derivative contracts sponsored its own DPC.

Credit risk of third-party counterparties is quantified by sophisticated models. Potential future market environment is simulated and valuation modules are used to project the mark-to-market of each swap contract. By combining market paths with credit paths (in which the creditworthiness of the counterparty is simulated), one can assess where capital is being deployed to cover for losses. The potential losses corresponding to each market path can be analyzed by combining the results of default simulations and the counterparty exposures. A consideration of losses across all market paths permits the construction of a distribution of potential credit losses. The credit enhancement to protect against losses at a given level of confidence may be analyzed. This risk model can also quantify the potential change in the portfolio's value over a period of time.

A DPC with a continuation structure generally receives collateral from the parent to cover its exposure to the parent resulting from the back-to-back trades. This collateral amount, after appropriate discount factors are applied, is equivalent to the net mark-to-market value of the DPC's portfolio of contracts with its parent. Upon the occurrence of certain events, however, the management of the DPC's portfolio will typically be passed on to a contingent manager.

In the short period prior to the transfer of portfolio management to the contingent manager, the value of the DPC's contracts with its parent could rise. Using the capabilities of the risk model, the potential increase in the DPC's credit exposure to the parent may be quantified.

In a termination structure, the value of the DPC's portfolio can change over the period beginning with the last regular valuation date and ending at the early termination valuation date upon occurrence of a termination trigger event. Again, the potential change in the portfolio's value may be determined at the desired level of confidence by using the same risk model.

DPCs are equipped with a liquidity model that covers short-term liquidity squeezes and with operational capital that covers operational risks.

Structured Investment Vehicles (SIVs)

SIVs are limited-purpose operating companies that take arbitrage opportunities by purchasing mostly highly rated medium- and long-term assets and funding themselves with cheaper short-term commercial paper (CP) and medium-term notes (MTNs) [3, 6, 7].

SIV-Lites combine features of both collateralized debt obligations (CDOs) and SIV technologies. They typically purchase high-grade asset-backed securities (ABSs), primarily residential mortgage-backed securities (RMBS), but may also include a small portion of commercial mortgage-backed securities (CMBS) or other ABSs and fund themselves by issuing short-term CP or repurchase agreements (REPOs) and MTNs. SIVs and SIV-Lites roll their senior short-term liabilities (REPOs and CPs) unless a market disruption event occurs or any other liquidation trigger is reached.

When analyzing SIVs or SIV-Lites, stochastic cash-flow models could be used to quantify the risks they are exposed to. The main risk factors of the asset portfolio, which include credit migration and market risk, are captured and projected by Monte Carlo simulation. Credit migration measures the new credit profile of the portfolio. Defaults net of recovery result in loss in the portfolio. Upgrades and downgrades directly affect the credit profile and the market value of the portfolio. Asset spreads and asset market values are projected in the capital model as well as credit migration. Historical spread data are used to calibrate the asset spread model. Rating migration, default correlation, and recovery assumptions are based on historical default studies and applied to the capital model. Market risk involves interest rates and foreign exchange rates (if there exists foreign currency valuation) modeling. SIVs or SIV-Lites are designed to be market risk neutral. Additional interest rate and foreign exchange rate sensitivity tests are usually used to this end. These tests mainly measure the change in the net asset portfolio value caused by a sudden change in the interest rates or foreign exchange rates. Liquidation risk is the most complicated risk factor to be modeled. Liquidation assumptions on the assets are required

when senior debts cannot be rolled over. Haircuts on the assets for the liquidation purpose are based on stressed historical asset price movements and applied appropriately when needed.

Since the stability of capital requirement is one of the key components in the risk management of SIVs or SIV-Lites, a certain large number of Monte Carlo paths are generated to test whether the model converges in a good manner.

A liquidity model might be used to monitor the vehicle's internal liquidity relative to the liabilities. Net cumulative outflow (NCO) tests are normally calculated for each rolling 1, 5, 10, and 15 business day period commencing on the next day of calculation through and including the day that is one year from the day of such calculation, for example, the vehicle needs to determine on a daily basis its 1, 5, 10, and 15 day peak NCO requirements over a period of one year.

SIVs and SIV-Lites also face management and operational risk that is covered by additional capital.

In mid-2007, SIVs and SIV-Lites experienced a number of difficulties owing to, among other things, the liquidity crunch and spread widening. Various SIVs and SIV-lites were downgraded by rating agencies. Some of them went into default. SIV exposures in rated funds have dropped dramatically since the last quarter of 2007.

Credit Derivative Product Companies (CDPCs)

CDPCs are special-purpose entities that sell credit protection under credit default swaps (CDS) or certain approved forms of insurance policies [1, 4]. Unlike traditional DPCs, which are engaged in interest rate, currency, and equity swaps and options, CDPCs sell protection on single-name obligors, such as corporate, sovereign, and asset-backed securities, or on tranching, structured finance obligations. CDPCs can also buy protections or enter interest rate and currency swaps. However, that is mainly for hedging purposes.

When analyzing CDPCs, sophisticated models are built to quantify the risk of CDPCs and analyze the amount of capital needed to meet the obligations for their counterparty ratings and various debt-note ratings, respectively. These obligations include payments on credit events, payments on senior fees

expenses, and potential termination payments upon counterparty defaults.

The key risk factors that CDPCs might be exposed to are credit risk for the reference entities, counterparty risk, and market risk. Credit risk is the primary concern for a CDPC. If the reference entity on which a CDPC sells protection defaults, the CDPC will suffer a loss. Historical default studies and rating analysis are applied to simulate the time to default of the reference portfolio. Correlation and recovery assumptions are used to size the loss. Single-period time-to-default models are efficient tools to model credit risk. However, multiperiod rating transition models become necessary when the credit qualities of the underlying assets are required to be featured in the capital model. When modeling the credit risk of CDO tranches, drilling down to the underlying obligors is more effective to model the correlation risk.

When a CDS counterparty experiences a default, the swap contracts that are associated with that counterparty are unwound and a termination payment (fair market value) may need to be calculated. Credit spread widening can lead to large termination payments upon counterparty default. Spread models quantify the potential deterioration in the creditworthiness of the assets as well as market volatility. Spread data are usually grouped in rating and industry categories.

Market standard valuation modules for single-name CDS contracts and tranching CDO transactions are generally used to calculate the fair market value upon the counterparty defaults. Besides credit spreads, interest rate projection is necessary for the discounting of future cash flows in order to calculate present values, the amount of coupon paid on the notes, and the amortization schedule for CDPCs that invest in prepayment-sensitive assets such as ABSs. A CDPC is exposed to foreign exchange risk when it makes or receives payments in more than one currency, and no hedge has been set up to neutralize the mismatch. In this case, the foreign exchange rate is modeled in the capital model. The cash-flow waterfall is added to the default model according to

the structure of the CDPCs. CDPCs also face liquidity risk. A liquidity model might be developed to size capital for short-term needs.

Like modeling SIVs or SIV-Lites, the model calibration is a complex exercise. A certain large number of Monte Carlo paths are generated to test appropriate model convergence.

CDPCs also face operational risks and CDPC management risk. Additional capital is typically assigned for these risks.

Since CDPCs have no market value triggers that would force them to sell assets or reduce leverage, they have not been affected as significantly by the subprime mortgage crunch as SIVs or SIV-Lites in 2007. However, CDPCs have been experiencing a hard time to find counterparties because of the volatility of the spreads.

References

- [1] *Criteria for Rating Global Credit Derivative Product Companies*, Standard & Poor's, www.ratingsdirect.com
- [2] Gupton G.M., Finger C.C. & Bhatia M., Morgan Guaranty Trust Company, (1997). *CreditMetrics*, Technical Document, April 1997.
- [3] Merrill Lynch (2005). *Fixed Income Strategy*, "SIVs are Running Strong", January 28, 2005.
- [4] Polizu, C., Jiang, J. & Venus, S. (2007). *Structured Finance ViewPoint on Quantitative Analytics: Creating Transparency To Better Manage Risk*, Standard & Poor's.
- [5] *Rating Derivative Product Companies S&P Structured Finance Criteria*, Feb 2000, www.ratingsdirect.com
- [6] de Servigny, A. & Jobst, N. (2007). *Quantitative Handbook of Structured Finance*, 1st Edition, McGraw-Hill.
- [7] *Structured Investment Vehicle Criteria* (published on March 13, 2002), www.ratingsdirect.com

Related Articles

Collateralized Debt Obligations (CDO); Credit Default Swaps; Securitization.

CRISTINA POLIZU & JENNIFER JIANG

Credit Portfolio Insurance

Credit portfolio insurance products such as credit constant proportion portfolio insurance (CPPI) and constant proportion debt obligations (CPDOs) utilize similar technologies that use leverage as a mechanism to enhance returns. Credit CPPIs and CPDOs evolved in the low credit spread environment witnessed recently, especially between the years 2004 to about mid-2007 as shown in Figure 1. For many investors, the tightness in investment grade spreads rendered investment grade return targets difficult to achieve other than through rule-based, leveraged credit strategies. Credit CPPIs and CPDOs are two alternative formats of gaining leverage in a similar class of investments,^a with the first credit CPPI products being introduced in 2004 followed by CPDO products in 2006.

Credit CPPI

Historically, the CPPI concept came into being in the context of equity portfolios. Some of the earlier works in this area are by Black and Jones [1], Black and Perold [2], and Black and Rouhani [3]. For a CPPI, the term *insurance* is a loosely defined expression that refers to principal protection of the issued CPPI notes. The principal component of the notes, called the *bond floor*, is secured by virtue of the fact that an appropriate portion of the note issuance proceeds is invested in risk-free assets. The remainder of the proceeds, called *cushion* or *reserve*, is used to take leveraged exposure to risky credit assets in the case of a credit CPPI. Combined earnings from the risk-free and risky assets generate the cashflows to pay the periodic interest and final principal on the CPPI notes.

A simple example of a CPPI would be a structure where proceeds from the CPPI note issuance are divided into two components—the first component is invested in a “risk-free” zero coupon bond with a face value equal to the principal of the CPPI note and the second component acts as a reserve account for taking exposure to more risky assets through credit default swaps (CDSs). The idea is that the risk-free bond will back the repayment of principal at maturity, while the more risky portfolio of CDS assets will be used to generate spread income and to pay the interest on the CPPI note.

Figure 2 describes the structure of a CPPI. The CPPI investor provides the initial funds equivalent to the notional amount of the issued CPPI notes and the proceeds are utilized to set up (i) a reserve account for investing in risky assets and (ii) a deposit account for investing in “risk-free” assets. Leverage is obtained on the risky portion of the balance sheet by taking positions that make the notional exposure to the risky assets to be some multiple of the reserve account balance. Although the choice of these risky assets can include all types of assets, for a credit CPPI the typical investment would be in the form of exposure to a credit index, standardized or bespoke synthetic CDOs, cash flow CDOs, equity default swaps (EDSs),^b and so on.

Figure 3 shows the expected evolution of the *bond floor* and the more risky portfolio value over the term of the CPPI. The actual value of the risky portfolio and *bond floor* at a particular point in time will, however, depend on the evolution of credit CPPI risk factors such as movements in credit migrations, default events, interest rates, and loss given default. Moreover, the performance of the CPPI is quite sensitive to the leverage multiplier, and to mitigate this risk, the CPPI is often required to reduce leverage through rebalancing as losses on the risky portfolio eat into the reserve account.^c

Credit CPPI Risk Factors

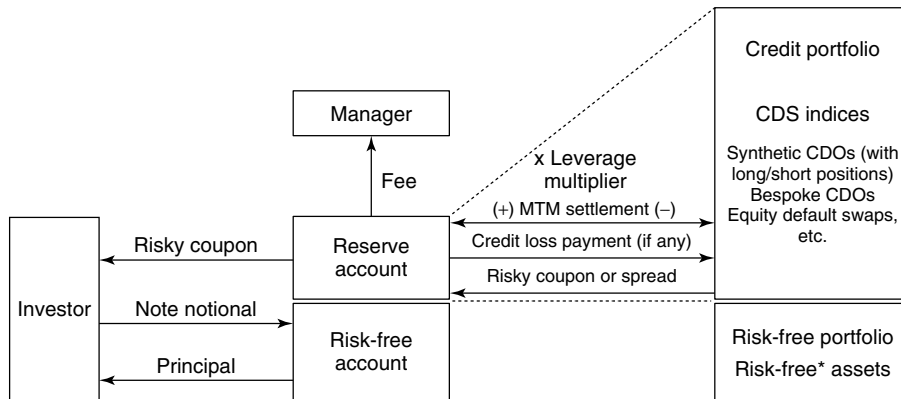
Gap risk: This is the possibility that rebalancing the risky portfolio may not happen at the speed of its market value deterioration. In such a scenario, a loss of principal could occur if a sudden loss in market value exceeds the reserve cash account in the form of the *cushion*. Moreover, the higher the leverage multiplier, the higher the gap risk. A less drastic scenario is when a loss on the risky portfolio nearly erases the *cushion* and leaves only risk-free assets in the portfolio—in that case although the principal is still protected, the promised CPPI note coupon can no longer be paid in full.

Default risk: Defaults in the more risky portfolio, both in terms of timing and loss given default, affect the expected returns and the principal protection of the CPPI notes. Some structures mitigate default risk by imposing restrictions on investments in subinvestment grade assets.

Interest rate risk: Changes in interest rates lead to fluctuations in the zero coupon *bond floor*, which



Figure 1 Evolution of the CDX North American Investment Grade (CDX.NA.IG) Index over the period July 2004 to July 2007



* These assets have no default risk but may be subject to interest rate or prepayment risk.

Figure 2 Cash flow diagram describing a credit CPPI structure

in turn impacts the reserve available for taking leveraged exposure to credit risk assets. Since the objective of the credit CPPI is to limit downside while maximizing returns through dynamic rebalancing of the portfolio between “risk-free” and more risky assets, changes in the *bond floor* due to interest rate movements affect the rebalancing strategies.

Performance risk: The CPPI manager’s ability in implementing the credit CPPI dynamic portfolio rebalancing strategy may also be important.

Modeling a Credit CPPI

A Monte Carlo simulation approach is an effective means of modeling the risk of a credit CPPI transaction. The four main features of a bottom-up

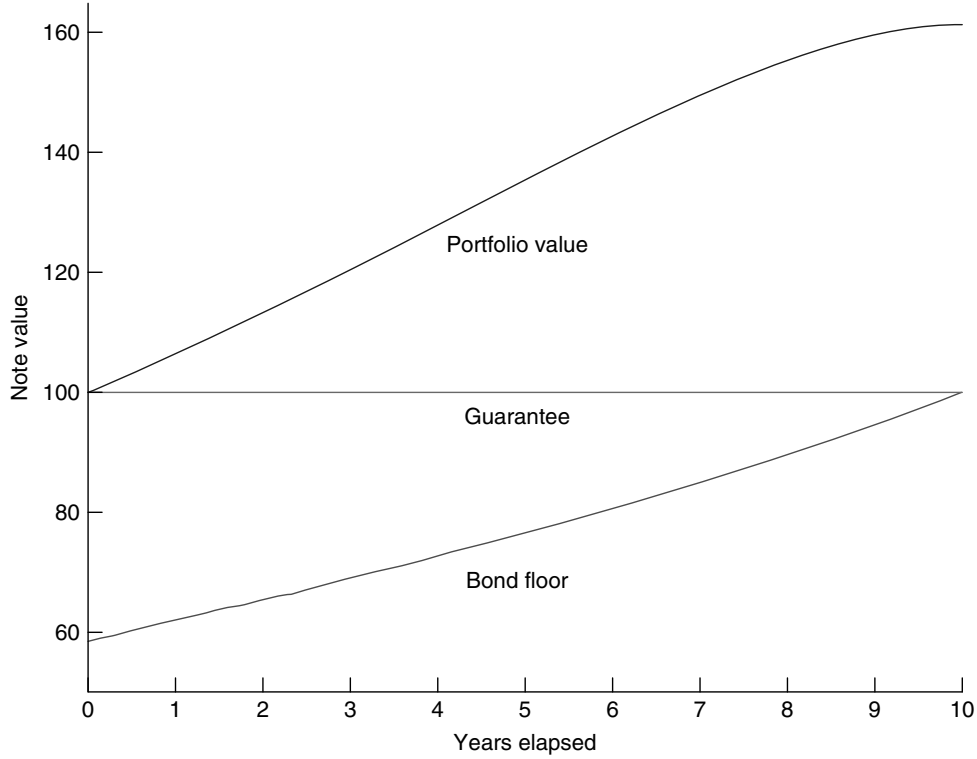


Figure 3 Expected value of the bond floor and the more risky portfolio over the life of a CPPI

credit CPPI model should include (i) single name CDS spread curve dynamics, (ii) correlation between spread movements and credit migrations across the names in the portfolio or index, (iii) losses due to defaults and associated recoveries, and (iv) interest rate dynamics for cash flow discounting, *bond floor* value estimation, and liability payment calculations.

Credit quality changes in the form of rating migrations may be modeled in a structural framework where the normalized asset return, A_k , of asset k can be decomposed into systematic and idiosyncratic components:

$$A_k = \sqrt{\rho}z + \sqrt{1 - \rho}\varepsilon_k \quad (1)$$

For a Gaussian copula framework, the systematic and idiosyncratic latent factors, that is, z and ε_k , respectively, would have standard normal distributions and ρ would be the asset correlation parameter. The probability of asset k migrating from its current rating level R_i to a new rating level R_j would be given by the following expression in this

case:

$$\sum_{m=1}^{j-1} \Pr(R_i \rightarrow R_j) \leq \Phi(A_k) < \sum_{m=1}^j \Pr(R_i \rightarrow R_j) \quad (2)$$

Credit spreads are often modeled as mean reverting processes, where the mean reversion speed and spread volatility could be made functions of rating levels. One such specification for the spread model could be

$$\begin{aligned} \log(S_k(t_{i+1})) &= \delta \log(S_k(t_i)) + (1 - \delta) \\ &\times \log(S^*(R_k(t_{i+1}))) + \sigma \sqrt{\delta} \Delta t \\ &\times (\sqrt{\rho_s} A_k + \sqrt{1 - \rho_s} \varphi_k) \end{aligned} \quad (3)$$

where $\delta = \exp(-\kappa \Delta t)$, κ is the mean reversion speed, σ is the spread volatility, S^* is the average spread associated with rating R , A_k is the asset return used in the credit migration simulation, ρ_s is the correlation between the asset return and the spread, and $\varphi_k \sim N(0,1)$.

One may also include a jump component in the above specification, or a regime switching module, for enhancing the possibility of more sudden and drastic spread moves for a more conservative estimation of gap risk. For example, Cont and Tankov [5] developed a model that includes jumps in the asset price process.

Lastly, an appropriate interest rate process based on any popular spot rate or forward rate model such as the one by Hull and White [7] or Heath *et al.* [6] can be incorporated in the above framework for cash flow discounting calculations.

CPDO

CPDOs use a similar leveraging technology as described previously for CPPIs but with two major exceptions. First, there is no guaranteed return of principal and, second, rebalancing in the form of leverage changes is in the opposite direction to that of a CPPI, that is, CPDO rebalancing is based on a “buy low, sell high” strategy where a spread widening leads to an increase in the notional exposure to the more risky portfolio, subject to the maximum leverage cap. This is designed to allow the structure to compensate for the loss in value due to the spread widening by increasing the spread income from more risky assets in future periods.

The economics behind CPDO transactions backed by exposure to investment grade credits such as an index CPDO is the fact that investment grade

names typically display an upward sloping term structure of spreads. The theory is that the CPDO can take exposure to a new on-the-run index series at a relatively high spread level (corresponding to $5\frac{1}{4}$ years maturity) but trade out of the series at a significantly lower spread level (corresponding to $4\frac{3}{4}$ years) as the index becomes off-the-run.^d In the context of CPDOs, this dynamic is known as the *roll down the curve* benefit and the aim of the CPDO is to maximize this advantage by employing leverage.

Figure 4 describes a typical CPDO structure. Upon initiation, all cash proceeds from the CPDO investor are placed in a deposit account and get invested in short-term liquid instruments that earn interest at the risk-free benchmark rate, typically LIBOR. The deposit account also acts as the reserve account for taking exposure to a more risky credit portfolio of CDS assets (or a CDS index) where the credit portfolio notional = CPDO note notional* leverage factor. The most popular choice for the risky assets is a portfolio providing equal exposure to the CDX North American Investment Grade (CDX.NA.IG) and the iTraxx Europe Investment Grade indices.^e Moreover, to mitigate default and liquidity risk, such index CPDOs are generally required to sell protection only for “on-the-run” index series so that exposure to index constituents’ names that get downgraded to below investment grade rating levels is automatically limited because of the substitution rules imposed when indices roll over every six months.

The deposit account is credited with interest from the liquid short-term assets, spread premia from

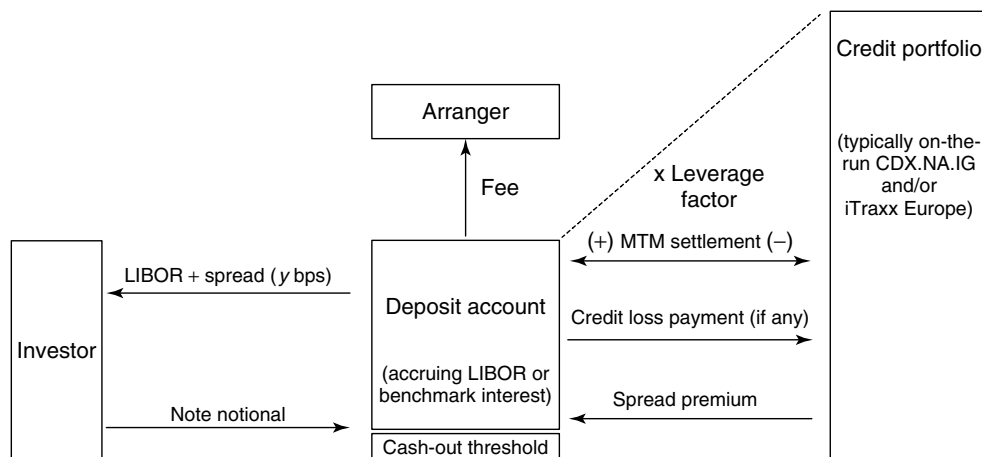


Figure 4 Cash flow diagram describing a CPDO structure

the CDS credit portfolio, and any positive mark-to-market (m-t-m) settlements on the credit portfolio. The deposit account gets debited with coupon payments to the CPDO investor, management fees to the arranger, any negative m-t-m settlements, or credit loss payments due to defaults in the CDS portfolio. The performance of the CPDO thus depends on how the *net asset value* (NAV) at time t , which is equal to the sum of the deposit account balance, the m-t-m of the CDS portfolio, and the accrued spread income, compares to the *target value* (TV) at time t , which is the amount if invested at the risk-free rate would be sufficient to meet the CPDO coupon and principal payment obligations, as well as any fees.^f The difference between $TV(t)$ and $NAV(t)$ is called the *shortfall*, that is,

$$\text{Shortfall}(t) = TV(t) - NAV(t) \quad (4)$$

The leverage employed in the CPDO transaction is a function of the ratio of the *shortfall* to the risky income calculated as the present value of the expected spread premium earnings. Note that this ratio quantifies the amount of the current shortfall when compared with the cash flow expected from the risky assets. More specifically, the transaction has predefined rules for estimating a target notional defined as

$$\begin{aligned} \text{Target notional} &= \text{Gearing factor} \\ &\times \frac{\text{Shortfall}}{PV(\text{expected CDS premia})} \end{aligned} \quad (5)$$

Some structures define a flat gearing factor, while others may employ a time-dependent gearing factor. Leverage is then simply the ratio of the target notional to the CPDO note notional. From equation (5), it can be seen that if spreads widen, causing the CPDO to incur m-t-m losses that increase the *shortfall* relative to the expected risky spread earnings, leverage increases. This is contrary to a CPPI leverage strategy where any m-t-m losses on the more risky portfolio lead to a delivering of the transaction.

At the commencement of a CPDO transaction, the *shortfall* is positive as $NAV(t_0)$ is simply the CPDO note notional proceeds *minus* any upfront structuring fees since the value of the CDS portfolio is zero at initiation and there are no accrued spread premium earnings yet. With time, however, the shortfall is

expected to decrease to zero, in which case the CPDO is said to have a “cash-in” event. This implies that the structure no longer needs to take on any risky exposure and the reserve deposit account is large enough to meet all future interest, principal, and fee payment obligations by simply investing in risk-free assets. In adverse market conditions, however, frequent and large m-t-m losses on the portfolio and/or defaults may lead to severe erosion of the deposit reserve account. In some instances, the NAV may drop below a certain threshold, typically defined at 10% of the issued CPDO note notional. In this case, the CPDO transaction is unwound and is said to have suffered a “cash-out” event. The only other event possible is when the transaction fails to redeem the CPDO note at par upon maturity, in which case the CPDO investor incurs a loss, though not as severe as under a “cash-out” event.

CPDO Risk Factors

Similar to CPPIs, the main risk factors for a CPDO structure are as follows:

Spread risk: Sudden (and correlated) widening of spreads can lead to large m-t-m losses that eat into the cash reserve account. With index CPDOs, this risk may also be in the form of rollover risk, whereby index constituent names that deteriorate in credit (and show spread widening) get downgraded and get replaced with tighter spread names. Since the CPDO is generally required to roll over its exposure to the on-the-run index series, the effect of the rollover is twofold. On the one hand, riskier names have to be effectively unwound at wider spreads, resulting in m-t-m losses. On the other hand, these names get replaced in the index with higher rated names, implying lower spread earnings after the roll that would have ordinarily compensated for the m-t-m losses.^g

Default risk: Similar to CPPIs, CPDO investors are exposed to default risk as protection sellers. However, generally speaking, this risk is greatly mitigated in index CPDOs due to the rollover every six months.

Liquidity risk: Since index CPDOs are required to roll their exposure to the index series every six months, they can be more prone to liquidity risk. Even though on-the-run index series are the most liquid, bid–offer spreads may widen at the time of rollover as dealers anticipate more one-sided trades.

Interest rate risk: In “normal” circumstances, a natural hedge exists since liabilities, that is, CPDO notes outstanding (paid LIBOR) are less than or equal to the cash reserve account (earns LIBOR). However, interest rate risk may become significant if large m-t-m (or default) losses eat into the reserve account, leading to a mismatch between the CPDO notes outstanding and the size of this reserve deposit account.

Modeling a CPDO

CPDOs can be modeled using a similar bottom-up Monte Carlo approach as described for CPPIs. For index CPDOs, the analysis should also include a means of modeling index changes to capture the impact of rollover risk on the CPDO note performance. This can be incorporated in the suggested framework by using the ratings-based credit migration model to analyze changes to the index by appropriate replacement of credits that get downgraded to below investment grade rating categories.

Cont and Jensen [4] provided an alternative modeling approach using a top-down method. In their framework, rollover effects are incorporated *via* jumps in the index spread at each roll date.

End Notes

^aSome investors also consider synthetic CDO tranches as an alternative means of gaining leverage in this class of investments. We, however, differentiate CPPIs and CPDOs from synthetic tranches since the latter are correlation products.

^bEDSs have risk characteristics that are similar to CDSs and hence a CPPI with exposure to EDSs is classified as a credit CPPI.

^cAlternatively, an increase in the more risky portfolio value may result in an increase in the notional exposure to the more risky assets up to a predefined maximum leverage

cap. For this reason, a credit CPPI strategy is also known as *buy high, sell low*.

^dThis assumes that the average credit quality of the index remains more or less the same over the six-month period.

^eThis equally weighted combination of the CDX and iTraxx indices is frequently referred to as the *GLOBOXX index*.

^fIn other words, the *target value* is the present value of all liabilities discounted at the risk-free rate.

^gIn actual practice, the CPDO takes exposure to the index directly and not individual names, so the effect of the index rollover is a sudden drop in the index spread as the more risky names get replaced with the less risky ones.

References

- [1] Black, F. & Jones, R. (1987). Simplifying portfolio insurance, *Journal of Portfolio Management* **14**, 48–51.
- [2] Black, F. & Perold, A. (1992). Theory of constant proportion portfolio insurance, *Journal of Economics, Dynamics and Control* **16**, 403–426.
- [3] Black, F. & Rouhani, R. (1989). *Constant Proportion Portfolio Insurance and the Synthetic Put Option: A Comparison*. Institutional Investor Focus on Investment Management, pp. 695–708.
- [4] Cont, R. & Jensen, C. (2009). *Constant Proportion Debt Obligations (CPDO): Modeling and Risk Analysis*. <http://ssrn.com/abstract=1372414>.
- [5] Cont, R. & Tankov, P. (2007). Constant proportion portfolio insurance in presence of jumps in asset prices, *Mathematical Finance* **19**(3), <http://ssrn.com/abstract=1021084>.
- [6] Heath, D., Jarrow, R. & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology, *Econometrica* **60**(1), 77–105.
- [7] Hull, J. & White, A. (1990). Pricing interest rate derivative securities, *Review of Financial Studies* **3**(4), 573–592.

Related Articles

Constant Proportion Portfolio Insurance; Credit Migration Models.

SAIYID S. ISLAM

Structured Finance

Rating Methodologies

Rating agencies are among the most important players in the securitization market and they have contributed significantly in the development of the market. Rating agencies have the expertise and access to necessary information to rate the multiple tranches of structured finance transactions. Their ratings are independent assessments of the credit and noncredit risks in a transaction. Investors who do not have the expertise and the required information to assess the credit quality of the issued structured finance products can use the ratings in their investment decision process. In addition, rating agencies provide transparency and increase the effectiveness in the securitization market. Finally, rating agencies have contributed significantly to the rapid development of the securitization market in the recent years.

Rating agencies have developed methodologies for structured finance ratings, which are significantly more complex than those they use to rate traditional instruments. Quantitative modeling is essential for rating structured finance securities. All three major agencies use a three-step rating process to model the assets and liabilities of a transaction. Although the rating methodologies across asset classes share the same basic principles, the modeling differs across asset classes and jurisdictions.

We examine the asset-backed securities (ABS) (see **Securitization**) and collateralized debt obligation (CDO) (see **Collateralized Debt Obligations (CDO)**) rating methodologies applied in the securitization market ([5, 6] for an introduction to securitization process). We focus on the quantitative aspects of the modeling that the major rating agencies apply in order to rate ABS and CDO tranches with a discussion on the importance and uses of the ratings in the securitization market.

Rating Methodologies

Rating agencies update their rating methodologies regularly to increase the accuracy and to incorporate new structures that appear in the market. For previous reviews on rating methodologies, see [1, 7].

Although the rating methodologies differ across rating agencies, all three major agencies, that is, Fitch, Moody's, and Standard & Poor's follow a three-step rating approach. The basic principles of this three-step approach are applied to structured finance transactions. In general, the three steps of a rating process are the following:

1. calculation of the default/loss distribution of the portfolio of assets;
2. generation of the asset cash flows using the portfolio loss characteristics; and
3. generation of the liability cash flows using the asset cash flows.

However, the modeling of each of the above steps can be very different depending on the type of structure (synthetic or cash) or the type of collateral (such as auto loans, mortgages, bonds, or others). The first two steps focus on the asset side and the third step on the liabilities side of the transaction. The next two sections discuss the basic principles of the assets and liabilities modeling that the agencies use to determine structured finance ratings.

Asset Side Modeling

In the first step of their respective methodologies, the agencies assess the credit risk in the underlying asset portfolio. They do so by modeling the default/loss distribution of the portfolio. The three main inputs driving these models are the following:

1. *Term structure of probabilities of default* (PDs) of each individual obligor in the portfolio across the life of the transaction.
2. *Recovery rates* (or losses-given-default (LGD), which is equal to one minus the recovery rate in the event of default).
3. *Asset correlations* within the portfolio, which determine default correlations and thus the likelihood of occurrence of joint defaults in a given period.

The main outputs of the credit risk assessment models are the loss characteristics of the portfolio, that is, the portfolio loss and default distributions.

Typical ABS portfolios usually contain large, granular, and homogeneous pools of assets. Hence, the systematic risk factors are typically the primary

drivers of the default distribution of the ABS portfolios. Moreover, model inputs, such as the default probabilities, the recovery rates, and the correlation assumptions, are usually estimated on a portfolio level. On the other hand, CDOs are typically nongranular portfolios of a small number of assets. Thus, the idiosyncratic risk is much more important in CDOs than the ABS portfolios. As a result of this, one has to carefully consider each obligor's individual probability of default and recovery rates.

As of February 2008, for rating CDO transactions, Standard and Poor's [10, 20] and Fitch [11, 14, 18] use very similar Monte Carlo simulation methodologies to calculate the loss/default distribution of CDO portfolios. Both the agencies use structural form models, where an obligor defaults if its assets value falls below its liabilities (also referred to as *default threshold*); see [16]. In these models, copulas are used to model the joint defaults of the obligors in the portfolio. In February 2007, Moody's introduced a Monte Carlo simulation approach [15] similar to the other agencies' methodologies to compute the expected loss for static synthetic CDOs and CDOs-squared. As of February 2008, Moody's continues to use the binomial expansion technique (BET) [3, 8, 21] for assessing the credit risk of other types of CDOs. For a comparison of the BET against the copula models, see [9]. In spite of the similarities in the agencies' methodologies, there are also some disparities in their approaches. Specifically, there are differences in the correlation structures and the recovery assumptions that the rating agencies employ.

The approaches that the rating agencies apply to all the other ABS transactions differ across asset classes and jurisdictions. These approaches are described next as of February 2008. Fitch uses a simulation-based methodology [2] appropriate for large, granular ABS portfolios composed of consumer and auto loans or leases for the European transactions, whereas it employs an actuarial approach to rate US ABS transactions; for example, it uses historical default rates to specify the credit risk of such portfolios. The key difference between the CDO and ABS simulation approaches is that the default probabilities inputs in the latter are specified at a pool level and not at an asset level. Fitch uses similar simulation-based methodology to the one that is used for CDOs in order to calculate the loss distributions of commercial mortgage-backed securities (CMBS) portfolios [4]. Fitch rates residential mortgage-backed securities

(RMBS) transactions by applying standard criteria at a country level. Historical information on a loan-by-loan level as well as on the level of the originator is used to arrive at probability defaults at a loan-by-loan level before aggregating to pool level. Moody's applies an actuarial approach to calculate the loss distribution in most ABS transactions. It analyzes historical data on pools of loans provided by the originators to calculate the expected loss and the volatility of the losses for the underlying asset portfolio. Once expected losses and volatility have been estimated, a lognormal distribution is used to approximate the loss distribution of the portfolio [13]. Standard and Poor's also estimates the default probability of ABS portfolio using historical default rates provided by the originator. Standard and Poor's estimates the probability of default of an RMBS portfolio at a country level [12]. More specifically, the agency estimates the defaults in a portfolio and the loss severity in a portfolio by calculating the weighted average foreclosure frequency and the weighed average loss severity. These estimates increase for the higher stressed ratings scenarios.

The loss characteristics of the asset portfolio along with the recovery rate and interest rate movement assumptions are the main inputs in the model, which generates the cash flows of the assets. In this second step, cash flows of the underlying assets are generated for different stress scenarios, which correspond to different ratings. Note that for many synthetic CDOs usually there is no need to measure any impact on the credit enhancement levels *via* the cash flow model. For these CDOs, the credit enhancement levels can be directly determined from the loss distribution of the underlying credit portfolio [14].

Liabilities Side Modeling

The asset cash flows and the default assumptions that have been determined in the first two steps of the rating process are the key inputs to the liabilities cash flow model. The purpose of the cash flow model is to determine whether the various tranches of the liabilities receive the principal and interest payments in accordance with the terms of the transaction. In order to achieve this aim, the cash flow modeling usually takes into account the following noncredit risk factors [1, 19], which are known to affect the performance of the transaction:

1. *Capital structure of the transaction, principal, and interest waterfall triggers*, for example, the mechanics through which the assets cash flows are allocated to pay the tranches and all the other transaction fees and expenses.
2. *Market risks* such as prepayment, interest rate, currency, and potential basis risks.
3. *Operational and administrative risks*, that is, the performance risks which are related to the participants in the transaction such as asset managers and servicers.
4. *Counterparty risk*, that is, the performance risks of credit enhancement providers and hedge counterparties.
5. *Legal and regulatory risks* that may rise from imperfect isolation of the securitized assets from the bankruptcy risk of the special purpose vehicle (SPV) under the applicable legal and regulatory framework.

The cash flow modeling procedures differ across the asset classes. Also, cash flow modeling is based heavily on the distinct characteristics of each specific transaction. Thus, in many instances, agencies apply deal-specific cash flow modeling. Usually, the ABS structures are more complicated than those of CDOs and thus cash flow modeling is, in general, the most important part of the ABS rating process. For more technical details on cash flow modeling, refer to [17].

References

- [1] BIS (2005). *The Role of Ratings in Structured Finance Issues and Implications*, Committee on the Global Financial System.
- [2] Bund, S., Dyke, H., Hölter, M., Weintraub, H., Akhavein, J., Ramachandran, B. & Cipolla, A. (2006). *European Consumer ABS Rating Criteria*, Fitch Ratings, 11 October.
- [3] Cifuentes, A. & O'Connor, G. (1996). *The Binomial Expansion Method Applied to CBO/CLO Analysis*. Moody's Special Report, 13 December.
- [4] Dominedo, G. & Currie, A. (2007). *Criteria for European CMBS Analysis*, Fitch Ratings, 12 September.
- [5] Fabozzi, F. & Choudry, M. (2004). *The Handbook of European Structured Financial Products*, Fabozzi Series, Wiley Finance.
- [6] Fabozzi, F., Davis, H. & Choudry, M. (2006). *Introduction to Structured Finance*, Frank J. Fabozzi Series, Wiley.
- [7] Fender, I. & Kiff, J. (2004). *CDO Rating Methodology: Some Thoughts on Model Risk and its Implications*. BIS Working paper, November.
- [8] Garcia, J., Dewyspelaere, T., Langendries, R., Leonard, L. & Van Gestel, T. (2003). *On Rating Cash Flow CDO's using BET Technique*, Dexia Group. Working paper version April 17th.
- [9] Garcia, J., Dwyspelaere, T., Leonard, L., Alderweireld, T. & Van Gestel, T. (2005). *Comparing BET and Copulas for Cash Flows CDO's*, Dexia Group, available in www.defaultrisk.com January 31.
- [10] Gikes, K., Jobst, N. & Watson, B. (2005). *CDO Evaluator Version 3.0: Technical Document*, 19-December-2005.
- [11] Jebjerg, L., Carter, J., Linden, A., Hrvatin, R., Zelter, J., Cunningham, T., Carroll, D. & Hardee, R. (2006). *Global Rating Criteria for Portfolio Credit Derivatives (Synthetic CDOs)*. FitchRatings, 11 October.
- [12] Johnstone, V., McCabe, S. & Kane, B. (2003). *Cash Flow Criteria for RMBS Transactions*. 20-November-2003.
- [13] Kanthan, K., DiRienzo, M., Eisbruck, J., Stesney, L., Weill, N. & Hemmerling, B. (2007). *Moody's Approach to Rating US Auto Loan-Backed Securities*. Moody's, 29 June.
- [14] Koo, M., Cromartie, J. & Vedova, P. (2006). *Global Rating Criteria for Collateralised Debt Obligations*, FitchRatings. 18 October 2006.
- [15] Marjolin, B., Lassalvy, L. & Sieler, J. (2007). *Moody's Approach to Modelling Exotic Synthetic CDOs with CDOROM*. Moody's, 1 February.
- [16] Merton, R. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.
- [17] Tick, E. (2007). *Structure Finance Modeling with Object-oriented VBA*, Wiley Finance.
- [18] Wilgen, A. (2006). *Global Rating Criteria for Cash Flow Collateralised Debt Obligations*, FitchRatings. October.
- [19] Wong, C., Gills, T. & Michaux, F. (2007). *Criteria: Principles-based Rating Methodology for Global Structured Finance Securities*, Standard & Poor's. 29-May.
- [20] Xie, M. & Witt, G. (2002). *Global Cash Flow and Synthetic CDO Criteria*, Standard & Poor's. 21-March.
- [21] Xie, M. & Witt, G. (2005). *Moody's Modeling Approach to Rating Structured Finance Cash Flow CDO Transactions*. 26 September.

Related Articles

ABS Indices; CDO Square; Collateralized Debt Obligations (CDO); Credit Risk; Rating Transition Matrices; Securitization.

AHMET E. KOCAGIL & VASILEIOS
PAPATHEODOROU

Nested Simulation

Stochastic default-intensity (“reduced-form”) models are widely applied in the pricing of single-name credit instruments such as credit default swap (CDS) and corporate bonds (*see Duffie–Singleton Model; Intensity-based Credit Risk Models*). Duffie and Gârleanu [5] demonstrate how collateralized debt obligations (CDO) tranches may be priced using a multiname extension of the stochastic intensity framework (*see Multiname Reduced Form Models*). This article remains influential as a conceptual benchmark, but practitioners generally find the computational burden of this model prohibitive for real-time trading.^a

Risk-management applications introduce additional challenges. Time constraints are less pressing than in trading applications, but the computational task may appear more formidable. When loss is measured on a mark-to-market basis, estimation *via* simulation of VaR and other risk measures calls for a nested procedure: In the outer step, one draws realizations of all default intensities up to the horizon, and in the inner step one uses simulation to reprice each instrument in the portfolio at the horizon conditional on the realized intensities. At first glance, simulation-based pricing algorithms would seem to be impractical in the inner step, because the inner pricing simulation must be executed for each trial in the outer step. This intuition is misleading because a relatively small number of trials in the inner step can suffice, particularly when the portfolio contains a large number of positions.

Model Framework

For simplicity in exposition, we consider a portfolio of K unfunded, synthetic CDO tranches.^b The reference names in the CDO collateral pools are drawn from a universe consisting of m obligors. Each CDO tranche is defined by a set of variables $(A_t, L_t, \Gamma, \Delta, s, T)$ where

- A_t is a vector $A_t = (a_{1t}, \dots, a_{mt})$ of exposures in the CDO pool to each name in the underlying universe, expressed in currency units. Exposure is zero for names not included in a pool.

- L_t is the cumulative default loss since origination as of time t .
- Γ and Δ are the original attachment and detachment points, respectively, for a tranche. These too are expressed in currency units. The residual face value of the tranche at time t is $F_t \equiv \max\{\Delta - L_t, 0\} - \max\{\Gamma - L_t, 0\}$.
- s is the spread on the tranche. We assume that the CDO issuer pays the tranche holder a continuous stochastic premium of sF_t .
- T is the maturity of the CDO.

Let λ_{jt} denote the stochastic default intensity for obligor j . The vector of default intensities is denoted by Λ_t . In models of multiname derivatives, such as CDOs and basket-default swaps, cross-sectional dependence in Λ_t is a central concern. For now, we simply assume that the joint process for Λ_t is specified under risk-neutral and physical measures and defaults are independent conditional on Λ_t (i.e., we rule out contagion and “frailty” in the sense of what is discussed in [6]). Conditional independence implies that default event risk is diversifiable, and so it should attract no risk premium in a large and efficient market. In this case, the risk-neutral intensity $\tilde{\Lambda}_t$ equals the empirical intensity Λ_t at each moment t , even though the two processes evolve into the future under different laws [12].

To keep the focus on credit risk, we assume that risk-free interest rates are constant at r . In this case, the price of position k at time t is a memory-less function of $(\Lambda(t), A_k(t), L_k(t), \Gamma_k, \Delta_k, s_k, T_k - t)$.

Simulation of Value-at-Risk

We now develop notation related to the simulation process. The simulation is nested: There is an “outer step” in which we draw histories up to the horizon H . For each trial in the outer step, there is an “inner step” simulation needed for repricing at the horizon. Loss is measured on a mark-to-market basis, inclusive of interim cash flows. We normalize the present time to zero and the model horizon is H .

Let M be the number of trials in the outer step. In each of these trials, we perform the following:

1. Draw a path for $\Lambda(t)$ for $t = (0, H]$ under the physical measure.
2. Conditional on the $\{\lambda_j(t)\}$, draw default times τ_j up to H and sort them in increasing order. For

each default in $(0, H]$, the following steps are followed:

- (a) Draw a recovery rate R_j for defaulted obligor j .
- (b) For each CDO with exposure to j , increment cumulative loss $L_k(t)$ at $t = \tau_j$ by $(1 - R_j)a_{k,j}(\tau_j)$ and decrement residual face value $F_k(\tau_j)$ accordingly. Record the accrued value at H of the default-leg payment, that is, $e^{r(H-\tau_j)}dF_k(\tau_j)$.
- (c) Adjust exposure vector $A_k(t)$ for termination of exposure to the defaulted obligor.

The information set generated by the outer step trial, denoted by ξ , consists of $\{\tau_j, R_j\}$ for obligors defaulting before time H and $\{\lambda_j(H)\}$ for the survivors.

3. Evaluate the accrued value at H of the premium-leg cash flows.^c
4. Evaluate the price of each position at H with N “inner step” trials for each position. Paths for $\Lambda(t)$ for $t = (H, T_k]$ are simulated under the risk-neutral measure.
5. Discount prices and cash flows back to time zero and subtract from current prices to get portfolio loss $Y(\xi)$.

Observe that the full dependence structure across the portfolio is captured in the period up to the model horizon. Inner step simulations, in contrast, are run independently across positions. This is because the value of position k at time H is simply a conditional expectation (given ξ_H and under the risk-neutral measure) of its own subsequent cash flows, and does not depend on the future cash flows of other positions. Independent repricing implies that pricing errors are independent across the positions, and so they tend to diversify away at the portfolio level. Furthermore, when the positions are priced sequentially, to price tranche k we need to only draw joint paths for the obligors in the collateral pool for CDO k . This may greatly reduce the memory footprint of the simulation.

We now consider the problem of efficient estimation of VaR for Y . For a target insolvency probability α , VaR is the value y_α given by

$$y_\alpha = \text{VaR}_\alpha[Y] = \inf\{y : P(Y \leq y) \geq 1 - \alpha\} \quad (1)$$

Under mild regularity conditions, Y is a continuous random variable so that $P(Y \geq y_\alpha) = \alpha$.

If we had analytical pricing formulae, then $Y(\xi)$ would be a nonstochastic function of ξ , and simulation would involve generating i.i.d. samples $Y(\xi_1), Y(\xi_2), \dots, Y(\xi_M)$. We would sort these draws as $Y_{[1]} \geq \dots \geq Y_{[M]}$, so that $Y_{[\alpha M]}$ provides an estimate of $y_{[\alpha]}$, where $[\alpha]$ denotes the integer ceiling of the real number α .

In the absence of an analytical pricing formula, $Y(\xi)$ is replaced by a noisy estimate $\tilde{Y}(\xi)$, which is obtained *via* the inner step simulations. In place of $Y_{[\alpha M]}$, we have the empirical quantile $\tilde{Y}_{[\alpha M]}$ as our estimate of y_α . Our interest is in characterizing the mean square error (MSE) $E(\tilde{Y}_{[\alpha M]} - y_\alpha)^2$, and then minimizing it. We decompose MSE into variance and squared bias:

$$E[(\tilde{Y}_{[\alpha M]} - y_\alpha)^2] = V[\tilde{Y}_{[\alpha M]}] + E[\tilde{Y}_{[\alpha M]} - y_\alpha]^2 \quad (2)$$

The variance is proportional to $1/M$, whereas the bias vanishes with $1/N$. It can be shown [11, 16] that

$$E[\tilde{Y}_{[\alpha M]}] - y_\alpha \approx \frac{\theta_\alpha}{Nf(y_\alpha)} \quad (3)$$

where

$$\theta_\alpha = \frac{-1}{2} \frac{d}{du} f(u) E[\sigma_\xi^2 | Y = u] \Big|_{u=y_\alpha} \quad (4)$$

and where σ_ξ^2 denotes the conditional variance of the mean-zero pricing error $(\tilde{Y}(\xi) - Y(\xi))$ (conditioned on ξ) and f is the density of Y . A result parallel to equation (4) appears in the literature on “granularity adjustment” of credit VaR to adjust asymptotic approximations of VaR for undiversified idiosyncratic risk [10, 14]. Except in contrived pathological cases, we will generally find $\theta_\alpha > 0$, so that $\tilde{Y}_{[\alpha M]}$ is biased upward as an estimate of VaR.

We suppose that the overall computational budget is fixed at B , and choose (N, M) to minimize the MSE of the estimator $\tilde{Y}_{[\alpha M]}$ subject to the budget constraint. Letting $B \rightarrow \infty$, we find that N^* grows in proportion to $(\theta_\alpha^2 B)^{1/3}$ and M^* grows in proportion to $(B/\theta_\alpha)^{2/3}$. That is, for large computational budgets, M^* grows with the square of N^* . Thus, marginal increments to B are allocated mainly to the outer step. It is easy to intuitively see the imbalance between N^* and M^* . Note that when N and M are of the same order $\sqrt{B}$, the squared bias term contributes much less to the MSE compared to the variance term. By increasing M at the expense of N , we reduce the variance till it matches up in contribution to the squared bias term.

As we increase the number of positions K , the conditional variance σ_ξ^2 falls, and θ_α falls proportionately. If the computation budget grows in proportion to K , then the optimal N^* falls to one, for large-enough K [11]. The intuition is that idiosyncratic pricing error is diversified away at the portfolio level, so a single inner step trial suffices.

Importance Sampling

By recentring the outer step simulation on the region of the state space in which large losses are more common, importance sampling (see **Variance Reduction**) can lead to orders of magnitude improvement in performance. Importance sampling has been applied to structural models of portfolio credit risk [2, 9], but the existing literature does not yet offer a well-developed importance sampling theory for reduced-form models.^d Bassamboo and Jain [1] did some initial work in this direction. They consider the case where the intensity Λ_t is an affine process. Observing that the intensity remains affine under a constant exponential twist, they developed an asymptotically efficient importance sampling change-of-measure to estimate the probability of a large number of defaults in a portfolio.

We now discuss two specifications for correlated intensity processes. Duffie and Gârleanu [5] consider the case where each intensity is the sum of common and idiosyncratic affine processes, that is,

$$\lambda_{i,t} = \lambda_t^* + \zeta_{i,t} \quad (5)$$

Here λ_t^* and $\zeta_{i,t}$ for each i are mutually independent nonnegative affine processes. Under suitable parameter restrictions, the resultant process $\lambda_{i,t}$ is also a nonnegative affine process. For moderately large pools, one expects that large pool losses are most likely to occur when the compensator of the common intensity process, that is, $\int_0^H \lambda_t^* dt$, is unusually large. When an exponential twist of $\theta > 0$ is applied to this variable, the processes remain affine under the new measure and the likelihood ratio has a straightforward representation using results from [7]. If we wish to estimate the α quantile of the overall loss distribution, we choose θ so that the mean of $\int_0^H \lambda_t^* dt$ under the new measure equals its α quantile under the original measure. The latter is unknown at the beginning of the simulation, but may be adaptively learnt as the simulation proceeds.

The Black–Karasinski specification is increasingly popular for modeling single-name default intensities [15] and is readily extended to multiname models. Let

$$d \log \lambda_i(t) = \kappa_i(\mu_i - \log \lambda_i(t)) dt + \sigma_i dW_i(t) \quad (6)$$

where κ_i , μ_i , and σ_i are constants. The diffusion $W_i(t)$ is decomposed into common and idiosyncratic components

$$W_i(t) = \sqrt{\rho_i} W^*(t) + \sqrt{1 - \rho_i} V_i(t) \quad (7)$$

where $W^*(t)$ and $V_i(t)$ for each i are mutually independent standard Brownian motions and ρ_i is a constant. An importance sampling scheme that alters the drift of $W^*(t)$ is straightforward to implement.

Discussion

These methods for nested simulation apply in a much more general setting. We could have structural or ratings-based pricing models in place of the stochastic intensity models, which could introduce stochastic interest rates, and allow for long and short positions. Indeed, we can allow for general derivative portfolios (not just portfolios of CDOs), in which case the intensities $\Lambda(t)$ are replaced by a vector of state variables that could include interest rates, commodity prices, equity prices, and so on. The asymptotic allocation of workload between outer and inner steps remains unchanged. Furthermore, the same analysis applies to estimation of expected shortfall and large loss probabilities, to pricing of compound options, and to the credit rating of structured products under parameter uncertainty.

Optimal allocation schemes and important sampling can be combined and extended in a variety of ways. First, for senior tranches of CDOs in particular, importance sampling within the inner step pricing (as well as the outer step) may offer large efficiency gains. Second, a *jackknife* procedure in the inner step simulation can be used to eliminate the $1/N$ order term in the bias [11]. Third, a scheme for *dynamic allocation* [11] can easily be implemented. An initial estimate of loss $\tilde{Y}^n(\xi)$ is obtained for a given ξ from a small inner step sample of n trials. If $\tilde{Y}^n(\xi)$ is large (and therefore in the neighborhood of VaR), the estimate is refined through drawing additional samples of the inner step. Similar to this is a *screening and restarting* scheme [3, 13] developed for expected

shortfall and other coherent risk measures. In the first phase, an initial sample of M outer step trials is obtained using a small number of inner step samples. These samples are screened or filtered to pick out the large loss samples that are likely to contribute to the expected shortfall. In the second phase, for the short-listed samples, inner steps are resampled to improve the statistical properties of the resultant estimator.

These various refinements can alter the trade-off between bias and variance, so that optimal N^* may grow at a slower or faster rate with the budget. Nonetheless, the essential lesson of the analysis is robust: in a large portfolio VaR setting, inner step pricing simulations can be run with few trials. Despite the high likelihood of grotesque pricing errors at the instrument level, the impact on estimated VaR is small and can be controlled.

Acknowledgments

The opinions expressed here are our own, and do not reflect the views of the Board of Governors or its staff.

End Notes

^aEckner [8] develops a semianalytic algorithm for [5] under somewhat restrictive assumptions.

^bObserve that a single-name CDS can be represented as a special case of such a tranche. With some additional notation, it would be straightforward to accommodate corporate bonds, cash flow CDOs, and other credit instruments.

^cThe implicit assumption here is that interim cash flows are reinvested in the money market until time H , but other conventions are easily accommodated.

^dChiang *et al.* [4] develop an efficient importance sampling technique for pricing basket-default swaps in a Gaussian copula model of times to default.

References

- [1] Bassamboo, A. & Jain, S. (2006). Efficient importance sampling for reduced form models in credit risk, in L.F. Perrone, B.G. Lawson, J. Liu & F.P. Wieland, eds, *Proceedings of the 2006 Winter Simulation Conference*, IEEE Press, Piscataway, NJ, pp. 741–748.
- [2] Bassamboo, A., Juneja, S. & Zeevi, A. (2008). Portfolio credit risk with extremal dependence: asymptotic analysis and efficient simulation, *Operations Research* **56**(3), 593–606.
- [3] Boesel, J., Nelson, B.L. & Kim, S.-H. (2003). Using ranking and selection to “clean up” after simulation optimization, *Operations Research* **51**(5), 814–825.
- [4] Chiang, M.-H., Yueh, M.-L. & Hsieh, M.-H. (2007). An efficient algorithm for basket default swap valuation, *Journal of Derivatives* Winter, 8–19.
- [5] Duffie, D. & Gârleanu, N. (2001). Risk and valuation of collateralized debt obligations, *Financial Analysts Journal* **57**(1), 41–59.
- [6] Duffie, D., Eckner, A., Horel, G. & Saita, L. (2009). Frailty correlated default, *Journal of Finance* **64**(5), 2087–2122.
- [7] Duffie, D., Pan, J. & Singleton, K.J. (2000). Transform analysis and asset pricing for affine jump diffusions, *Econometrica* **68**, 1343–1376.
- [8] Eckner, A. (2009). Computational techniques for basic affine models of portfolio credit risk, *Journal of Computational Finance* **13**(1), 1–35.
- [9] Glasserman, P. & Li, J. (2005). Importance sampling for portfolio credit risk, *Management Science* **51**(11), 1643–1656.
- [10] Gordy, M.B. (2004). Granularity adjustment in portfolio credit risk measurement, in G.P. Szego, ed., *Risk Measures for the 21st Century*, John Wiley & Sons.
- [11] Gordy, M.B. & Juneja, S. (2008). *Nested Simulation in Portfolio Risk Measurement*. FEDS 2008-21, Federal Reserve Board, April 2008.
- [12] Lando, D., Jarrow, R.A. & Yu, F. (2005). Default risk and diversification: theory and empirical implications, *Mathematical Finance* **15**(1), 1–26.
- [13] Lesnevski, V., Nelson, B.L. & Staum, J. (2008). An adaptive procedure for estimating coherent risk measures based on generalized scenarios, *Journal of Computational Finance* **11**(4), 1–31.
- [14] Martin, R. & Wilde, T. (2002). Unsystematic credit risk, *Risk* **15**(11), 123–128.
- [15] Pan, J. & Singleton, K.J. (2008). Default and recovery implicit in the term structure of sovereign CDS spreads, *Journal of Finance* **63**(5), 2345–2384.
- [16] Shing-Hoi, L. (1998). *Monte Carlo Computation of Conditional Expectation Quantiles*. PhD thesis, Stanford University.

Further Reading

Gordy, M.B. & Juneja, S. (2006). Efficient simulation for risk measurement in a portfolio of CDOs, in L.F. Perrone, B.G. Lawson, J. Liu & F.P. Wieland, eds, *Proceedings of the 2006 Winter Simulation Conference*, IEEE Press, Piscataway, NJ.

Related Articles

Credit Portfolio Simulation; Credit Risk; Large Pool Approximations; Monte Carlo Simulation; Multiname Reduced Form Models; Value-at-Risk; Variance Reduction.

MICHAEL B. GORDY & SANDEEP JUNEJA

CDO Tranches: Impact on Economic Capital

With the recent rapid growth in the collateralization debt obligation (CDO) market since its inception in the mid-1990s, many financial institutions, in particular banks, are holding CDO tranches of many different transactions in their credit portfolios. Thus, evaluation of the impact of CDO tranches on the economic capital of a credit portfolio is becoming increasingly important. This article discusses how to measure the impact of CDO tranches on economic capital and capital allocation in credit portfolios. The economic capital of a credit portfolio is the amount of capital reserved to pay for any unexpected losses up to a confidence level as required by the financial institution. The purpose of economic capital is to buffer the effect of large losses in the portfolio. Capital allocation measures the incremental economic capital requirement of a portfolio as a result of adding an asset, such as a CDO tranche. Economic capital and capital allocation are calculated from the probability distribution of portfolio losses. Monte Carlo simulation has been widely applied to calculate the probability distribution of portfolio loss in credit portfolios comprised of bonds and loans. However, it is only recently that Monte Carlo simulations are being used in credit portfolios that comprise CDO tranches in addition to bonds and loans. The methodology presented here addresses this important application in a consistent framework.

The article is organized as follows. In the section Monte Carlo Simulation of Credit Portfolios Comprised of Bonds, Loans, and Collateralization Debt Obligation (CDO) Tranches, a brief introduction is given to the methodology for calculating portfolio loss distribution of credit portfolios comprised of bonds, loans, and CDO tranches. The section Economic Capital and Capital Allocation discusses the calculation of economic capital and capital allocation from the portfolio loss distribution. In the section An Example of Calculating Portfolio Loss Distribution, an example is presented for calculating the probability distribution of portfolio loss for a portfolio comprised of loans and one synthetic CDO tranche. From the loss distribution, economic allocation to the tranche is calculated and compared with the capital

allocation of a loan of the same maturity and similar credit quality.

Monte Carlo Simulation of Credit Portfolios Comprised of Bonds, Loans, and Collateralization Debt Obligation (CDO) Tranches

In credit portfolio management, Monte Carlo methods are the industry standard approach for calculating risk figures from a portfolio loss distribution. Even though the synthetic index-based market has seen considerable development of analytic and semianalytic approaches, these methods do not fully apply to credit portfolio management. This is mainly a consequence of the heterogeneity of underlying assets and the need for default indicators at individual asset levels.

Following standard industry convention, we use a bottom-up approach to model correlated defaults in a pool of names by applying an asset value factor model first introduced by Merton [1] and Vasicek [4]. The calculation of correlated defaults in a pool composed of loans and CDO tranches consists of two basic steps. The first step is to determine the default or credit migration of bonds and loans in the portfolio together with the underlying assets of CDO tranches over a horizon. The second step is valuing the CDO tranches at horizon. For a CDO tranche, its value would be the sum of the cash flows received over the horizon plus a forward value calculated on the basis of the future credit state of its underlying assets at horizon. Cash flows received by a synthetic tranche over the horizon would consist mainly of interest payments, while those received by a cash CDO tranche would consist of interest payments and principal repayments. The loss of a tranche over a horizon is then calculated as the difference between the tranche value at horizon and its current value. This yields a loss distribution at horizon for a credit portfolio comprised of bonds, loans, and CDO tranches, from which its economic capital and capital allocation are calculated.

Although the methodology discussed here for the calculation of economic capital in a credit portfolio is conceptually simple, it is highly complex in practice because it requires nested Monte Carlo simulations. A nested Monte Carlo simulation consists of an outer simulation and an inner simulation at every scenario

of the outer simulation. In the outer simulation, systematic and idiosyncratic risk factors are drawn

within the interval defined by its lower edge, $k_j^{r, \text{lower}}$, and upper edge, $k_j^{r, \text{upper}}$, as follows:

$$\left[k_j^{r, \text{lower}}, k_j^{r, \text{upper}} \right] = \left[\Phi^{-1} \left(PD_j + \sum_k^{r-1} T_{j,k} \right), \Phi^{-1} \left(PD_j + \sum_k^r T_{j,k} \right) \right] \quad (2)$$

at each scenario over a horizon to calculate the defaults and new credit states of nondefaulted assets. The purpose of the inner simulation is to value a CDO tranche conditional on the credit states of its underlying assets at horizon. Valuation of the bonds and loans based on the future credit states of their obligors is also required at a horizon, but their valuation generally does not require a simulation. Even though development of efficient techniques to perform nested Monte Carlo simulations is an interesting and important area of research, a more thorough discussion would be well beyond the scope of the present article.

For the purpose of illustrating the Monte Carlo simulation methodology, we discuss an example of calculating the probability distribution of portfolio loss for a credit portfolio comprised of N loans and one CDO tranche, T_r , over a horizon of 1 year. Furthermore, we assume that the CDO is backed by these loans of the portfolio.

To determine the defaults and credit migrations of the loans at horizon, we calculate their asset value correlation with a single-factor model. The asset value and credit state of a loan in this article refer to those of its obligor. In the single-factor model framework, the sensitivity of the asset value of a loan, X_i , to the systematic risk factor is given in terms of the correlation parameter ρ_i as follows:

$$X_i = \rho_i Y + \sqrt{1 - \rho_i^2} Z_i \quad (1)$$

where Y and Z_i are the systematic risk factor and the idiosyncratic risk factor of the loan, respectively. Both Y and Z_i are independent and have standard normal distribution.

Furthermore, we assume that the calculation of the default and credit migration of a loan at horizon is based on the change of its asset value as given by equation (1). Specifically, a loan is defaulted if its asset value falls below its default threshold, which is defined by its default probability to horizon. Analogously, a loan i initially in the state j is migrated to a state r if its asset value at horizon falls

where $\Phi^{-1}[\bullet]$ is the inverse of the standard normal distribution and $T_{j,k}$ is the probability for the transition from state j to state k . The calculation of the defaults and credit migrations of the portfolio's loans from this first step would be used to determine the loss of the portfolio over the horizon.

The contribution to the loss of portfolio value consists of the loss of the loans defaulted over the horizon and the loss in value of the loans which are not defaulted and the CDO tranche. The total loss of the portfolio at horizon is calculated as follows:

$$L_M = \sum_N \omega_i \times I_i^D \times LGD_i + \sum_{N, T_r} \omega_i \times (1 - I_i^D) \times (PV_{i, t_0} \times DF_{t_0 \rightarrow H} - PV_{i, H}) \quad (3)$$

where ω_i is the weight of the notional value of the loan relative to the total notional value of the portfolio. LGD_i is the fixed percentage loss of the notional value of the loan at default. I_i^D is the default indicator of the loan, which is defined as $I_i^D = 1$, if $X_i < k_i^D$ and $I_i^D = 0$, if $X_i \geq k_i^D$. $PV_{i, H}$ and PV_{i, t_0} are the values of asset i at H and t_0 , respectively, and $DF_{t_0 \rightarrow H}$ is the discount factor from t_0 to H . The first sum of equation (3) consists of the loss of the loans defaulted over the horizon and the second sum consists of the loss in value of the loans which are not defaulted and the CDO tranche.

Valuation of a loan at the horizon assumes that the future credit state of a loan sufficiently determine its value. Similarly, valuation of a CDO tranche at horizon assumes that the future credit states of its underlying assets at horizon sufficiently determine its value although the valuation could require another simulation.

Although the calculation of loss distribution discussed here is based on a single-factor asset value model, extension of the calculation to a multifactor asset model is fairly straightforward. Thus, employing a multifactor asset model in the calculation of loss distribution allows one to apply the methodology discussed in this article to portfolios with heterogeneous asset compositions.

Economic Capital and Capital Allocation

Economic capital of a credit portfolio is defined as the loss exceeding the portfolio expected loss (EL) for the quantiles of the loss distribution to a given confidence level α , that is, $EC_p(\alpha) = q(\alpha) - EL$ with $q(\alpha) = \min\{x : P(L < x) \geq \alpha\}$. This confidence level is interpreted as the probability that the credit portfolio of a financial institution would suffer a loss, which will use up the institution's capital. Since capital exhaustion implies institutional failure, the confidence level α is equivalent to the default risk of an institution.

In the calculation of capital allocation for the CDO tranche of the credit portfolio discussed in the previous section, we apply the methodology of expected shortfall. The expected shortfall^a of an asset, either of the loans or the CDO tranche, is defined in terms of the portfolio loss distribution as follows:

$$ES_i = E[L_{i,H} | L_{\text{portfolio},H} > q(\alpha)] - E[L_{i,H}] \quad (4)$$

where $L_{i,H}$ and $L_{\text{portfolio},H}$ are the losses of asset i and the portfolio at H . Then, the capital allocation of the CDO tranche in terms of its expected shortfall ES_{tranche} and the expected shortfalls of the loans in the portfolio is calculated as follows:

$$EC_{\text{tranche}}(\alpha) = EC_p(\alpha) \frac{ES_{\text{tranche}}}{\sum_{N,T_r} ES_i} \quad (5)$$

An Example of Calculating Portfolio Loss Distribution

We discuss an example of calculating the portfolio loss distribution over a horizon of 1 year for a credit portfolio of loans and a CDO tranche.

Data and Modeling Assumptions

A credit portfolio is assumed to be composed of 125 loans and one synthetic CDO tranche. Each loan is a unique entity, which is taken from the entities of the iTraxx Europe S8 index. The iTraxx Europe index is a credit default swap (CDS) index composed of the most liquid 125 CDSs referencing European investment grade credits. Each loan assumes a notional amount of €1 million, a maturity of 1 year, and a fixed recovery rate of 25%. The synthetic CDO is based on

a portfolio of 125 CDSs. Each CDS assumes a 5-year maturity, the same maturity as the CDO. Each CDS references the entity of one of the loans of the credit portfolio. Therefore, this example considers a case where there is a strong overlap between the credit portfolio's assets and the underlying assets of the CDO.

The nested Monte Carlo simulation as discussed in the section Monte Carlo Simulation of Credit Portfolios Comprised of Bonds, Loans, and Collateralization Debt Obligation (CDO) Tranches is applied to calculate the contribution of the synthetic CDO tranche to the portfolio loss distribution. At each scenario of the outer simulation, we determine which loan is defaulted and a future credit state for the one which is not defaulted. Then, since the loans mature at horizon, we calculate the loss in the notional value of the portfolio from the defaults of the loans over the horizon. To calculate the loss of the synthetic CDO tranche, we value the tranche at horizon based on the future credit states of its underlying assets and compare the value at horizon to its current value. In this example, the future credit states of the underlying assets are exactly those of the loans of the credit portfolio.

The standard industry practice is to use the risk-neutral default probabilities of the underlying assets based on their future credit states to value the synthetic CDO tranche at horizon. One would calculate a mark-to-market (MTM) value of the tranche by using the risk-neutral default probabilities of the underlying assets calibrated to their market spreads and the correlation parameter calibrated to the pricing of tranches of similar structures.^b The calculation is fairly straightforward once the parameters are determined. However, calculating an MTM value of the tranche at horizon is much more challenging because one must determine the future credit states of its underlying assets and apply them to derive their forward risk-neutral default probabilities.

Instead of calculating the loss of the CDO tranche at horizon from its MTM values, we approximate its loss at each scenario of the outer simulation by the conditional expected loss of the tranche. The conditional expected loss of the tranche at horizon in each scenario of the outer simulation is calculated as the average loss over the scenarios of the inner simulation. We employ the following assumptions in the calculation of conditional expected loss at horizon. First, the future credit state of a loan is

4 CDO Tranches: Impact on Economic Capital

represented by an S&P credit rating. The S&P credit rating of a loan at horizon is calculated from its current S&P credit rating with the empirical S&P transition matrix [2]. Second, the best estimation of the forward default probability of a loan at horizon is the forward default probability derived from the default term structure based on its S&P rating. We are assuming that default is a Markov process, but properly inhomogeneous in time. Lastly, we include only the loss of principal in the calculation of conditional expected loss, thereby neglecting any loss from interest payments to the tranche. In addition, we assume a zero interest rate. Under these assumptions, we use a semianalytical approach [3] to calculate the conditional expected loss of the CDO tranche at horizon instead of performing an inner simulation.

We calculate economic capital for the credit portfolio as the loss of its value exceeding the portfolio expected loss at the confidence level of 99% and calculate capital allocation based on expected shortfall methodology at the confidence level of 95%. We assume a correlation parameter of 31% for the underlying assets. We consider a synthetic CDO tranche with an attachment point of 0.0, 3, or 6% and a corresponding detachment point of 3, 6, or 9%. Table 1 shows the expected loss of a tranche as of today and at horizon and Table 2 shows the economic capital allocation of the tranche at different notional amounts. Expected loss of the tranche as of today is calculated with a 5-year cumulative default probability for the underlying assets based on their current S&P ratings. Expected loss of the tranche at horizon is calculated by averaging the conditional expected loss of the tranche at horizon over the scenarios of the outer simulation of the nested Monte Carlo simulation.

Analysis of Results

Now, we would like to discuss the results for the *EL* of the tranche and its capital allocation as shown in Tables 1 and 2. Table 1 clearly shows that

Table 1 Tranche statistics

	0–3%	3–6%	6–9%
EL today	34.71%	2.61%	0.23%
EL at horizon	38.81%	3.43%	0.33%

Table 2 Economic capital allocation to the CDO tranche

Notional (MM)	Exposure (MM)	0–3%	3–6%	6–9%
10	0.3	28.13%	10.46%	1.36%
20	0.6	32.04%	10.48%	1.37%
30	0.9	35.49%	11.12%	1.38%
40	1.2	38.44%	11.41%	1.39%
50	1.5	40.94%	11.69%	1.39%

the *EL* values of the tranche at horizon are larger than the values as of today. In particular, for the 6–9% tranche, the difference is as much as 50%. However, the difference decreases with decreasing tranche subordination and becomes much smaller for the 0–3% tranche. This should suggest that calculating the correlated defaults of the underlying assets plays an important role in determining these results. The *EL* of the 6–9% tranche is more sensitive to default clusters in the portfolio than the *EL* of the 0–3% tranche.

Calculating the *EL* of a tranche as of today and at horizon requires modeling the assets defaults in the pool over the life of a transaction. The calculation of the *EL* of a tranche as of today is based on the approach of Merton [1]. We calculate the defaults of the underlying assets over a horizon of 5 years. The timing of the defaults is not required because the calculation of the expected loss assumes a zero interest rate. It is a one-step calculation of default since both the systematic and idiosyncratic risk factors are drawn once at each scenario of the Monte Carlo simulation. However, the calculation of the *EL* of the tranche at horizon assumes two steps although the calculation of the defaults in each of these steps is also based on the approach of Merton. It is a two-step calculation of default because both the systematic and idiosyncratic risk factors are drawn independently twice in the outer and inner simulations of the nested Monte Carlo simulation. As a result of calculating the default in a single step in the calculation of *EL* as of today and in two steps in the calculation of *EL* at horizon, one should expect the results to be different because of the difference in capturing the correlated defaults in the portfolio.^c Since the calculation of portfolio loss distribution requires using this two-step approach in calculating defaults, we should adopt this approach in the calculation of *EL* of the tranche as of today and at

Table 3 Loan statistics

	AA–	BBB+	BBB–	BB
EL today	0.30%	1.45%	3.71%	7.49%
EL at horizon	0.33%	1.58%	3.69%	7.49%

horizon. Consequently, both *ELs* would be calculated to have the same value.

The results in Table 3 also show that the *EL* of a 5-year loan at horizon is larger than its *EL* as of today for different S&P ratings although the difference is less than 10%. This is simply the consequence of using empirically measured S&P transition matrix and default probability term structures. However, since in comparing a CDO tranche to a loan of similar credit risk we have opted to measure credit risk by an asset's *EL* at horizon, this kind of discrepancy from using empirically measured transition matrix and default probability term structures should not affect the comparison of the results of this article.

In Table 2, we present the tranche's capital allocations at different notional values. Capital allocation is reported as per unit exposure of the tranche. Since the maximum loss of a tranche depends on its thickness defined as the difference between its detachment point and attachment point, it is more meaningful to report capital per unit of exposure rather than per unit of notional amount. These results clearly show that capital allocation per unit of exposure increases with increasing exposure. However, the increase is much larger for the 0–3% tranche than for the 6–9% tranche. This is because the 6–9% tranche is mostly sensitive to systematic risk, while the 0–3% tranche is sensitive to both systematic and idiosyncratic risks. An asset that is mostly sensitive to systematic risk and an asset in a very granular portfolio should have capital requirements that scale approximately the same extent with the size of exposures. Capital allocation per unit of exposure of an asset in a very granular portfolio is approximately constant independent of the asset's exposure size. Thus, as shown in Table 2, the capital allocation of a senior tranche increases slowly with exposure size as compared to a junior tranche.

Now, we would like to compare the capital allocation of a tranche with that of a loan of similar credit quality and the same maturity. In a separate calculation, we substitute the CDO tranche in the credit

Table 4 Economic capital allocation to a loan substituting the CDO tranche

Exposure (MM)	AA–	BBB+	BBB–	BB
0.3	0.87%	3.88%	5.94%	11.03%
0.6	1.11%	4.49%	6.95%	12.55%
0.9	1.18%	4.85%	7.53%	13.35%
1.2	1.24%	5.03%	7.95%	13.66%
1.5	1.24%	5.10%	8.24%	14.49%

portfolio with a loan of 5 years' maturity and re-run the Monte Carlo simulation to calculate the capital allocation of the substituted loan. Table 4 presents the capital allocations for the substituted loans with different S&P ratings.

To compare the capital allocation of a CDO tranche with that of its loan equivalent, we define a loan equivalent of a CDO tranche as a loan that has the same value of *EL* at horizon as the CDO tranche. We assume that a CDO tranche and its loan equivalent have similar credit risks as of today. For example, the loan with AA–rating of Table 4 is a loan equivalent of the 6–9% tranche. However, for the 3–6% tranche, we cannot find a loan from Table 4 with its *EL* at horizon matching that of the CDO tranche. For the purpose of comparing the capital allocation of the 3–6% tranche to its loan equivalent, we scale the capital allocations of the loan with BBB–rating by a ratio of the tranche's *EL* at horizon over the loan's *EL* at horizon. We assume that the scaled capital allocations are for the loan equivalent of the 3–6% tranche.

In Table 5, we compare the capital allocations of the 3–6% and 6–9% tranches with those of their loan equivalents. The results clearly show that the capital

Table 5 Capital allocations: CDO tranche *versus* loan equivalent

Exposure (MM)	AA–	6–9%	Scaled of BBB–*	3–6%
0.3	0.87%	1.36%	5.52%	10.46%
0.6	1.11%	1.37%	6.46%	10.48%
0.9	1.18%	1.38%	7.00%	11.12%
1.2	1.24%	1.39%	7.39%	11.41%
1.5	1.24%	1.39%	7.66%	11.69%

* Capital allocations calculated by multiplying a constant factor of 0.93 with the capital allocations of the loan with a BBB–rating

allocation of a CDO tranche can be larger than that of its loan equivalent by as much as 80%.

We can understand these results in terms of the increase in the systematic risk of the credit portfolio as a result of adding a CDO tranche or a loan. In general, a credit portfolio with larger systematic risk would require higher economic capital. A CDO transaction backed by a pool of assets is mostly sensitive to systematic risks since its exposures to idiosyncratic risk factors are already significantly reduced through the diversification of the assets in the pool. Therefore, adding CDO tranches to a credit portfolio could significantly increase the portfolio's systematic risk compared to adding a loan equivalent. As a result, one should expect that holding a CDO tranche in a credit portfolio would require extra economic capital compared to holding a loan of similar credit rating and the same maturity.

Conclusion

In this article, we presented the methodology based on nested Monte Carlo simulation to measure the impact of CDO tranches on economic capital and capital allocation of credit portfolios. One of the advantages of using the methodology presented here is in the calculation of the correlated defaults of the assets in a portfolio. Calculation of correlated defaults is important in calculating economic capital for a portfolio especially for those holding CDO tranches because CDO tranches are mostly sensitive to systematic risk. As an example, we applied the methodology of calculating capital allocation to a synthetic CDO tranche in a credit portfolio. This portfolio also comprises of the loans to which the underlying assets of the CDO are referencing. The results of our calculation show that the capital allocation of adding a CDO tranche to a credit portfolio can be much larger than that of adding a loan of similar credit quality and the same maturity. In some cases, the increase in capital allocation is as much as 80%. Our finding clearly suggests that by treating a CDO tranche as a loan equivalent, only a poor approximation of economic capital is obtained. This explains why the methodology presented in this article is necessary to measure the impact of CDO tranches in credit portfolios.

Acknowledgments

I would like to thank David Cao, Michele Freed, Saiyid Islam, and Liming Yang for many useful discussions. I especially want to thank Bill Morokoff and Christoff Goessl for careful reading of the manuscript and providing me with many useful comments. The views expressed in this paper are the author's own and do not necessarily represent those of Market & Investment Banking of the UniCredit Group, Lloyds TSB Group or Moody's KMV. All errors remain my responsibility.

End Notes

^aThe other approach of calculating capital allocation is based on marginal Value-at-Risk contribution. For a recent review of the two approaches of calculating capital allocation, please refer to the following paper: Glasserman, P. (2006). Measuring marginal risk contribution in credit portfolios, *Journal of Computational Finance* **9**, 2.

^bSee Burtshell, X., Gregory, J. & Laurent, L.-P. (2005). *A Comparative Analysis of CDO Pricing Models*, Working paper, BNP-Paribas.

^cIn valuing a synthetic CDO tranche, one can use a single-step default model because the valuation is based on the calibration of asset correlation parameters to the pricing of similar tranches in the market.

References

- [1] Merton, R.C. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.
- [2] Okunev, P. (2005). *A Fast Algorithm for Computing Expected Loan Portfolio Tranche Loss in the Gaussian Factor Model*, LBNL-57676.
- [3] Standard & Poor's (1996). *Credit Week-April*.
- [4] Vasicek, O. (2002). Loan portfolio value, *Risk* **15**, 160–162.

Related Articles

Collateralized Debt Obligations (CDO); Gaussian Copula Model; Monte Carlo Simulation; Value-at-Risk.

YIM T. LEE

Equity–Credit Problem

The equity–credit problem, perhaps the most significant problem of corporate finance, is the problem of linking the value of a firm’s equity with that of its debt. Etymologically, the word “credit” is linked to credence, belief, faith, and confidence. As such it is passive and absolute; the only contingency that may affect it is default. One lends money to another on the faith that he/she will pay back the same. This introduces an asymmetry. The debtor may renege and the creditor will try to put his/her assets or company in bankruptcy. Equity, by contrast, is symmetric. Participating in somebody’s equity is agreeing to share the profits, the losses, and the risks of the enterprise. Its only contingency is what action (not passion) the future has in stock for the shareholders. The word *equity* means “fairness” and the word for *share*, in French, is *action*.

The pricing and hedging of convertible bonds (*see* **Convertible Bonds**) is an example of equity-to-credit problem: the optimal conversion policy intimately fuses the two. Once the hazard rate (*see* **Hazard Rate**) is made explicit alongside the price process of the underlying stock, it is followed by the question what kind of correlation can prevail between the two. Trying to model this correlation is the quantitative side of the equity-to-credit problem.

Derivative pricing models are almost never used normatively, going from theoretical parameter to derivative theoretical value. They are used in reverse, going from derivative *market* price to implied parameters (*see* **Implied Volatility Surface**). From this, it becomes apparent that a single volatility parameter cannot explain the variety of prices of options of different strikes and different maturities (the volatility smile), not mentioning that credit default swaps (CDSs, *see* **Credit Default Swaps**) are also written on the company and traded. The equity-to-credit problem, as it is understood today, is to find a unifying derivative pricing framework where the full volatility surface of equity options and the full term-structure of CDS spreads can be explained.

Corporate events, which may have been lost under the impassive writing of derivative payoffs, thus reemerge as the script, waiting for our examination, that the prices of equity derivatives and the prices of credit derivatives now *jointly* compose. The

volatility skew of out-of-the-money puts, the term structures of volatility and credit spread, all of which are commonly observed on single names, cannot be handled by Black–Scholes–Merton (BSM) or by any of the models that tried to improve on it simply by tweaking its assumptions (local volatility (*see* **Local Volatility Model**), Heston stochastic volatility model (*see* **Heston Model**), etc.). These price structures are best (and most robustly) explained by a *regime-switching model* (*see* **Regime-switching Models**). In this model, the volatility of the underlying share and the hazard rate can suddenly switch between different and very dissimilar regimes, accompanied by substantial jumps of the underlying equity price. Default itself is one such regime, except that it is a regime of a very extreme sort. Takeover and restructuring are other regimes. Poisson processes trigger the various switches, and the regime changes then lend themselves naturally to a reinterpretation in terms of *corporate events* [1]. The equity-to-credit problem, when it is reinterpreted in terms of derivative pricing, imprints back corporate events on market prices.

A Regime-switching Model

There exist K volatility regimes, indexed from 1 to K , and one default regime, indexed by d . We assume that the regimes are observable and known at any time. By convention, regime 1 can be set as the present regime when we use the model, either to value contingent claims or to calibrate the model against their market prices. We let the discrete variable u_t describe the current regime at time t . The dynamics of the regime is driven by a continuous time Markov process with a finite number of states or regimes. The important correlation between the stock price and its volatility, as well as between the stock price and the hazard rate, is captured by a possible jump on the stock price as a regime switch occurs. Here, we chose to model jumps by a very simple fixed percentage size jump; it would be possible to generalize to a density of jumps.

- $\hat{\lambda}_{k \rightarrow l}$ is the (risk neutral) intensity of going from regime k to regime l . This occurs with a percentage jump $y_{k \rightarrow l}$ on the stock price.
- $\hat{\lambda}_{k \rightarrow d}$ is the (risk neutral) intensity of going from regime k to default. The corresponding jump on

the stock price is $y_{k \rightarrow d}$. When $y_{k \rightarrow d} = -1$, the stock goes to zero upon default.

- We let v_k and v_d be the volatility in regime k and in default, respectively. The jumps indexed from $i = K$ to J correspond to stock price jumps inside a regime, which may be the default regime. Both the size and the intensity of the jumps depend on the regime.
- In every regime, we index the jumps in the following way. For i between 1 and $(K - 1)$, the jump corresponds to a regime switch, which is also denoted as $y_{k \rightarrow l}$ if the current regime is k and if the jump goes into regime l . For i between K and J , the jump is a pure stock price jump within a regime. For $i = (J + 1)$, the jump corresponds to default, also denoted as $y_{k \rightarrow d}$ if the current regime is k .
- There is no recovery from the default regime so that $\hat{\lambda}_{d \rightarrow k} = 0$. However, there may be jumps in the stock price within the default regime.

For simplicity, we consider a nonstochastic risk-free term structure, and we denote the deterministic instantaneous forward rate at time t as r_t . We consider the most general setting with both continuous and discrete time dividends, although both are usually not present at the same time. r_t^f is the nonstochastic continuous dividend rate at time t (this notation stems from the foreign exchange market).

Pricing Equations

We consider a general derivative instrument with value $F(t, S_t, u_t)$ at time t when the stock price is S_t and the current regime is u_t . We have the following partial decoupling between the nondefault regimes and the default one.

- The value $F(t, S_t, d)$ in the default regime can be computed on a stand-alone basis, without knowledge of F in the nondefault regimes.
- The value of F in the nondefault regimes gives rise to K -coupled equations, which, in general, involve the value of F in the default regime, unless the derivative is a call and the stock goes to zero upon default.

Stand-alone Default Regime

We start by evaluating the value of the general derivative F in the default regime. It is given by

the following stand-alone equation, which does not depend on the value of F in the nondefault regimes:

$$\begin{aligned} \frac{\partial F}{\partial t} + \frac{1}{2} v_d^2 S_t^2 \frac{\partial^2 F}{\partial S_t^2} + \left(r - r^f - \sum_{j=K}^J \hat{\lambda}_j y_j \right) S_t \frac{\partial F}{\partial S_t} \\ + \sum_{j=K}^J \hat{\lambda}_j (F(t, S_t(1 + y_j), d) - F(t, S_t, d)) = rF \end{aligned} \quad (1)$$

with appropriate boundary conditions. The sums $\sum_{j=K}^J \hat{\lambda}_j y_j$ and $\sum_{j=K}^J \hat{\lambda}_j \Delta F$ correspond to stock price jumps inside the default regime.

Coupled Nondefault Regimes

For every nondefault regime u_t ,

$$\begin{aligned} \frac{\partial F}{\partial t} + \frac{1}{2} v_{u_t}^2 S_t^2 \frac{\partial^2 F}{\partial S_t^2} + \left(r - r^f - \sum_{j=1}^{J+1} \hat{\lambda}_j y_j \right) S_t \frac{\partial F}{\partial S_t} \\ + \sum_{j=K}^J \hat{\lambda}_j (F(t, S_t(1 + y_j), u_t) - F(t, S_t, u_t)) \\ + \sum_{l \neq u_t} \hat{\lambda}_{u_t \rightarrow l} (F(t, S_t(1 + y_{u_t \rightarrow l}), l) \\ - F(t, S_t, u_t)) \\ + \hat{\lambda}_{u_t \rightarrow d} (F(t, S_t(1 + y_{u_t \rightarrow d}), d) \\ - F(t, S_t, u_t)) = rF \end{aligned} \quad (2)$$

again with appropriate boundary conditions. A few remarks concerning this equation are given below:

- We note that in the general case the value of F in the default regime is needed here.
- The sum $\sum_{j=1}^{J+1} \hat{\lambda}_j y_j$ corresponds to a sum on all stock price jumps that may occur from the current regime u_t , both inside the regime and from the regime u_t to another one, including default. Although the notation is not explicit, the terms of this sum depend on the current regime u_t .
- The sum $\sum_{j=K}^J \hat{\lambda}_j \Delta F$ corresponds to the stock price jumps inside the current regime u_t .
- The sum $\sum_{l \neq u_t} \hat{\lambda}_{u_t \rightarrow l} \Delta F$ corresponds to the changes in regime, from the current regime u_t to the other nondefault regimes indexed by l .

- The last term $\hat{\lambda}_{u_t \rightarrow d} \Delta F$ corresponds to a jump from the current regime u_t to the default.

Dividends

In the absence of arbitrage, the derivative must be continuous across the payment of dividends. At a time t_i when a fixed dividend $d_i(S_{t_i-})$ is paid by the company, we have

$$\begin{aligned} F(t_i-, S_{t_i-}, u_{t_i}) &= F(t_i, S_{t_i}, u_{t_i}) \\ &= F(t_i, S_{t_i-} - d_i(S_{t_i-}), u_{t_i}) \end{aligned} \quad (3)$$

where u_{t_i} can be any regime, including default.

The same reasoning applies for proportional dividends. At a time t_j when a proportional dividend $\delta_j S_{t_j-}$ is paid by the company, we have

$$\begin{aligned} F(t_j-, S_{t_j-}, u_{t_j}) &= F(t_j, S_{t_j}, u_{t_j}) \\ &= F(t_j, (1 - \delta_j)S_{t_j-}, u_{t_j}) \end{aligned} \quad (4)$$

where again u_{t_j} can be any regime, including default.

Credit Default Swaps

Definitions

We describe the contingent cash flows of the CDS, when the current time is t and the current nondefault regime is u_t .

- The nominal of the CDS is 1.
- The payment frequency of the premium is Δt , a fraction of a year, which can be 1 for one year, 1/2 for a semester, 1/4 for a quarter, or 1/12 for one month.
- We let $N \geq 1$ be the number of remaining premium payments, assuming no default up to maturity of the CDS.
- The first premium payment occurs at time t_1 , with the constraint that $0 < (t_1 - t) \leq \Delta t$.
- The remaining premium payments occur after t_1 with a constant interval Δt , which means that the dates of premium payments are denoted as $t_i = t_1 + (i - 1)\Delta t$ for i between 1 and N .
- The maturity T of the CDS is the last premium payment, that is, $T = t_N = t_1 + (N - 1)\Delta t$.

- We define $t_0 = t_1 - \Delta t$, which is either the date of the last premium payment before t or the issue date of the CDS.
- S is the premium of the CDS paid until default with the payment frequency Δt after the first payment date t_1 . If default occurs at time τ and the last premium payment was at time t_i , then the buyer of insurance still owes the accrued premium, defined as the linear fraction of the next premium payment, that is, $S(\tau - t_i)/\Delta t$.
- Let R be the recovery, between 0 and 1. Upon default, the party insured receives $(1 - R)$ at the time τ of default, if default occurs before maturity T , and nothing otherwise.
- For a time t and a regime u_t , we let $F^{\text{CDS}}(t, u_t; t_1, \Delta t, N, S, R)$ be the value at time t when the regime is u_t of the CDS with premium S and recovery R , whose sequence of premium payment dates is described by t_1 , the date of the first one, Δt the constant time between two premium payments after the first one, and N the total number of premium payments. Its maturity is $T = t_1 + (N - 1)\Delta t$. This value assumes the point of view of the holder of the CDS. He/she is the party who seeks insurance against default risk and who pays the streams of premia in exchange for a compensation in the case of default. When $u_t = d$, we assume that default occurred at time t exactly, and $F^{\text{CDS}}(t, d; t_1, \Delta t, N, S, R)$ values the benefit of the insurance upon default. At a time t_i of premium payment, we assume that $F^{\text{CDS}}(t_i, u_t; t_1, \Delta t, N, S, R)$ corresponds to the value ex premium payment, that is, just after the payment of the premium S . When $(t_1, \Delta t, N)$, the maturity T , the premium S , and the recovery R are understood, we simply write $F^{\text{CDS}}(t, u_t)$.

Value in Default

Let $F^{\text{CDS}}(t, d)$ be the value of the CDS at time t assuming that default has occurred at time t . It is the difference between the compensation immediately received and the accrued premium still owed:

$$F^{\text{CDS}}(t, d) = 1 - R - S \frac{(t - [t])}{\Delta t} \quad (5)$$

Backward Equation

We derive a backward equation for the value $F^{\text{CDS}}(t, u_t)$ of the CDS. At maturity T and for every nondefault regime k , we have

$$F^{\text{CDS}}(T, k) = 0 \quad (6)$$

$$F^{\text{CDS}}(T-, k) = -S \quad (7)$$

since maturity T corresponds to the last premium payment. At a time t different from any premium payment date and for every nondefault regime k , we have

$$\begin{aligned} \frac{\partial F^{\text{CDS}}}{\partial t}(t, k) + \sum_{l \neq k} \hat{\lambda}_{k \rightarrow l} \\ \times \left(F^{\text{CDS}}(t, l) - F^{\text{CDS}}(t, k) \right) \\ + \hat{\lambda}_{k \rightarrow d} \left(F^{\text{CDS}}(t, d) - F^{\text{CDS}}(t, k) \right) \\ = r F^{\text{CDS}}(t, k) \end{aligned} \quad (8)$$

where the value in default $F^{\text{CDS}}(t, d)$ has been derived in equation (5). At a time $t_i = t_1 + (i - 1)\Delta t$ where the premium S is paid, we have

$$F^{\text{CDS}}(t_i-, k) = F^{\text{CDS}}(t_i, k) - S \quad (9)$$

which means that the value of the CDS increases by S after each premium payment. This yields a system of K -coupled backward equations.

The regime-switching model represents *corporate events* in terms of jumps. The volatility parameters v_k and v_d , the regime-switching intensities $\hat{\lambda}_{k \rightarrow l}$ and $\hat{\lambda}_{k \rightarrow d}$, and the corresponding jump sizes $y_{k \rightarrow l}$ and $y_{k \rightarrow d}$ are all inferred by calibration of the model to the market prices of equity options and CDS spreads. This is achieved by solving an inverse problem characterized by the pricing equations above. For this reason, the regime-switching model is perfectly suited

for pricing the convertible bond (see **Convertible Bonds**).

References

- [1] Ayache, E. (2004). *The Equity-to-Credit Problem. The Best of Wilmott 1, Incorporating the Quantitative Finance Review, 2004*, John Wiley & Sons, Chichester, West Sussex, pp. 79–107.

Further Reading

- Ayache, E., Forsyth, P.A. & Vetzal, K.R. (2002). Next generation models for convertible bonds with credit risk, *Wilmott* December, 68–77.
- Ayache, E., Forsyth, P.A. & Vetzal, K.R. (2003). The valuation of convertible bonds with credit risk, *Journal of Derivatives* **11**, 9–29.
- Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.
- Ferguson, N. (2008). *The Ascent of Money: A Financial History of the World*, Allen Lane, Penguin Books, London, 119–175.
- Goldman, S. (1994). *Valuing Convertible Bonds as Derivatives*. Quantitative Strategies Research Notes, Goldman Sachs.
- Hull, J. (2008). *Options, Futures, and Other Derivatives*, 7th Edition, Prentice-Hall, Englewood Cliffs, New Jersey.
- MacKenzie, D. (2003). An equation and its worlds: bricolage, exemplars, disunity and performativity in financial economics, *Social Studies of Science* **33**, 831–868.
- Smith, C.W. (2007). Markets as definitional practices, *The Canadian Journal of Sociology* **32**(1), 1–39.
- Thorpe, E.O. & Kassouf, S.T. (1967). *Beat the Market*, Random House, New York.
- Tsiveriotis, K. & Fernandes, C. (1998). Valuing convertible bonds with credit risk, *Journal of Fixed Income* **8**, 95–102.
- Wilmott, P. (2006). *Paul Wilmott on Quantitative Finance*, 2nd Edition, John Wiley & Sons, London.

Related Articles

Convertible Bonds; Credit Default Swaps; Hazard Rate; Regime-switching Models; Structural Default Risk Models.

ÉLIE AYACHE

Convertible Bonds

A convertible bond (CB) is a *corporate bond* that may be converted into a certain number of shares of the issuing company, also known as the *underlying equity*. By offering the investor an option to participate in its future growth, the company is able to finance itself at a lower cost. Alternatively, the CB provides a means to issue equity in a deferred fashion when the market context or timing may not be right for the company.

Like any corporate bond, the CB is issued with a nominal, *principal*, or face value, to be redeemed at maturity, sometimes at a premium, and pays interest, typically periodic coupons, fixed or floating. In some cases (in particular, when no coupons are paid), the principal accretes continuously at a rate guaranteeing a given *yield to maturity* (YTM), and the final redemption value is the result of the accretion. The CB can be redeemed earlier than maturity at the option of the issuer (issuer's call), often with penalties in *call protection* periods (make-whole premium, guaranteed yield). Alternatively, the holder can opt out of the convertible and get his/her money back at certain prespecified discrete dates, or *put* dates. Coupons can be fixed or floating. The number of shares against which the CB can be converted is the *conversion ratio*. The value the holder obtains following conversion is the *conversion value*: ratio $\times$ share value. The *conversion price* is the ratio of nominal over conversion ratio. It is the value of the underlying above which conversion value becomes greater than the nominal, thus making the conversion potentially attractive.

The interplay of options that are embedded in the CB (part held by the investor, part by the issuer) translates into a hierarchy of conditions bounding and constraining its value. Given the many ways in which the principal may be redeemed (coupons, redemption at a premium, accretion, mixture of the two) and the fact that the process of redemption may be interrupted prematurely, either by the holder's option to convert (or to put) or by the issuer's call, we first need to define the intermediate notion of the holder's *claim*. This is the amount the investor is entitled to claim at any intervening time between issue date and maturity, should the bond terminate at that time, in order to

secure the same yield on investment as the YTM.

$$\text{claim}(t) = \sum_{T \geq t_i \geq t} \frac{C_i}{(1 + \text{YTM})^{(t_i - t)}} + \frac{\text{redemption}}{(1 + \text{YTM})^{(T - t)}} \quad (1)$$

where C_i is the coupon payment at time t_i ; t_i is a coupon date (in number of years since a reference date); YTM is the annualized yield to maturity guaranteed by the claim; and redemption is the amount redeemed at the maturity date.

$$\text{redemption} = \text{principal (possibly accreted)} + \text{redemption premium} \quad (2)$$

At time $t = 0$, the claim is, by definition, the issue price, which is also the initial amount of money that gets invested, and this defines the annualized yield to maturity, YTM. At maturity, the claim is the redemption amount.

The claim serves as the basis of all three options characterizing the CB. Conditions on the CB value are enforced in the following sequence order.

Whenever the issuer has the right to exercise his/her early call option, he/she will do so optimally as soon as the CB value exceeds the amount he/she would otherwise be contractually bound to redeem, should he/she opt for early redemption (early redemption price). This entails the following constraint:

$$V(S, t) \leq \text{early redemption price} \quad (3)$$

The early redemption price is usually defined as a percentage of the claim at the time or early call. Alternatively, it can be set in such a way as to guarantee a certain yield up to the date of call, which may be different from the yield to maturity.

The holder's conversion option overrides the issuer's call. It is expressed by the following condition:

$$V(S, t) \geq \text{conversion ratio} \times S \quad (4)$$

and applied hierarchically after the issuer's call.

At maturity T , note that the combination of the early redemption option (which is no longer "early" and no longer an option) and the conversion option yields the following final payoff condition:

$$V(S, T) = \max(\text{conversion ratio} \times S, \text{redemption}) \quad (5)$$

Finally, the put condition is the last one to be applied. If the investor finds, at put dates, that the CB value is still lower than the put strike (even after allowing for the option to convert), he/she will exercise his/her put. Thus

$$V(S, t_p) \geq \text{put price} \quad (6)$$

where t_p is a put date.

Similar to the early redemption price, put price is defined as a percentage of the claim or through a guaranteed yield to put.

Options versus Convertible Securities

The convertible bond is a hybrid security belonging to the category of *corporate debt* (as such, subject to credit risk) and to the category of *convertible security*. Convertible securities are claims, issued by a company that can be *converted* into the underlying equity of the company during their lifetime. Warrants belong to this general category. They feature a strike price and are very similar to American call options.

The terminology of warrants differs from that of options for reasons that run deeper than casual jargon. We do not *exercise* a warrant; we *convert* it. The warrant is not the right to *buy* the underlying share; we say it is *convertible* into that share. Indeed, the warrant is issued by the company, whereas the call option is written by an independent counterparty; *option buyers* and *option sellers* exchange their contracts independently, in the over-the-counter (OTC) market or on a listed exchange. To put it bluntly, the difference between convertible securities and options is the whole difference between being an investor in the company and taking an external view (bet) on the evolution of the price of its underlying equity. It is the difference between *corporate events* and *incorporeal events* (which is discussed later).

The number of convertible bonds that a corporation can issue cannot exceed the total number of shares that it would eventually have to issue upon conversion. Issuing convertibles alters its capital structure and dilutes the earnings of the existing shareholders. Options, by contrast, can be written in unlimited amounts, all the more so because the procedure of cash settlement no longer compels the option seller to actually deliver the underlying share.

Dynamics of Abstraction

It is important to emphasize the difference between convertible securities and options, if only for the reason that the first are of earthly nature (they tend to remain and “deposit”), whereas the latter are volatile (they tend to “evaporate”). This material difference has ramifications bordering on ethics. Indeed, options lend themselves easily to the logic, therefore to the charge, of betting and gambling. Not everybody is aware that the fathers of volatility arbitrage, Sheen Kassouf and Ed Thorp, first exercised their skills and “system” in the convertibles area, and not in options [7]. Options had been banned in the aftermath of the 1929 market crash, and it was not until the early 1970s that they regained some liquidity and Black and Scholes were able to study them in order to derive their formula. The Chicago Board Options Exchange played no little role, subsequently, in bringing options back in favor, truly resurrecting their markets. (The belief had caught on that the Black–Scholes–Merton (BSM) formula, and the option trading it involved, were not about speculation and gambling after all, but about *efficient pricing* [6].)

As a matter of fact, the BSM paradigm put all derivatives on an equal footing. It led to the mathematization of their valuation, thus leveling out all their genealogical or ethical differences. From the valuation perspective, everybody now viewed derivatives as *pure payoff structures*, that is to say, as cash flow events that were simply triggered by the price trajectory of the underlying. From then on, the only dynamics that mattered was the price dynamics of the underlying. Find the right stochastic process to model the underlying behavior, and all the derivative pricing problems will be solved. This was the beginning of the *quant* era: the heedless, unstoppable sophistication of valuation models and payoff structures.

This pricing paradigm treated convertible securities, and more specifically convertible bonds, no differently than equity options. The convertible bond was identified with a corporate bond, whose valuation posed no greater difficulty than the rest of fixed income, bundled with an equity option, whose valuation posed no greater difficulty than BSM, or so everybody thought. Owing to the American-styled conversion feature, the suggestion was to price the

CB by dynamic programming techniques, for example, Cox's binomial tree (see **Binomial Tree; American Options**). The procedure is initialized with the terminal payoff condition: either convert into equity or have the principal redeemed. Rolling backward, the cash payments accruing from the fixed-income part are simply added in the nodes at coupon dates and they become an integral part of the value of the CB. The procedure keeps comparing this value to conversion value, as it progresses to the present date, in order to check for early conversion [5, 9].

Credit risk started posing a problem though, when it was observed that the convertible bond had effectively a shorter maturity than the advertised one due to the probability of early conversion. How could the fixed-income component of the CB be valued using the same credit spread as the corporate bond of same official maturity? Is not the CB in effect less subject to credit risk than the pure bond because it is likely to wind up as equity, and this, "of course," bears no credit risk?

In terms of the pricing procedure, this is the question of the rate with which to discount the value of the CB in the tree. Should it be the risk-free rate (as in BSM), the risky rate (risk-free + credit spread), or a mixture of the two? Some have suggested using a weighted average of the two rates, depending on the probability of early conversion [4]. The delta of the CB would be the estimate of this probability. Others have proposed to split the CB into a credit-risky component and a riskless component (as the hybrid nature of the instrument naturally suggests) and the dynamic programming technique would now determine the splitting, because this obviously depends on the optimal conversion policy [8].

Ultimately, the right approach was to abandon the whole idea of patching together two pricing paradigms that had nothing in common: BSM (with dynamic hedging and dynamic programming) and fixed income (with static discounting using credit spreads). Rather, credit risk should be brought to bear right at the heart of the dynamics, and the process of the underlying equity itself revised. The vague notion of credit risk was thus reduced to the definite, probabilistic occurrence of a *default event*.

Default risk was modeled through a Poisson process superimposed on the traditional BSM diffusion. Its intensity, or hazard rate, measured the instantaneous probability of default, conditionally on default

not having occurred earlier. The Poisson process triggers a massive jump both in the underlying equity price and in the derivative price. Typically, the convertible bond would jump to its recovery value, modeled as a fraction of the claim at the time of default. All equity and credit derivatives could now be valued in this reduced-form, unified framework. To that end, the dynamic hedging argument of BSM had to be generalized to cover the jump to default. A second, credit-sensitive, hedging instrument would thus be needed on top of the underlying. There was no need to split the CB any longer, as the insertion of the hazard rate in the pricing equation, combined with the explicit expression of the fate of the CB after default (recovery), mechanically determined the "rate" with which to discount values in the tree or in the PDE (see **Partial Differential Equations**) [2, 3].

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r + \lambda) S \frac{\partial V}{\partial S} = rV + \lambda(V - R \text{ claim}(t)) \quad (7)$$

where σ is the diffusion coefficient, λ is the hazard rate, r is the risk-free interest rate, and R is the recovery ratio.

The Equity-Credit Problem

This revision of the underlying dynamics set up the stage for the *equity-to-credit problem*. As the hazard rate was now recognized as the second factor alongside the diffusion coefficient, it could potentially be made stochastic and correlated with the process of the underlying [1].

Equity-to-credit, or rather, credit-to-equity dynamics is the real dynamics ruling the CB. *Conversion is but a movement from credit to equity*. Debt, passively binding creditor and debtor, is converted into activity, project, enterprise, participation in the upside (see **Equity-Credit Problem**). Recall that conversion is an altogether more consequential transmutation than just exercising a right. The entire convertible can be reread in the light of this difference. The owner of a convertible security is more involved in the company than the buyer of an option. As such, he/she has to be *protected*. *Dividend protection, takeover protection, call protection*, and so on, are among the many clauses that make for the increasingly thicker CB prospectuses nowadays. Dividend protection dates

back to the warrants that Thorp and Kassouf used to trade. It protects the holder of the convertible security against the fall of the underlying share following dividend announcements, by readjustment of the conversion ratio or by a “pass-through” procedure. Takeover protection entitles the holder to sell back the convertible (takeover put) or to convert it against a greater number of shares than initially promised (ratchet clause), when a change of control or takeover is announced. Alternatively, the conversion ratio can be adjusted in proportion with the ratio of the price of the share of the company taking over and of the company being taken over.

As for the events that may trigger a change in the CB structure (early conversion by the holder, early redemption by the issuer, reset of the conversion price in case of resettable CBs, etc.), there is not one of them that is not “slowed down” and “mashed down” by a host of averaging clauses (as if the convertible had truly to *digest* the event). Only if the underlying trades 20 days out of 30 above a certain trigger level is the issuer entitled to call back the bond (soft call), or is the holder entitled to convert it (contingent conversion); the conversion price is reset at the average of closing prices of the underlying over a certain number of days before the reset date, etc.

From the point of view of holders of options (as opposed to convertible securities), all that matters is the price process of the underlying equity, which screens off all the deeper corporate changes. Dividend announcements, takeover announcements, and so on, are among many factors that may otherwise affect the price of the underlying, so why would option holders regard them any differently than the rest of price shocks and jumps that they prepare to face anyway? In their view, events are incorporeal and are

only expressed in terms of underlying price changes: barriers and triggers may very well be breached punctually and without averaging clauses, so long as everybody agrees.

Holders of convertible securities, by contrast, are the *patient* readers of grave and significant corporate events, and the convertible is the book that binds and encodes those events.

References

- [1] Ayache, E. (2004). The equity-to-credit problem, in *The Best of Wilmott 1, Incorporating the Quantitative Finance Review*, 2004, P. Wilmott, ed., John Wiley & Sons, Ltd, Chichester, West Sussex, pp. 79–107.
- [2] Ayache, E., Forsyth, P.A. & Vetzal, K.R. (2002). Next generation models for convertible bonds with credit risk, *Wilmott Magazine* December, 68–77.
- [3] Ayache, E., Forsyth, P.A. & Vetzal, K.R. (2003). The valuation of convertible bonds with credit risk, *Journal of Derivatives* **11**, 9–29.
- [4] Goldman Sachs (1994). Valuing convertible bonds as derivatives, in *Quantitative Strategies Research Notes*, Goldman Sachs.
- [5] Hull, J. (2008). *Options, Futures, and other Derivatives*, 7th edition, Prentice-Hall, Englewood Cliffs, New Jersey.
- [6] MacKenzie, D. (2003). An equation and its worlds: bricolage, exemplars, disunity and performativity in financial economics, *Social Studies of Science* **33**, 831–868.
- [7] Thorp, E.O. & Kassouf, S.T. (1967). *Beat the Market*, Random House, New York.
- [8] Tsiveriotis, K. & Fernandes, C. (1998). Valuing convertible bonds with credit risk, *Journal of Fixed Income* **8**, 95–102.
- [9] Wilmott, P. (2006). *Paul Wilmott on Quantitative Finance*, 2nd edition, John Wiley & Sons, London.

ÉLIE AYACHE

Bond

A *bond*, or a *debt security*, is a financial claim by which the *issuer*, or the borrower, is committed to paying back to the *bond holder* or the lender, the cash amount borrowed, called *principal*, plus periodic coupon interests calculated on this amount during a given period. It can have either a standard or a nonstandard structure. A *standard bond* is a fixed coupon bond without any embedded option, delivering its coupons on periodic dates and principal on the maturity date. Nonstandard bonds include, among others, zero-coupon bonds, floating rate notes, inflation-linked bonds, callable and puttable bonds, and convertible bonds.

An example of a standard bond would be a US treasury bond with coupon interest 4%, maturity date November 15, 2017, and a nominal issued amount of \$20 billion, paying a semiannual interest of \$400 million ($\$20 \text{ billion} \times 4\%/2$) every six months until November 15, 2017 included, as well as \$20 billion on the maturity date.

A bond issuer has a direct access to the market, and so avoids borrowing from investment banks at higher interest rates. A bond holder has the status of a creditor, unlike an equity holder, who has the status of an owner of the issuing corporation. This is the reason why a bond is less risky than an equity.

A bond issue is mainly characterized by the following components [4]:

- *The issuer name*: For example, France for a treasury bond issued in France.
- *The issuer type*: This is mainly the economic sector the issuer belongs to.
- *The issuer domicile*.
- *The issuance market*: The issuance market may differ from the issuer domicile. For example, the Eurodollar market corresponds to bonds denominated in USD and issued in any country other than the United States.
- *The bond currency denomination*.
- *The maturity date*: This is the date on which the principal amount is due.
- *The coupon rate*: It is expressed in percentage of the principal amount.
- *The coupon type*: It can be fixed, floating, or a mix of the two.
- *The coupon frequency*: Most commonly, it is semiannual in the United States, the United Kingdom, and Japan, and annual in the Euro zone, except for Italy, where it is semiannual.
- *The day-count type*: The most common types are actual/actual, actual/365, actual/360, and 30/360. Actual/actual (actual/365, actual/360) means that the accrued interest between two given dates is computed using the exact number of calendar days between the two dates divided by the exact number of calendar days of the ongoing year (365, 360). The term 30/360 means that the number of calendar days between the two dates is computed, assuming each month count as 30 days.
- *The interest accrual date*: This is the date when interest begins to accrue.
- *The settlement date*: This is the date on which payment is due in exchange for the bond. It is equal to the trade date plus a number of working days (generally, one to three, depending on the country).
- *The issuance price*: This is the percentage price paid at issuance.
- *The spread at issuance*: The spread in basis points to the benchmark treasury curve or the swap curve.
- *The rating*: A ranking of a bond's quality and its record in paying interest and principal. The three major rating agencies are Moody's, Standard and Poor's, and Fitch.
- *The outstanding amount*: This is the amount of the issue still outstanding.
- *The par or nominal or principal amount*: The face value of the bond.
- *The redemption value*: Expressed in percentage of the nominal amount, it is the price at which the bond is redeemed on the maturity date.
- *The identifying code*: The most popular ones are the *ISIN* (International Securities Identification Number) and the *CUSIP* (Committee on Uniform Securities Identification Procedures) numbers.

Bond Price and Yield to Maturity

Bonds are usually quoted in price, yield, or spread against an underlying benchmark bond or reference swap rate. The price of a bond is always expressed in percentage of its nominal amount. The *quoted price*

(or *market price*) of a bond is usually its *clean price*, that is, its *gross price* minus the *accrued interest*. When an investor purchases a bond, he is actually entitled to receive all the future cash flows of this bond, until he no longer owns it. If he buys the bond between two coupon payment dates, he logically must pay it at a price reflecting the fraction of the next coupon that the seller of the bond is entitled to receive for having held it until the sale. This price is called the *gross price* (or *dirty price* or *full price*). It is computed as the sum of the clean price and the portion of the coupon that is due to the seller of the bond. This portion is called the *accrued interest*. It is computed from the settlement date on.

The quoted *yield to maturity* of a bond is the discount yield that equalizes its gross price times its nominal amount to the sum of its discounted cash flows. As an illustration, let us consider a French bond paying an annual coupon of 7%, residual maturity 8.5 years and market price 106.459%. The accrued interest is equal to $7\% \times 0.5 = 3.5\%$ (there have been six months since the last coupon payment). The bond annual yield to maturity is such that

$$106.459\% + 3.5\% = \sum_{i=0}^8 \frac{7\%}{(1+R)^{i+0.5}} + \frac{100\%}{(1+R)^{8.5}} \quad (1)$$

which can be solved to yield $R = 6\%$.

The price–yield to maturity relationship is inverse and convex. The inversion property means that the higher (lower) the price, the lower (higher) the yield to maturity. The convexity property means that the previous property is not symmetrical. Actually, for the same absolute change in the yield to maturity, the bond price will increase more than decrease. Convexity is an attractive property that tends to increase with bond maturity.^a A bond price is also dependent on time to maturity. All else being equal, as the bond residual maturity decreases, its price converges to the redemption value. This so-called *pull-to-par phenomenon* is particularly noticeable for bonds with coupons far from their yield to maturity. The price–yield to maturity relationship is given by the following formula [1], where we assume that the

bond has just paid a coupon:

$$P = \sum_{i=1}^{\delta T} \frac{CF_i}{\left(1 + \frac{R}{\delta}\right)^i} \quad (2)$$

Here, P denotes the bond dirty price, T its maturity expressed in years, CF_i its cash flows, R its yield to maturity, δ is 1 for an annual bond, 2 for a semiannual bond, 4 for a quarterly bond, and so forth.

In other words, the yield to maturity is the internal rate of return of the cash flows produced by the bond, using a constant discount rate across all cash flows. The yield to maturity may therefore be interpreted as an “average” discount rate throughout the life of the bond or, equivalently, as the discount rate that would prevail if the yield curve happened to be flat at date t (which of course is not generally the case). It may be easily computed by trial and error or using built-in functions on Excel.

Under certain technical conditions, there exists a one-to-one correspondence between the price and the yield to maturity of a bond. Therefore, giving a yield to maturity for a bond is equivalent to giving a price for the bond. It should be noted that this is precisely what is actually done in the bond market, where bonds are most often quoted in yield to maturity. While bond yield to maturity is a useful concept, it should be approached with some care, given that it represents a weighted average of discount yields across maturities. Indeed, unless the term structure of interest rates is flat, there is no reason why one would consider the yield to maturity on, say, a 10-year bond as the relevant discount rate for a 10-year cash flow horizon. In fact, the relevant discount rate is the 10-year pure discount (or zero coupon) rate.

The yield to maturity, as the name suggests, can be viewed as the expected rate of return on a bond and allows comparison of two bonds of the same issuer with close maturities with each other. Let us consider a 10-year bond and a 10.25-year bond issued by the French treasury, both quoted at par value (i.e., 100%). The former (the latter) yields 4.50% (4.55%). An investor will most likely prefer the second one, as it yields 5 more basis points (0.05%) than the other (a nonnegligible premium) for a maturity that is slightly longer (three months). However, the yield to maturity on an investment will be achieved only under the following constraints:

- Once bought, bond securities have to be held until maturity.
- Bond coupons (equal in our example to the yield to maturity) have to be reinvested at the bond yield to maturity.

The first assumption is restrictive in the sense that it excludes early bond sales. If interest rates go down, an investor will be willing to sell the bond before maturity so as to take advantage of capital gains. This condition is only acceptable for long-term investors (pension funds, insurance companies), who are traditionally buy-and-hold investors. The second assumption is not realistic for two reasons:

- During the life of a bond, interest rates increase and decrease, but never remain unchanged. Consequently, postulating a unique reinvestment rate is not particularly relevant.
- Taking their respective yield to maturity for two bonds as reinvestment rate, that is, assuming that two different cash flows falling on the same date can be reinvested at different interest rates, boils down to questioning the uniqueness of the reinvestment rate at a given time for the same investment horizon [2].

In general, the expected return on a bond nearly always differs from its yield to maturity, except for zero-coupon bonds, which embed no reinvestment risk, as they deliver no intermediary cash flows before maturity. In practice, instead of using the yield to maturity, investors compute the *total rate of return* on a bond, which is equal to the sum of the difference between the sale price and the purchase price and the coupons reinvested at the interest rate corresponding to each reinvestment period, divided by the purchase price. So as to determine the future reinvestment rates as well as the yield at which the bond will be sold, investors often apply various scenarios of evolution of interest rates (worst case, best case, and neutral scenarios)—this is known as *scenario analysis*. To each scenario corresponds a specific total rate of return.

Duration and Convexity

While the concept of yield to maturity has its flaws, it is quite useful in the construction of hedges against interest rate risk. Indeed, as we have seen, the bond

price can generally be written as a unique function of the yield to maturity, so it is sensible to treat the yield to maturity as the main stochastic risk factor driving the price of the bond. If the yield happens to increase, the bond value will decrease, generating (in case of early unwinding) potentially significant losses against which one can choose to be immunized. For this purpose, market participants rely on an interest rate risk measure, whose popularity is inseparable from that of the yield to maturity: duration. This concept, which was developed in 1938 by the American economist Frederick Robertson Macaulay (1882–1970), only became famous in the bond markets from the 1970s onward, a period precisely characterized by a large increase in interest rates and hence a sharp rise in interest rate risk. Macaulay had realized that a bond maturity was an insufficient measure of its effective life because it effectively took only the final cash flow into account. As an alternative to the outright maturity, he suggested that bond interest rate risk be characterized by the average “length” of the bond cash flows—the so-called *Macaulay duration* [3]. These days, a more common measure for interest rate risk is the so-called *modified duration*, which is simply defined as the absolute value of the first derivative of the bond price with respect to its yield, divided by the bond price itself. The Macaulay duration can be computed as the modified duration multiplied by the factor $1 + \frac{R}{\delta}$. A third measure of duration—the *dollar duration*—equals the modified duration times the bond price. Notice that dollar duration measures directly the change in value for a small change in the yield to maturity, whereas modified duration measures percentage change in value.

All duration measures are related to the slope of the bond-yield function at the current yield R , and as such can be interpreted as related to first-order terms of a Taylor expansion of the bond price in its yield. Duration is an appropriate measure of risk for small parallel moves in yield. However, when one wants to quantify the impact of large parallel yield curve moves, it should be accompanied by a second-order measure, called *convexity*, which is the second derivative of the bond price with respect to yield to maturity, divided by the bond price. Formally, we can accomplish this by simply adding another term to the Taylor expansion.

For this, let us denote the bond modified duration by S , its convexity by C , the change in the yield to

maturity by ΔR , and the change in the bond price by ΔP , such that

$$\frac{\Delta P}{P} \sim -S \times \Delta R + 0.5 C \times (\Delta R)^2$$

Duration and convexity, which are based on the flat shape of the term structure of interest rates and on exclusively parallel shifts in interest rates, are faced with two significant hurdles: yield curves are practically never perfectly flat and they are not solely affected by parallel movements.

Interestingly, the Macaulay duration of a bond or bond portfolio is the investment horizon such that investors will not care if interest rates drop or rise as long as changes are small. In other words, capital gain risk is offset by reinvestment risk, as shall be shown now in the following example.

Consider a three-year standard bond with a 5% yield to maturity and a \$100 face value, which delivers a 5% coupon rate. Coupon frequency and compounding frequency are assumed to be annual. The bond price is \$100 and its Macaulay duration is equal to 2.86 years. We assume that the yield to maturity changes instantaneously and stays at this level during the life of the bond. Whatever the change in the yield to maturity, we show in Table 1 that the sum of the bond price and the reinvested coupons after 2.86 years is always the same and equal to 114.972.

The main properties of the three duration measures are as follows:

- The Macaulay duration of a zero-coupon bond equals its time to maturity.
- Holding the maturity and the yield to maturity of a bond constant, the lower a bond coupon rate, the higher its Macaulay or modified or dollar duration.
- Holding the coupon rate and the yield to maturity of a bond constant, its Macaulay or modified

duration increases with time to maturity, as dollar duration decreases.

- Holding other factors constant, the lower a bond yield to maturity, the higher its Macaulay or modified duration and the lower its dollar duration.
- Duration is a linear operator. In other words, the duration of a portfolio P (D_P) invested in n bonds i denominated in the same currency with weights w_i is the weighted average of all bond durations (D_i):

$$D_P = \sum_{i=1}^n w_i D_i \quad (3)$$

This relationship holds for all definitions of duration (Macaulay, modified, dollar).

There are two commonly used measures of second-order interest rate risk. We have already encountered the convexity (C) but market practitioners also use the *dollar convexity*, defined as the bond convexity times the bond price. Dollar convexity is used to quantify the absolute change in a bond price due to convexity for a given change in the yield to maturity.

The main properties of the convexity and dollar convexity measures are as follows:

- For a given bond, the change in value due to the convexity term is always positive.
- Holding the maturity and the yield to maturity of a bond constant, the lower the coupon rate, the higher its convexity and the lower its dollar convexity.
- Holding the coupon rate and the yield to maturity of a bond constant, its convexity and dollar convexity increase with its time to maturity.
- Holding other factors constant, the lower a bond yield to maturity, the higher its convexity and dollar convexity.
- Convexity is a linear operator. In other words, the convexity of a portfolio P invested in n bonds denominated in the same currency with given weights is the weighted average of all bond convexities.

Table 1

Yield to maturity (%)	Bond price	Reinvested coupons	Total
4	104.422	10.550	114.972
4.5	104.352	10.620	114.972
5	104.282	10.690	114.972
5.5	104.212	10.760	114.972
6	104.142	10.830	114.972

End Notes

^aNote that the convexity of a zero-coupon bond is approximately equal to the square of its maturity.

References

- [1] Fabozzi, F.J. (1996). *Fixed-Income Mathematics*, 3rd Edition, McGraw-Hill, New York.
- [2] La Bruslerie, H. de (2002). *Gestion obligataire*, 2nd Edition, Economica, Paris.
- [3] Macaulay, F.R. (1938). *The Movements of Interest Rates, Bond Yields and Stock Prices in the United States since 1856*, National Bureau of Economic Research, New York.
- [4] Martellini, L., Priaulet, P. & Priaulet, S. (2003). *Fixed-Income Securities: Valuation, Risk Management and Portfolio Strategies*, Wiley Finance, Chichester.

Related Articles

Caps and Floors.

STÉPHANE PRIAULET

LIBOR Rate

LIBOR stands for London Interbank Offered Rate. It provides a measure of banks' borrowing costs and represents one of the most widely referenced interest rates in the world. LIBOR is owned by the British Bankers' Association (BBA); see [1] for information about BBA and its publications. The *LIBOR rate* is a rate at which contributing banks believe they could raise unsecured funds for a short term. LIBOR panel banks are those with the best credit ratings, and this benchmark describes availability of funds to banks with similar creditworthiness. LIBOR is used as the basis for settlement of interest rate contracts, both those that are traded on the exchanges worldwide (interest rate futures, *see Eurodollar Futures and Options*, and futures options), and the over-the-counter (OTC) transactions. LIBOR supports a swap market estimated at over \$300 trillion and a loan market estimated at over \$10 trillion. All LIBOR rates are the benchmarks set in the London market.

LIBOR is set for 10 different currencies at 11 am London time. The LIBOR rates therefore reflect the relative availability of the corresponding currencies' funds in European markets.

The actual calculation of the LIBOR rates is done by Reuters for the BBA, and the process is overseen by several committees. Sixteen contributor banks are selected for each of the four major currencies: US dollar (USD), sterling (GBP), euro (EUR), and yen (JPY). Between 8 and 12 contributor banks are selected for the other six currencies which are Australian dollar (AUD), Canadian dollar (CAD), Swiss franc (CHF), Danish krone (DKK), New Zealand dollar (NZD), and Swedish krona (SEK).

The process starts with polling the contributor banks. The rate submitted must be formed from the bank's perception of its cost of funds in the interbank market. Hence the quotes are obtained from the money market trading desks (who manage the banks' cash positions), and not from interest rate derivative traders. After the quotes are collected, the 25 percentile of the highest and the 25 percentile of the lowest quotes are discarded. This makes it virtually impossible for a contributor to skew the result. The remaining two quartiles of the quotes are averaged producing the number, which is then published as *LIBOR fixing*.

LIBOR rates are fixed each London business day for a set of short-term maturities up to 12 months. A fixed LIBOR rate refers to a certain interest rate period, starting at the so-called value date and ending at the maturity date (Figure 1). There is an offset between the fixing date and the value date for all currencies except sterling. More details about the rules and conventions governing LIBOR fixing can be obtained from the BBA documentation [1].

Forward Rate Agreements (FRAs)

A forward rate agreement (FRA) is a contract between two parties to fix a future interest rate based on a principal. It is an OTC product and represents an interest rate derivative in which one party (the buyer) pays a fixed interest rate and receives a floating interest rate, which is equal to some underlying rate called the *reference rate*. The most commonly used underlying rate is LIBOR.

FRAs are available for a variety of periods: starting from a few days to terms of several years. However, most of the liquidity in the FRA market is concentrated within 1 year, and those products are regarded as money-market instruments. FRAs are typically agreed on BBA-terms.

FRA's term period is normally specified by a pair of numbers separated by dash, forward slash, or very often by a cross: 0–3, 3/9, 3 × 6, and so on. The first number refers to the starting month and the second number denotes the ending month of the FRA term, counting from the current month. If no day of the month is given, the spot start is assumed; otherwise, the FRA term starts on a given day of the starting month. Figure 2 shows all dates that specify the FRA contract:

Trade date The date on which the contract is traded.

Value date spot Derived in the same way as the value date of the reference rate fixing.

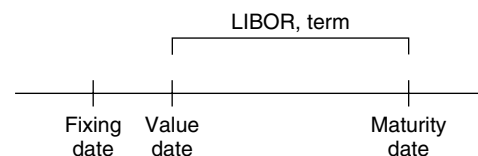


Figure 1 LIBOR rate dates

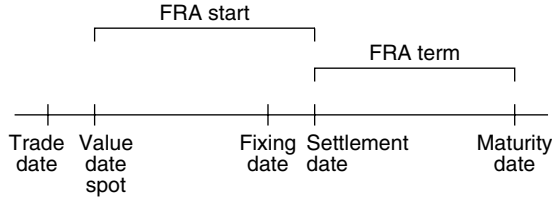


Figure 2 FRA dates

Settlement date The date on which the contract term commences. This is also the date when the amount due is paid by one party to the other.

Fixing date The date on which the reference rate is observed. For a LIBOR FRA, the relationship between the fixing date and the settlement date is the same as that between the fixing date and the value date shown in Figure 1.

Maturity date The end date of the FRA term.

The amount due is determined on the fixing date as a difference between interest rate payments for the buyer and the seller. The payment amount is

$$\frac{(L - R_0) \cdot \tau \cdot (\pm N)}{1 + L \cdot \tau} \quad (1)$$

where L is the reference rate fixing; R_0 is the FRA rate; N is the principal amount ($\pm =$ buy/sell); and τ is the time fraction of the FRA term.

The FRA settlement amount represents the difference in the values of deposits accrued to the maturity date; however, as the payment is exchanged on the settlement date, the appropriate discounting factor is applied, that is, $1 + L \cdot \tau$ in the denominator of (1).

The pricing of an FRA is reduced to forecasting the fixing of the reference rate. The present value of the FRA contract can be written as the expected value—under the pricing measure—of the payoff given by equation (1), assuming that the fixing occurs at T and S is the maturity (per one unit of principal):

$$PV(0) = E \left\{ \frac{P(T, S) \cdot [L(T, T, S) - R_0] \cdot \tau}{B(T)} \right\} \quad (2)$$

where $P(T, S)$ is the value of discount bond at T ; $B(T)$ is the value of a money market account; and $L(t, T, S)$ is the forward reference rate at t .

We observe that since $P(T, S) \cdot L(T, T, S) \cdot \tau \equiv 1 - P(T, S)$, the expression inside the expectation in equation (2) is a linear combination of traded assets of which discounted prices are martingales. Thus the FRA's values today can be expressed in terms of the forward rates:

$$\begin{aligned} PV(0) &= P(0, T) - P(0, S) - R_0 \cdot P(0, S) \cdot \tau \\ &= P(0, S) \cdot [L(0, T, S) - R_0] \cdot \tau \end{aligned} \quad (3)$$

In modeling literature, as in the derivation above, it is common to assume that discounting is done at rates consistent with forward LIBOR rates. In reality, however, discounting should reflect the true cost of financing of the corresponding derivative contract for a given bank. Such financing rate usually differs, and sometimes significantly, from the prevailing LIBOR forward rates. See [2, 3] for further details of derivative pricing using separate forward and discount curves.

References

- [1] Available at: <http://www.bba.org.uk> (2008).
- [2] Henrard, M. (2007). The irony in the derivatives discounting, *Wilmott Magazine* July, 92–98.
- [3] Traven, S. (2008). *Pricing Linear Derivatives with a Single Discount Curve*. working paper, unpublished.

Further Reading

Available at: <http://www.isda.org> (2008).

Related Articles

Bermudan Swaptions and Callable Libor Exotics; Bond; Caps and Floors; LIBOR Market Model.

SERGEI TRAVEN

Eurodollar Futures and Options

A Eurodollar rate is an interest rate on US dollar deposits that are held in banks outside the United States. The Eurodollar futures contract is a contract whose underlying is the three-month Eurodollar interest rate. When launched in 1981 at the Chicago Mercantile Exchange (CME), the Eurodollar futures were the world's first cash-settled futures contract. Since then, they have become the most actively traded short-term interest rate contract. At the time of writing, an average of more than 1.5 million contracts are traded each day and open interest exceeds 20 million positions (see [2]).

Each futures contract has a notional or “face value” of \$1 000 000. On the expiry date of the contract, the futures price is determined by the London Interbank Offered Rate (LIBOR; see **LIBOR Rate**), which is applied to Eurodollar deposits for a three-month period starting from the third Wednesday of the delivery month. If R is the fixing of the three-month LIBOR at expiration, expressed with quarterly compounding, the futures final settlement price is defined as $100 - R$, and the payoff of one futures contract is $1\,000\,000 \times (1 - 0.25 \times R\%)$. This results in futures price dropping when yield rises.

Eurodollar futures are closely related to the forward rate agreements (FRAs; see **LIBOR Rate**). However, the futures are marked-to-market daily, and for a long Eurodollar futures position, the margin payments are financed at a higher cost when rates rise and invested at a lower rate when rates decline. To compensate for the disadvantage of being long Eurodollar futures against FRA, the futures price must be lower and the futures rate ($= 100 - \text{futures price}$) must be higher than the corresponding forward rates. Empirical study by Burghardt [1] shows that this difference, known as “convexity bias”, may exceed 15 basis points for futures contracts with five years to expire. The exact magnitude of the bias is model dependent, for example, Hull [4] used the approximation $\frac{1}{2}\sigma^2 t_1 t_2$, where σ is the annualized standard deviation of the short-term interest rate, t_1 is the time of maturity of the futures contract, and t_2 is the time of maturity of the deposit. Readers should refer to [5, 6] for the theoretical framework (also see [8] and the reference therein for more results).

Eurodollar futures contracts are listed in a March quarterly expiration cycle. At any given time, there are forty quarterly expiries listed, spanning out 10 years, plus the four nearest serial months (non quarterly months). In addition to the outright individual contracts, the CME also lists a variety of products that enable one to initiate positions more efficiently on a particular segment of the yield curve. Some examples are orders known as *Packs and Bundles*. Packs are the simultaneous purchase or sale of an equally weighted, consecutive series of four Eurodollar futures. Bundles are consecutive series of futures beginning with the front contract. Both packs and bundles are quoted by the average of net price changes from the previous day's close of the constituent contracts.

CME lists American-style call and put options on Eurodollar futures. There are several types of such options. Quarterly Eurodollar options expire in the same month as the underlying futures. Serial options are listed for the two nearest serial months with the next quarterly futures as the underlying contracts. Midcurve options are options with short expiration on longer-dated futures. They expire one, two, or four years before the underlying quarterly futures expire. Option premiums are paid in full at the time of purchase, the so-called “stock-type” settlement.

Eurodollar options can be priced under the Black–Scholes model with dividend yield equal to the risk-free rate (see **Black–Scholes Formula**), and numerical methods are needed to evaluate the early exercise premium (see **American Options**). More complexity arises when considering the underlying futures that are highly correlated with the risk-free rate. We refer interested readers to [3] for a more detailed treatment of this subject.

Eurodollar futures and options are also traded at the Singapore Futures Exchange (SGX) and Euronext. The contracts traded at the CME and SGX are identical. Euronext adopts the “futures-type” settlement for options, where unlike the “stock-type”, premium is not paid up front but marked to market. Oviedo [7] showed that early exercise is not optimal under “futures-type” settlement; therefore, Euronext options can be priced “as if” European-style.

References

- [1] Burghardt, G. (2003). *The Eurodollar Futures and Options Handbook*, McGraw-Hill.

- [2] Chicago Mercantile Exchange. *CME Eurodollar Future Brochure*, CME website, available at: http://www.cme.com/files/eurodollar_futures.pdf
- [3] Henrard, M. (2005). *Eurodollar Futures and Options: Convexity Adjustment in HJM One-Factor Model*, available at SSRN: <http://ssrn.com/abstract=682343>
- [4] Hull, J.C. (2006). *Options, Futures, and Other Derivatives*, Pearson Prentice Hall.
- [5] Hunt, P. & Kennedy, J. (2000). *Financial Derivatives in Theory and Practice*, John Wiley and Sons.
- [6] Musiela, M. & Rutkowski, M. (2000). *Martingale Methods in Financial Modeling*, Springer.
- [7] Oviedo, R. (2006). *The Suboptimality of Early Exercise of Futures-Style Options: A Model-Free Result, Robust to Market Imperfections and Performance Bond Requirements*, available at SSRN: <http://ssrn.com/abstract=825104>
- [8] Piterbarg, V.V. & Renedo, M.A. (2006). Eurodollar futures convexity adjustments in stochastic volatility models, *Journal of Computational Finance* **9**(3), 71–94.

Related Articles

American Options; Black–Scholes Formula; LIBOR Rate.

YUHUA YU

Bond Options

Bond options are contracts giving their buyer the right but not the obligation to buy (call option) or sell (put option) an underlying bond at a prespecified price (the strike price). The strike for options on bond prices is generally quoted in terms of clean price (i.e., not including the accrued interest). Upon exercise, the bond buyer pays the accrued interest to the bond seller. In some cases, the contract for options on bonds may specify a strike yield instead of a strike price. Since price and yields move in opposite directions, a call (put) option on price will correspond to a put (call) option on yield.

Option contracts have a termination date referred to as the *expiration date* and are classified based on the possible dates in which the option holder can exercise the option. European-style options can be exercised only at the contract's expiration date. Bermudan-style options may be exercised on a pre-specified set of dates up to (or including) the expiration date, whereas American-style options can be exercised at any time up to (and including) the expiration date.

Most of the activity for options on cash bonds is over the counter (OTC) and has government bonds as the underlying, with the most common being options on long-term treasury bonds. The exchange-traded market for options on cash bonds is virtually nonexistent. Among the existing contracts, we mention the European options on treasury yields for 13-week treasury bills, 5- and 10-year treasury notes, and 30-year treasury bonds traded on the Chicago Board of Exchange (CBOE). These contracts are cash-settled, that is, there is no delivery of the physical underlying bond at exercise.

Closely related contracts are options on bond futures. These are exchange traded and significantly more liquid than options on cash bonds. In a bond futures contract, the side with the short position can deliver any bond in a predefined basket. For example, in the 30-year US treasury bond futures contract, \$100,000 of any US treasury bond that is not callable for at least 15 years and has a maturity of at least 15 years can be delivered. When a call (put) option on a bond futures is exercised, the buyer receives a cash payoff and is assigned

a long (short) position in the underlying futures contract. The seller is assigned the corresponding opposite position. The most popular options on futures contracts are those on 5- and 10-year US treasury notes and 30-year US treasury bonds traded at the Chicago Board of Trade (CBOT). Exchange-traded options on bond futures are typically American style.

Besides being traded as separate contracts, bond options frequently appear as provisions in the bond contract itself. In this case, because they cannot be traded separately, they are referred to as *embedded options*. The most common types are call provisions, giving the issuer the right but not the obligation to buy back the bond at a predetermined price. Bonds containing a call provision are known as *callable bonds*. Analogously, puttable bonds give the bond holder the right but not obligation to sell the bond back to the issuer at a predetermined strike price. A callable bond buyer is effectively buying the underlying noncallable bond and selling a call option on a bond to the issuer. In the case of a puttable bond, the issuer is selling a put option to the bond holder along with the underlying nonputtable bond. As with regular options, the embedded call and put options can be European, Bermudan, or American style.

Finally, as shown later, there is a close relationship between options on bonds and the most commonly traded OTC options on interest rates, namely, caps, floors, and swaptions. A cap (floor) is equivalent to a portfolio of European put (call) options on zero coupon bonds, whereas a swaption can be seen as an option on a fixed coupon bond.

We restrict our discussion in the following to the pricing of options on default-free bonds. The treatment of options on defaultable bonds is substantially more complex, since it requires consideration of both interest rate and credit components and their interaction. For more information, we refer to [6] and references therein.

European Options

We denote the price of a default-free zero coupon bond at time t maturing at T by $P(t, T)$. Consider European call and put options with strike K and expiration T_e written on a zero-coupon bond with maturity $T_m \geq T_e$. Their payoffs at time T_e are

given as

$$\max\{\omega(P(T_e, T_m) - K), 0\} = [\omega(P(T_e, T_m) - K)]^+ \quad (1)$$

with $\omega = 1$ for a call and $\omega = -1$ for a put.

Default-free bullet bonds pay, with certainty, pre-determined amounts c_i at times $T_i, i = 1 \dots, M$. We assume, for convenience, that the last payment also includes the repayment of principal. Their price can be trivially expressed in terms of prices of zero coupon bonds. More generally, defining for any given time $T, l(T) := \max\{i \in 1, \dots, M : T_i < T\}$, we can express the time $t \leq T$ value of the cash flows to be received after T as

$$B_{cp}(t, T, \{c\}) = \sum_{i=l(T)+1}^M c_i P(t, T_i) \quad (2)$$

Call and put options on this bond expiring at time T and with strike price K are options on a portfolio of zero coupon bonds having payoff

$$\max\{\omega(B_{cp}(T, T, \{c\}) - K), 0\} \quad (3)$$

with $\omega = 1$ for a call and $\omega = -1$ for a put.

Modeling typically proceeds in one of two ways. The first involves directly modeling the stochastic evolution of a finite number of bond prices (or yields) and is in essence an extension of the Black–Scholes–Merton approach [3, 10] to the pricing of option on bonds. The second involves modeling the stochastic evolution of the whole-term structure of interest rates, or equivalently, the simultaneous evolution of zero coupon bonds of all maturities. A rather general framework for term structure modeling was proposed by Heath, Jarrow, and Morton (HJM) [7] (*see also Heath–Jarrow–Morton Approach*), who modeled the instantaneous forward rates as diffusion processes and established the appropriate conditions that need to be satisfied to ensure the absence of arbitrage opportunities. In the term structure modeling approach, the dynamics of bonds and other interest rate instruments follow as a consequence of the stochastic model imposed on the term structure of interest rates.

Black–Scholes Model

Historically, the first approach for valuing bond options made direct use of the Black–Scholes–Merton model (*see also Black–Scholes Formula*). It assumed a geometric Brownian motion process with constant instantaneous volatility for the price process of the underlying bond and a deterministic short-term interest rate. The time 0 price of call and put options on a European bond option with strike price K and expiring at time T_e , under this model, is

$$\begin{aligned} \text{Call}(0, K, T_e, \{c\}) &= e^{-RT_e} \text{Bl}(K, e^{RT_e} B(0, T_e, \{c\}), \sigma_B \sqrt{T_e}, 1) \\ \text{Put}(0, K, T_e, \{c\}) &= e^{-RT_e} \text{Bl}(K, e^{RT_e} B(0, T_e, \{c\}), \sigma_B \sqrt{T_e}, -1) \end{aligned} \quad (4)$$

where

$$\begin{aligned} \text{Bl}(K, F, v, \omega) &:= F \omega \Phi \left(\omega \frac{\ln(F/K) + \frac{1}{2} v^2}{v} \right) \\ &\quad - K \omega \Phi \left(\omega \frac{\ln(F/K) - \frac{1}{2} v^2}{v} \right) \end{aligned} \quad (5)$$

Φ denotes the standard Gaussian cumulative distribution function and R is the continuously compounded spot interest rate for time T_e and σ_B the bond price volatility.

Similar to the convention used in the equity market, option prices are also quoted in terms of their implied volatility, namely, the constant volatility to be imputed in Black's formula so as to reproduce the option price. In some situations, the market quotes implied yield volatilities rather than implied price volatilities. The conversion from a yield volatility to a price volatility is given as

$$\sigma_B = D y \sigma_Y \quad (6)$$

where σ_B is the bond's price volatility, D is its modified duration, and σ_Y is its yield volatility.

There are some obvious drawbacks to the Black–Scholes–Merton approach applied to bond options.

First, contrary to stocks, bond prices tend to their face value at maturity (the so-called “pull to par” effect), implying that their instantaneous volatility must go to zero as they approach maturity. Under a constant volatility assumption, the model is only appropriate in situations where the time to expiration is significantly shorter than the underlying bond’s maturity. Another drawback is that given the lognormal distribution of prices, there is a nonzero probability that the price of the bond will be larger than the sum of the future cash flows, implying negative yields. Finally, the fact that bonds are modeled independently of each other does not guarantee the absence of arbitrage between them.

Black’s 1976 Model

In 1976, Black [2] extended the Black–Scholes–Merton analysis to commodity contracts. An important point was the shift in focus toward forward rather than spot quantities. Under this perspective, one can try and model the underlying bond’s forward price F for delivery at time T_e as a geometric Brownian motion with deterministic volatility $\sigma_F(t)$. Let R be the time 0 continuously compounded spot rate for time T_e . The pricing formulas for European bonds become

$$\begin{aligned} \text{Call}(0, K, T_e, \{c\}) &= e^{-RT_e} \text{Bl}(K, F, \Sigma_F \sqrt{T_e}, 1) \\ \text{Put}(0, K, T_e, \{c\}) &= e^{-RT_e} \text{Bl}(K, F, \Sigma_F \sqrt{T_e}, -1) \end{aligned} \quad (7)$$

with $\Sigma_F^2 := 1/T_e \int_0^T \sigma_F^2(s) ds$ the squared volatility of the bond’s forward price. Note that in equation (4) we take the underlying as the bond’s spot price with its corresponding volatility, whereas in equation (7) the underlying is the bond’s forward price with its corresponding volatility. Since the forward price for time T_e is given by the ratio of the spot price and the zero coupon bond maturing at T_e , these two volatilities are not the same.

A further change in perspective led to the use of Black’s model applied to forward rates rather than to forward prices. In the case of zero coupon bonds, define the simple compounded forward rate:

$$L(t; T_e, T_m) := \frac{1}{\tau} \left[\frac{P(t, T_e)}{P(t, T_m)} - 1 \right], \quad t \leq T_e \quad (8)$$

where τ denotes the year fraction for $(T_e, T_m]$. By using a standard change of measure (*see also Forward and Swap Measures*) it is easy to show that

$$\begin{aligned} P(0, T_e) E^{T_e} \{ [P(T_e, T_m) - K]^+ \} \\ = P(0, T_m) \tau K E^{T_m} \left\{ \left[\frac{1 - K}{\tau K} - L(T_e; T_e, T_m) \right]^+ \right\} \end{aligned} \quad (9)$$

where E^T denotes expectation with respect to a measure having as numeraire the zero coupon bond maturing at T . Therefore, a call option on a zero coupon bond can be regarded as a put option on the forward rate $L(T_e; T_e, T_m)$. We can now use Black’s model on the forward rate L assuming that it follows geometric Brownian motion with volatility $\sigma_L(t)$. The time 0 price of a call option with strike K and time to expiration T_e on a zero coupon bond maturing at T_m can then be written as

$$\begin{aligned} \text{Call}_{zc}(K, T_e, T_m) \\ = \tau K P(0, T_m) \\ \times \text{Bl}(K', L(0, T_e, T_m), \Sigma_L \sqrt{T_e}, -1) \end{aligned} \quad (10)$$

with $K' = \frac{1-K}{\tau K}$ and $\Sigma_L^2 := 1/T_e \int_0^{T_e} \sigma_L^2(s) ds$.

A similar argument can be used to obtain the price of put options on zero coupon bonds in terms of the prices of call options on the forward rate $L(T_e; T_e, T_m)$. In the OTC market for London Interbank Offered Rate (LIBOR) rate based interest rate derivatives, it became market standard to express options in terms of forward rates and to use Black’s formula for valuation (*see also Risk-neutral Pricing*). There is also an analogous relationship between options on coupon bonds and swaptions, which are options on forward swap rates.

Alternative models having the bond price as the only state variable have been proposed to account for some of the shortcomings of the Black–Scholes model. Ball and Torous [1] modeled the rate of return on a discount bond by a Brownian bridge process, which guarantees that the bond price converges to its face value at maturity. Schaefer and Schwartz [13] proposed a model where the price volatility of a coupon bearing bond is proportional to the duration and, therefore, decreases as the bond approaches

maturity. Further attempts in this direction are reviewed by Rady and Sandmann [12].

Term Structure Models

As we have mentioned, the HJM approach is a rather general framework for modeling the full-term structure, guaranteeing no arbitrage. To make progress and obtain explicit results for the price of options on bonds, one needs to choose concrete realizations. Among the most popular are the so-called *short-rate models*, which, as their names indicate, focus on the modeling of the instantaneous short rate $r(t)$ (see also **Term Structure Models**) defined as

$$r(t) = \frac{-d \log P(t, T)}{dT} \Big|_{T=t} \quad (11)$$

These models have a long history, dating back to Merton [10] and Vasicek [14]. In general, the short-rate process is assumed to follow a continuous Markov process, in which case prices of zero coupon bonds become a function of the short rate $r(t)$. The payoff for a European call option expiring at T on a coupon bond can then be written as

$$\left(\sum_{i=1}^n c_i P(r(T), T, T_i) - K \right)^+ \quad (12)$$

If the zero coupon bond price is a continuous decreasing function of $r(t)$, the option will be exercised if and only if $r(T) < r^*$, where r^* is the solution of

$$\sum_{i=1}^n c_i P(r^*, T, T_i) = K \quad (13)$$

In this case, one can decompose the value of the option into a sum of options on zero coupon bonds by rewriting the payoff as

$$\sum_{i=1}^n c_i (P(r(T), T, T_i) - P(r^*, T, T_i))^+ \quad (14)$$

The aforementioned decomposition was discovered by Jamshidian [8] and is sometimes called the *Jamshidian formula*. Several popular term structure models allow for an exact solution for the price of zero coupon bonds in terms of the model

parameters. Among them are the Vasicek/Hull–White [14] (see also **Cox–Ingersoll–Ross (CIR) Model**) and the Cox–Ingersoll–Ross (CIR) [5] (see **Gaussian Interest-Rate Models**) models. In their one-factor version, these two models satisfy the requirement that discount bonds are monotonically decreasing in the short rate, allowing us to obtain the prices of options on coupon bonds in terms of the prices of options on zero coupon bonds.

Bermudan and American Bond Options

The pricing of Bermudan and American options is significantly more complex than their European counterpart. There are typically no closed-form solutions and one needs to resort to numerical methods for computing their value. A common approach is to calibrate a term structure model to European options and subsequently use the calibrated model to price American and Bermudan securities. The latter step would typically involve a numerical scheme such as Monte Carlo simulation or a finite difference method. Among the most popular models used are the one- and two-factor short-rate models (see also **Term Structure Models**) as well as the LIBOR market model (see **LIBOR Market Model**) family [4, 9, 11]).

References

- [1] Ball, C.B. & Torous, W.N. (1983). Bond price dynamics and options, *Journal of Financial and Quantitative Analysis* **19**, 517–531.
- [2] Black, F. (1976). The market model of interest rate dynamics, *Mathematical Finance* **7**, 127–154.
- [3] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**(3), 637–654.
- [4] Brace, A., Gatarek, D. & Musiela, M. (1997). The market model of interest rate dynamics, *Mathematical Finance* **7**, 127–154.
- [5] Cox, J.C., Ingersoll, J.E. & Ross, S.A. (1985). A theory of the term structure of interest rates, *Econometrica* **53**, 385–407.
- [6] Duffie, D. & Singleton, K.J. (2003). *Credit Risk: Pricing, Measurement and Management*. Princeton Series in Finance.
- [7] Heath, D., Jarrow, R.A. & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation, *Econometrica* **60**(1), 77–105.

- [8] Jamshidian, F. (1989). An exact bond option formula, *The Journal of Finance* **44**(1), 205–209.
- [9] Jamshidian, F. (1997). Libor and swap market models and measures, *Finance and Stochastics* **1**(4), 293–330.
- [10] Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**(1), 141–183.
- [11] Miltersen, K.R., Sandmann, K. & Sondermann, D. (1997). Closed form solutions for term structure derivatives with log-normal interest rates, *The Journal of Finance* **52**, 409–430.
- [12] Rady, S. & Sandmann, K. (1994). The direct approach to debt option pricing, *The Review of Futures Markets* **13**(2), 461–515.
- [13] Schaefer, S. & Schwartz, E. (1987). Time-dependent variance and the pricing of bond options, *The Journal of Finance* **42**(5), 1113–1128.
- [14] Vasicek, O. (1977). An equilibrium characterisation of the term structure, *Journal of Financial Economics* **5**, 177–188.

Further Reading

- Black, F., Derman, E. & Toy, W. (1990). A one-factor model of interest rates and its application to treasury bond options, *Financial Analysts Journal* **46**, 24–32.
- Brigo, D. & Mercurio, F. (2006). *Interest-Rate Models: Theory and Practice. With Smile, Inflation and Credit*, Springer Finance.

Related Articles

Black–Scholes Formula; Caps and Floors; Cox–Ingersoll–Ross (CIR) Model; Forward and Swap Measures; Gaussian Interest-Rate Models; LIBOR Market Model; LIBOR Rate; Put–Call Parity; Term Structure Models.

MARCELO PIZA

Caps and Floors

Definitions and Notation

Let us consider a set of times $0 =: T_0, T_1, \dots, T_M$ and denote by $P(t, T)$ the discount factor at time t for the generic maturity T .

A *cap* is a contract that pays at each time $T_i, i = a, \dots, b, a \geq 1$, the difference, if positive, between the LIBOR rate set at the previous time T_{i-1} and a given rate K specified by the contract, which is shortly referred to as a *strike*. In formulas, assuming unit notional, the time- T_i payoff is

$$\tau_i \max\{L(T_{i-1}, T_i) - K, 0\} = \tau_i [L(T_{i-1}, T_i) - K]^+ \quad (1)$$

where τ_i denotes the year fraction for the interval $(T_{i-1}, T_i]$ and the LIBOR rate $L(T_{i-1}, T_i)$ is the simply compounded rate at time T_{i-1} for maturity T_i , namely,

$$L(T_{i-1}, T_i) := \frac{1}{\tau_i} \left[\frac{1}{P(T_{i-1}, T_i)} - 1 \right] \quad (2)$$

(see also **LIBOR Rate**).

Each single option in a cap is called a *caplet*. A cap is then a strip of caplets.

Analogously, a *floor* is a contract that pays at each time $T_i, i = a, \dots, b, a \geq 1$, the difference, if positive, between a strike K and the LIBOR rate set at time T_{i-1} :

$$\tau_i \max\{K - L(T_{i-1}, T_i), 0\} = \tau_i [K - L(T_{i-1}, T_i)]^+ \quad (3)$$

Each single option in a floor is called a *floorlet*. A floor is then a strip of floorlets.

A cap (floor) is said to be at-the-money (ATM) if its price is equal to that of the corresponding floor (cap). It is said to be in-the-money (ITM) or out-of-the-money (OTM) if its price is, respectively, higher or lower than that of the corresponding floor (cap).

As a consequence of the put-call parity, which, on each T_i , reads as

$$\tau_i \max\{L(T_{i-1}, T_i) - K, 0\}$$

$$\begin{aligned} & - \tau_i \max\{K - L(T_{i-1}, T_i), 0\} \\ & = \tau_i [L(T_{i-1}, T_i) - K] \end{aligned} \quad (4)$$

(see also **Put-Call Parity**), the difference between a cap and a floor with the same payment times $T_i, i = a, \dots, b, a \geq 1$, and strike K is an interest rate swap (IRS) where, at each time T_i , the floating rate $L(T_{i-1}, T_i)$ is exchanged for the fixed rate K . Therefore, a cap (floor) is ATM if and only if the related IRS has zero value, that is, if and only if its strike equals the underlying (forward) swap rate:

$$K = K_{\text{ATM}} := \frac{P(0, T_{a-1}) - P(0, T_b)}{\sum_{i=a}^b \tau_i P(0, T_i)} \quad (5)$$

Moreover, the cap is ITM if $K < K_{\text{ATM}}$ and OTM if $K > K_{\text{ATM}}$. The converse holds for a floor.

Pricing Formulas

It is market practice to price caps and floors with sums of corresponding Black's formulas (see **Black-Scholes Formula**). Precisely, if the cap (floor) pays on dates $T_i, i = a, \dots, b, a \geq 1$, and has a strike K , its market formula is given by

$$\begin{aligned} \text{Cap}(K, T_a, T_b; \sigma_b) &= \sum_{i=a}^b P(0, T_i) \tau_i \\ &\quad \times \text{Bl}(K, L(0, T_{i-1}, T_i), \sigma_b \sqrt{T_{i-1}}, 1) \\ \text{Floor}(K, T_a, T_b; \sigma_b) &= \sum_{i=a}^b P(0, T_i) \tau_i \\ &\quad \times \text{Bl}(K, L(0, T_{i-1}, T_i), \sigma_b \sqrt{T_{i-1}}, -1) \end{aligned} \quad (6)$$

where

$$\begin{aligned} \text{Bl}(K, L, v, \omega) &:= L \omega \Phi \left(\omega \frac{\ln(L/K) + \frac{1}{2} v^2}{v} \right) \\ &\quad - K \omega \Phi \left(\omega \frac{\ln(L/K) - \frac{1}{2} v^2}{v} \right) \end{aligned}$$

Φ denotes the standard Gaussian cumulative distribution function, and $L(0, T_{i-1}, T_i)$ is the simply

compounded forward LIBOR rate at time 0 for the interval $[T_{i-1}, T_i]$:

$$L(t; T_{i-1}, T_i) := \frac{1}{\tau_i} \left[\frac{P(t, T_{i-1})}{P(t, T_i)} - 1 \right], \quad t \leq T_{i-1} \quad (7)$$

The parameter σ_b is called the *cap- (floor-) implied volatility*. It is a single volatility parameter that must be inputted in each Black's formula in equation (6) to reproduce the market price.

A cap and a floor with the same maturity and strike must be priced with the same implied volatility, as a consequence of the put-call parity (4).

Justification for the Market Formulas

Practitioners used the market formulas (6) for years without a formal proof that they were indeed arbitrage free. Such a proof was first given by Jamshidian [3], followed by Miltersen *et al.* [4] and Brace *et al.* [1] with their celebrated lognormal LIBOR market models (*see also LIBOR Market Model*).

Let us consider a T_i -maturity caplet with strike K . The floorlet case is analogous. By classic risk-neutral valuation (*see also Risk-neutral Pricing*), the no-arbitrage price of the caplet payoff (1) at time 0 is

$$\begin{aligned} & E \left\{ e^{-\int_0^{T_i} r(t) dt} \tau_i [L(T_{i-1}, T_i) - K]^+ \right\} \\ &= E \left\{ e^{-\int_0^{T_i} r(t) dt} \tau_i [L(T_{i-1}; T_{i-1}, T_i) - K]^+ \right\} \end{aligned} \quad (8)$$

where E denotes the risk-neutral expectation and $r(t)$ the instantaneous short rate at time t . Switching to the T_i -forward measure Q^{T_i} , whose associated numeraire is the zero-coupon bond $P(t, T_i)$ (*see also Forward and Swap Measures*), the caplet price becomes

$$\begin{aligned} & E \left\{ e^{-\int_0^{T_i} r(t) dt} \tau_i [L(T_{i-1}; T_{i-1}, T_i) - K]^+ \right\} \\ &= P(0, T_i) \tau_i E^{T_i} \left\{ [L(T_{i-1}; T_{i-1}, T_i) - K]^+ \right\} \end{aligned} \quad (9)$$

where E^{T_i} denotes expectation under Q^{T_i} . The Black formula for the caplet in question can then be

obtained by assuming that under such a measure:

$$dL(t; T_{i-1}, T_i) = \varsigma_i L(t; T_{i-1}, T_i) dW_i(t) \quad (10)$$

where ς_i is the (constant and deterministic) caplet volatility and W_i is a standard Brownian motion under Q^{T_i} . Assuming a driftless dynamics for the forward rate $L(t; T_{i-1}, T_i)$ under Q^{T_i} is the only admissible choice since, by its own definition (7), the forward rate is a martingale under the T_i -forward measure (tradable asset divided by the numeraire).

A more detailed descriptions of these arguments can be found in **LIBOR Market Model**.

Market Quotes and Smiles

Given the pricing formulas (6), it is also a market practice to quote caps and floors through their implied volatility σ_b . Such a volatility is typically a function of the strike price, too: $\sigma_b = \sigma_b(K)$, meaning that caps (floors) with the same maturities but different strikes can be priced with different implied volatilities. This is called *smile effect*, due to the typical shape of the market implied volatility curves and surfaces.

The market quotes cap-/floor-implied volatilities for a number of maturities (up to 30 years for the main currencies). An example of cap implied volatility surface from the USD market is shown in Figure 1, where payment times are three-month spaced.

Let us denote by T_{j_k} , $k = 1, \dots, N$, the N cap maturities in a given market, and assume that $j_N = M$. From the corresponding cap quotes, one can strip the implied caplet volatilities ς_i , for a given strike K , by recursively solving, for $k = 1, \dots, N$,

$$\text{Cap}(K, T_a, T_{j_k}; \sigma_{j_k})$$

$$\begin{aligned} &= \sum_{i=a}^{j_k} P(0, T_i) \tau_i \text{Cpl}(K, T_{i-1}, T_i; \varsigma_i) \\ &= \sum_{i=a}^{j_k} P(0, T_i) \tau_i \text{Bl}(K, L(0, T_{i-1}, T_i), \varsigma_i \sqrt{T_{i-1}}, 1) \end{aligned} \quad (11)$$

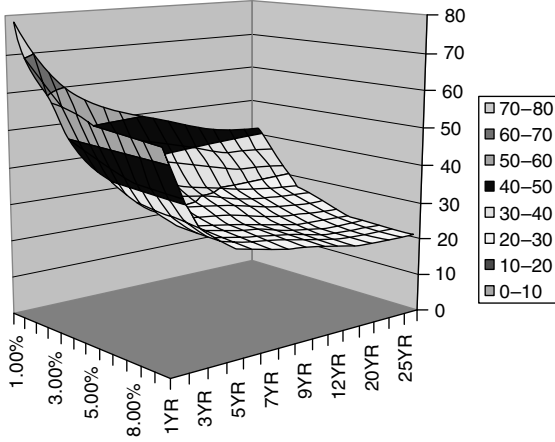


Figure 1 USD cap volatilities as of February 13, 2009. Strikes are in percentage points

Unfortunately, the number N of traded (cap) maturities is typically (much) smaller than the number of underlying (caplet) payments times. Therefore, stripping the caplet volatilities is neither trivial nor uniquely defined. A common approach is based on assuming, for each given strike, specific interpolations along the maturity dimension. The simplest choice is to assume that the ς_i are constant on the intervals defined by the cap maturities, namely, $\varsigma_i = \varsigma_h$ when $T_{j_{k-1}} < T_i, T_h \leq T_{j_k}$. This allows the recursive stripping of caplet volatilities from equation (11) by solving sequences of equations, each with a unique unknown.

The caplet-implied volatilities stripped from the caps of Figure 1 by assuming piece-wise constant interpolation along the maturity dimension are shown in Figure 2 for a limited range of maturities.

Beyond Black's Formula

The lognormal dynamics (10) imply that caplets with the same maturity T_i but different strikes are priced with the same implied volatility ς_i . However, as just seen, the implied volatilities quoted by the market change depending on the strike (smile effect). A popular alternative to equation (10), allowing for implied volatility smiles, is the stochastic alpha beta rho (SABR) model of Hagan *et al.* [2], where each forward LIBOR rate $L(t; T_{i-1}, T_i)$ is assumed to

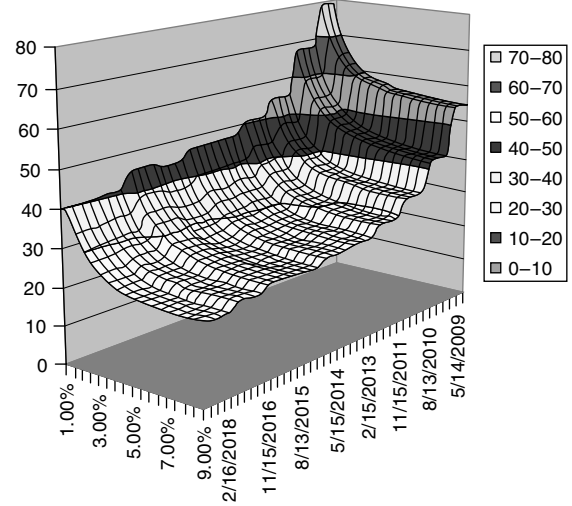


Figure 2 USD caplet volatilities as of February 13, 2009. Strikes are in percentage points

evolve under the corresponding forward measure Q^{T_i} according to

$$\begin{aligned} dL(t; T_{i-1}, T_i) &= V_i(t) L(t; T_{i-1}, T_i)^{\beta_i} dZ_i(t) \\ dV_i(t) &= \epsilon_i V_i(t) dW_i(t) \\ V_i(0) &= \alpha_i \end{aligned} \quad (12)$$

where Z_i and W_i are Q^{T_i} -standard Brownian motions with $dZ_i(t) dW_i(t) = \rho_i dt$ and where $\beta_i \in (0, 1]$, ϵ_i and α_i are positive constants, and $\rho_i \in [-1, 1]$.

Cap prices in the SABR model are given by the following closed form approximation:

$$\begin{aligned} \text{Cap}^{\text{SABR}}(K, T_a, T_b) &= \sum_{i=a}^b P(0, T_i) \tau_i \text{Bl}(K, L(0, T_{i-1}, T_i), \\ &\quad \sigma_i^{\text{SABR}}(K, L(0; T_{i-1}, T_i)) \sqrt{T_{i-1}}, 1) \end{aligned} \quad (13)$$

where

$$\sigma_i^{\text{SABR}}(K, L) = \frac{\alpha_i}{(LK)^{\frac{1-\beta_i}{2}} \left[1 + \frac{(1-\beta_i)^2}{24} \ln^2\left(\frac{L}{K}\right) + \frac{(1-\beta_i)^4}{1920} \ln^4\left(\frac{L}{K}\right) + \dots \right]^{\frac{z}{x(z)}}} \times \left\{ 1 + \left[\frac{(1-\beta_i)^2 \alpha_i^2}{24(LK)^{1-\beta_i}} + \frac{\rho_i \beta_i \epsilon_i \alpha_i}{4(LK)^{\frac{1-\beta_i}{2}}} + \epsilon_i^2 \frac{2-3\rho_i^2}{24} \right] T_{i-1} + \dots \right\} \quad (14)$$

with $z := \frac{\epsilon_i}{\alpha_i} (LK)^{\frac{1-\beta_i}{2}} \ln\left(\frac{L}{K}\right)$ and $x(z) := \ln \left\{ \frac{\sqrt{1-2\rho_i z + z^2} + z - \rho_i}{1-\rho_i} \right\}$. An analogous formula holds for floors.

The success of the SABR model is mainly due to the existence of an analytical formula for the implied volatilities σ_i^{SABR} , which is flexible enough to recover typical market smiles. In fact, it is a widespread practice to construct cap smiles by using the SABR functional form σ_i^{SABR} , assuming specific patterns for the parameters α_i , β_i , ϵ_i , and ρ_i between a caplet maturity and the next (e.g., piece-wise linear interpolation).

References

- [1] Brace, A., Gatarek, D. & Musiela, M. (1997). The market model of interest rate dynamics, *Mathematical Finance* **7**, 127–154.
- [2] Hagan, P.S., Kumar, D., Lesniewski, A.S. & Woodward, D.E. (2002). Managing smile risk, *Wilmott Magazine* September, 84–108.

- [3] Jamshidian, F. (1996). Sorting out swaptions, *Risk March*, 59–60.
- [4] Miltersen, K.R., Sandmann, K. & Sondermann, D. (1997). Closed form solutions for term structure derivatives with log-normal interest rates, *The Journal of Finance* **52**, 409–430.

Further Reading

Brigo, D. & Mercurio, F. (2006). *Interest-Rate Models: Theory and Practice. With Smile, Inflation and Credit*, Springer Finance.

Related Articles

Black–Scholes Formula; Forward and Swap Measures; LIBOR Market Model; LIBOR Rate; Put–Call Parity; Risk-neutral Pricing.

FABIO MERCURIO

Constant Maturity Swap

In-arrears swaps, averaging swaps, and constant maturity swaps (CMS) are simple extensions of vanilla interest rate swaps. Unlike vanilla swaps, these securities are sensitive to volatility *via* so-called convexity adjustments (*see* **Convexity Adjustments**). We demonstrate two methods for calculating convexity adjustments: the standard Black–Scholes framework and the replication method.

Market uncertainty is described *via* a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, Q)$ where $\mathcal{F}_t$ is a filtration associated with a Q -Brownian motion W_t . Q is the so-called risk-neutral probability measure, under which money-market-discounted asset prices are martingales. $\mathbb{E}_t[X] = \mathbb{E}[X|\mathcal{F}_t]$ is the conditional expectation under Q of the integrable random variable X .

Notations

We use the following notations within the article:

- r_t is the short rate at time t .
- $\beta_{t,T} = \exp(\int_t^T r_s ds)$ is the continuously rolled money-market account between t and T . Note that $\beta_{t,T} = \frac{\beta_{0,T}}{\beta_{0,t}}$ and that $\beta_{t,T}$ is $\mathcal{F}_T$ -measurable.
- $P(t, T)$ is the value at t of 1\$ paid at T . We have

$$P(t, T) = \mathbb{E}_t \left[\frac{\beta_{0,T}}{\beta_{0,t}} \right] = \mathbb{E}_t \left[e^{-\int_t^T r_s ds} \right] \quad (1)$$

- $G(w, f, N) = \sum_{i=1}^{f \cdot N} \frac{\delta(f)}{(1 + \delta(f)w)^i}$ is the cash level of the swap, or an approximation to the *present value of a basis-point*, or PV01, of a swap with different payment frequencies $f \in \{1, 2, 4\}$ and different maturities $N \in \mathbb{N}$. Note that $\delta(f) \approx \frac{1}{f}$ is the day-count fraction associated with a given day-count basis such as “Act360”, “30–360”, and so on. Going forward, we assume the frequency to be annual (i.e., $f = 1$), and T to be fixed and equal to N ; then we simply write $G(w, 1, N) = G(w)$.
- $L(t, T_1, T_2)$ is the value at t of the forward LIBOR rate, paid at T_2 , and determined at the settlement date T_1 , as described in **LIBOR Rate**.
- $L^f(t, T_1, T_2)$ is the futures rate with settlement date T_1 . Because of margin calls, the value

$L^f(t, T_1, T_2)$ differs from $L(t, T_1, T_2)$, as detailed in **Eurodollar Futures and Options**.

- Q^T denotes the T -forward probability measure (*see* **Forward and Swap Measures**). It is the risk-neutral probability measure associated with the numeraire $P(t, T)$ (*see* **LIBOR Rate**). For any $\mathcal{F}_T$ -measurable random variable X , the change of measures is given by

$$\mathbb{E}_t \left[\frac{\beta_{0,t}}{\beta_{0,T}} X \right] = P(t, T) \mathbb{E}_t^T[X] \quad (2)$$

where $\mathbb{E}_t^T(\cdot)$ is the conditional expectation under Q^T .

- The Q -Brownian motion, W_t , changes into W_t^T under Q^T *via* the Girsanov theorem.
- $P^f(t, T, T') = \frac{P(t, T')}{P(t, T)}$ is the forward zero-coupon, that is, the t -value of 1\$ paid at T' but borrowed at time $T < T'$. Note that $P^f(t, T, T')$ is a Q^T -martingale as $P(t, T)$ is a traded asset. More details on forward measures can be found in **Forward and Swap Measures** or in [5].
- $w_{t,T}$ is the forward swap rate of an N -year LIBOR swap paying annual coupons at dates $T_1, \dots, T_N$ (*see* **LIBOR Rate for the definition of a LIBOR swap**). We define T_0 to be the fixing date T . We have

$$w_{t,T} = \frac{P(t, T) - P(t, T_N)}{\sum_{i=1}^N \theta_i P(t, T_i)} = \frac{1 - P^f(t, T, T_N)}{\sum_{i=1}^N \theta_i P^f(t, T, T_i)} \quad (3)$$

The quantity θ_i is the day-count fraction for period $[T_{i-1}, T_i]$, $i = 1, \dots, N$.

- $A(t) = \sum_{i=1}^N \theta_i P(t, T_i)$ is the spot annuity paying θ_i at times $T_1, \dots, T_N$.
- $A^f(t, T) = \sum_{i=1}^N \theta_i \frac{P(t, T_i)}{P(t, T)}$ is the forward annuity paying θ_i at time $T_1, \dots, T_N$ forward in time, as of T . Note that $A^f(T, T) = A(T)$.
- The probability measure Q^{A_T} (*see* **Forward and Swap Measures**) associated with the numeraire $A(\cdot)$ is defined *via* the following measure change. For any $\mathcal{F}_T$ -measurable random variable X ,

$$\mathbb{E}_t \left[\frac{\beta_{0,t}}{\beta_{0,T}} A(T) X \right] = A(t) \mathbb{E}_t^{A_T}[X] \quad (4)$$

Q^{A_T} is usually called the *annuity*, or *swap*, measure. Note that for $N = 1$, $Q^{A_{T_1}}$ coincides with Q^{T_1} .

- The Q -Brownian motion, W_t , changes into $W_t^{A_T}$ under Q^{A_T} via the Girsanov theorem.
- A swaption is an option to enter into a swap at a future date T . We define K as the fixed rate in the swap. The price at $t < T$ of a cash-settled swaption is given by $CSS_{t,T}(K, \zeta) = \mathbb{E}_t \left[\beta_{t,T}^{-1} G(w_{T,T}) (\zeta \cdot w_{T,T} - \zeta \cdot K)_+ \right]$ with $\zeta = 1$ for a payer swaption and $\zeta = -1$ for a receiver swaption.
- In this section, we define the payout of a CMS caplet referencing the swap rate $w_{\cdot,T}$ with strike K as $(w_{T,T} - K)_+$, and paying at the fixing date T . So its price is $CMS_{caplet,t,T}(K) = \mathbb{E}_t[\beta_{t,T}^{-1}(w_{T,T} - K)_+]$. We define the CMS floorlet as $CMS_{floorlet,t,T}(K) = \mathbb{E}_t[\beta_{t,T}^{-1}(K - w_{T,T})_+]$. For a definition of standard caps and floors, see **Caps and Floors**: a standard (LIBOR) caplet with maturity T_{i+1} is given by $Caplet_{t,T_{i+1}}(K) = \mathbb{E}_t[\beta_{t,T_{i+1}}^{-1}(L(T_i, T_i, T_{i+1}) - K)_+]$.
- In CMS and LIBOR swaps, we distinguish two different types of payments: the nonstandard case (in-arrears), where the floating cash flow fixes on the same date as it pays, that is, T , and the standard case, (in advance), where the fixing takes place, say, 3 or 6 months before the payment.

Constant Maturity Swaps

Constant maturity swaps, constant maturity treasuries (CMT), and modified schedule LIBOR swaps are extensions of the standard fixed-for-floating swaps. Unlike the standard vanilla interest rate swaps (see **LIBOR Rate**) that specify an exchange of a fixed coupon for a LIBOR rate, CMS instruments pay the swap rate that is reset at each period *versus* either a fixed coupon or a LIBOR rate plus a spread. For CMT swaps, the structure is identical, but instead of the swap rate, the yield of a government bond is referenced. These swaps are often suggested as a way to benefit from a steepening or a flattening of the yield curve, while still being hedged against its parallel moves. To illustrate, let us consider the following CMS instrument: one pays 3-month LIBOR rate and receives the 10-year swap rate, denoted by $CMS_{10Y}(t)$, that is reset (observed) every 6 months. If, during the life of the swap (for instance 10 years)

the curve steepens, so that $CMS_{10Y}(t)$ increases as time t goes by relative to the LIBOR rate, the holder realizes a positive carry, that is, the amount of received cash flows exceeds the amount of paid cash flows. Given that LIBOR and CMS rates are reset at each period, a sudden parallel shift of the curve has no effect on the present value of the swap because of the mutual offset of both legs. As instruments that express views on the nonparallel moves of the yield curve, both CMS and CMT instruments are very popular and generally very liquid, and also often serve as building blocks for more complicated derivatives such as options on CMS rates or CMS spreads (differences between CMS rates of different tenors). Given their widespread use, accurate pricing of the underlying CMS cash flows is important. We describe the nonstandard CMS first, as the standard CMS is the nonstandard CMS adjusted for a payment lag.

There are two main methods for valuing CMS instruments: a standard Black–Scholes approach and a replication method. The first, a more traditional approach, assumes the swap rate to be log-normal, allowing the standard Black–Scholes framework to be used (see **Black–Scholes Formula for this model**). The key benefit of this approach is that it leads to a closed-form formula for the value of a CMS cash flow, also allowing options on CMS rates and other derivatives to be priced with the Black–Scholes formula.

We will show that the value of the CMS rate cash flow differs from the forward swap rate as it depends on volatility. This difference, called *convexity adjustment*, is often explained by the fact that in a CMS cash flow, the rate is paid once, rather than on its “natural” schedule. More details about convexity can be found in [4], [6], and [9].

Before describing these pricing methods in some detail, we introduce some notations. We know that in a CMS cash flow, the swap rate is paid once, at T , so the CMS value at $t < T$ is simply given by

$$\begin{aligned} P(t, T) \cdot CMS_t &= \mathbb{E}_t[\beta_{t,T}^{-1} w_{T,T}] \\ &= P(t, T) \mathbb{E}_t^T[w_{T,T}] \end{aligned} \quad (5)$$

so that

$$CMS_t = \mathbb{E}_t^T[w_{T,T}] \quad (6)$$

Note that $w_{\cdot,T}$ is not a Q^T -martingale, so $\mathbb{E}_t^T[w_{T,T}] \neq w_{t,T}$. The difference between CMS_t and

the forward swap rate $w_{t,T}$ is the convexity adjustment. Let us write the formula under the annuity measure Q^{A_T} , the probability measure under which $w_{\cdot,T}$ is a martingale. Using equation (4), we get

$$CMS_t = \frac{A(t)}{P(t,T)} \mathbb{E}_t^{A_T} \left[\frac{w_{T,T}}{A(T)} \right] \quad (7)$$

We see that CMS_t is a function of both the curve level via $\frac{A(t)}{P(t,T)}$ and the covariance between $A(T)$ and $w_{T,T}$. This strongly hints that the convexity adjustment is a function of the volatility of the forward swap rate $w_{T,T}$.

CMS Convexity Adjustment: the Simple Approach

The aim of this approach is to find a simple, closed-form approximation, assuming that the forward swap rate is log-normal: it is consequently often referred as the *Black* convexity adjustment. As $w_{\cdot,T}$ is a Q^{A_T} -martingale, we have $\frac{dw_{u,T}}{w_{u,T}} = \sigma dW_u^{A_T}$ at any time u between t and T . We assume that the variance of the swap rate under its swap measure is equal to its variance under the T -forward measure, that is, $\mathbb{E}_t^T [w_{T,T} - w_{t,T}]^2 = \sigma^2 (T-t) w_{t,T}^2$ (more precisely, we assume the convexity adjustment to be an adjustment of second order). We further assume that at any time u between t and T , $A^f(u, T) \approx G(w_{u,T})$, so that we approximate $w_{u,T}$ with the par yield of the forward starting bond—a common assumption used by market participants to, for example, compute the swap PV01. Since $A^f(\cdot, T)$ is a Q^T -martingale, so it is approximately $G(w_{\cdot,T})$. Performing a Taylor expansion to second order of $G(w_{T,T})$ around $w_{t,T}$, we obtain

$$\begin{aligned} G(w_{T,T}) &= G(w_{t,T}) + (w_{T,T} - w_{t,T}) G'(w_{t,T}) \\ &\quad + \frac{1}{2} (w_{T,T} - w_{t,T})^2 G''(w_{t,T}) \\ &\quad + o(w_{T,T} - w_{t,T})^2 \end{aligned} \quad (8)$$

Given also that $\mathbb{E}_t^T [G(w_{T,T})] = G(w_{t,T})$, we apply the expected value operator $\mathbb{E}_t^T$ to both sides of (8) to give us

$$CMS_t = w_{t,T} \left(1 - \frac{1}{2} w_{t,T} \cdot \sigma^2 \cdot (T-t) \cdot \frac{G''(w_{t,T})}{G'(w_{t,T})} \right)$$

$$\simeq w_{t,T} \exp \left(-\frac{1}{2} w_{t,T} \cdot \sigma^2 \cdot (T-t) \cdot \frac{G''(w_{t,T})}{G'(w_{t,T})} \right) \quad (9)$$

From the above formula, we see that the convexity adjustment $\exp \left(-\frac{1}{2} w_{t,T} \cdot \sigma^2 \cdot (T-t) \cdot \frac{G''(w_{t,T})}{G'(w_{t,T})} \right) - 1$ is positive given that G is decreasing, convex ($G' < 0$ and $G'' > 0$), and increases with the implied volatility σ .

CMS Convexity Adjustment: the Replication Approach

The simple approach does not capture the fact that the implied volatilities exhibit volatility smile. The replication method we develop in this section captures the smile effect, as noted by Amblard *et al.* [1].

The simple adjustment has a further drawback that it does not provide a way to construct a hedge for a CMS cash flow. Indeed, to hedge the CMS using the Black–Scholes adjustment, one can compute the delta and vega, but this is not a static hedge. The replication approach rectifies this issue as well.

The replication approach has the disadvantage of being nonparametric, but on the other hand, it takes into account the volatility smile and exhibits an explicit static hedge for the CMS rate cash flows and options in terms of payer and receiver swaptions. The computation is done in two steps. First we compute a CMS caplet and a CMS floorlet at a given strike k (see Notations for a definition of CMS caplets and floorlets; *see also Caps and Floors for a definition of caps and floors*). The CMS rate is then obtained *via* the call–put parity.

We use a numerical integration, commonly used to replicate European options with complex payoffs using vanilla European options (see [3]). Indeed, for any real C_2 -function V of the swap rate w and any real number x , we have

$$\begin{aligned} V(w) &= V(x) + (w - x) \cdot V'(x) \\ &\quad + 1_{\{w > x\}} \int_x^w V''(K) (w - K)_+ dK \\ &\quad + 1_{\{w < x\}} \int_w^x V''(K) (K - w)_+ dK \end{aligned} \quad (10)$$

Observing that the boundaries for the integrals can be extended to $+\infty$ and $-\infty$, respectively, we have

$$\begin{aligned} V(w) &= V(x) + (w - x) \cdot V'(x) \\ &\quad + 1_{\{w > x\}} \int_x^{+\infty} V''(K)(w - K)_+ dK \\ &\quad + 1_{\{w < x\}} \int_{-\infty}^x V''(K)(K - w)_+ dK \end{aligned} \quad (11)$$

so by setting $x = 0$

$$\begin{aligned} V(w) &= V(0) + wV'(0) \\ &\quad + 1_{\{w > 0\}} \int_0^{+\infty} V''(K)(w - K)_+ dK \\ &\quad + 1_{\{w < 0\}} \int_{-\infty}^0 V''(K)(K - w)_+ dK \end{aligned} \quad (12)$$

Let $k > 0$ be the strike of CMS caplet or CMS floorlet.

To compute the CMS caplet, let us define $V(w) = \frac{(w - k)_+}{G(w)}$. Then as $V(0) = V'(0) = 0$:

$$\begin{aligned} (w - k)_+ &= 1_{\{w > 0\}} \int_0^{+\infty} V''(K)G(w)(w - K)_+ dK \\ &\quad + 1_{\{w < 0\}} \int_{-\infty}^0 V''(K)G(w)(K - w)_+ dK \end{aligned} \quad (13)$$

Equation (13) expresses the general decomposition of a call payoff $(w - k)_+$ on a continuous set of payoffs $(G(w)(w - K)_+)_{K \geq 0}$ and $(G(w)(K - w)_+)_{K \leq 0}$.

By replacing w with $w_{T,T}$, multiplying both sides by $\beta_{t,T}^{-1}$, taking the $\mathcal{F}_t$ -conditional expectation under the risk-neutral measure, and using the definition of a CMS caplet in the section Notations, we obtain

$$\begin{aligned} CMScaplet_{t,T}(k) &= \int_0^{+\infty} V''(K) \mathbb{E}_t[\beta_{t,T}^{-1} G(w_{T,T})(w_{T,T} - K)_+] dK \\ &\quad + \int_{-\infty}^0 V''(K) \mathbb{E}_t[\beta_{t,T}^{-1} G(w_{T,T})(K - w_{T,T})_+] dK \end{aligned} \quad (14)$$

On the right-hand side, we recognize the prices of cash-settled swaptions as described in the section Notations. Note that $G(w_{T,T})$ is an approximation for $A(T)$ in the cash-settled swaption formula. So we have

$$\begin{aligned} CMScaplet_{t,T}(k) &= \int_0^{+\infty} V''(K) CSS_{t,T}(K, \zeta = 1) dK \\ &\quad + \int_{-\infty}^0 V''(K) CSS_{t,T}(K, \zeta = -1) dK \end{aligned} \quad (15)$$

If we assume $k > 0$, the second integral vanishes as $V''(K) = 0$ if $K < k$. To be tractable, this decomposition must be discretized and the integrals ultimately truncated:

$$CMScaplet_{t,T}(K) \approx \sum_{i=0}^{N_{\text{cap}}} \alpha_i^{\text{cap}} CSS_{t,T}(K_i, x = 1) \quad (16)$$

We start from $K_0 = k + \varepsilon$, with ε very small to avoid the discontinuity at $K = k$, and set $K_i = K_0 + i \Delta K$. The strike step ΔK is also arbitrary. We have $\alpha_i^{\text{cap}} = 2h'(K_i) + (K_i - k)h''(K_i)$ with $h(x) = \frac{1}{G(x)}$. In practice, the choice of N_{cap} or, equivalently, the last payer swaption strike used for replication can be computed dynamically: we add in that case an extra strike until we reach $\alpha_i^{\text{cap}} CSS_{t,T}(K_i, x = 1) < \text{ChosenPrecision}$. Depending on the assumption chosen for the smile extrapolation far out of the money, this can lead to some divergence of the replication algorithm. In that case, we can use an arbitrary cap, for example $K_{\text{max}}^{\text{cap}} = 50\%$.

The CMS floorlet is computed in a similar way, replacing $V(w)$ by $\frac{(k - w)_+}{G(w)}$. Note that we have $V(0) \neq 0$ and $V'(0) \neq 0$, so we have two extra terms:

$$\begin{aligned} CMSfloorlet_{t,T}(k) &= V(0) \mathbb{E}_t[\beta_{t,T}^{-1} G(w_{T,T})] \\ &\quad + V'(0) \mathbb{E}_t[\beta_{t,T}^{-1} w_{T,T} G(w_{T,T})] \\ &\quad + \int_0^{+\infty} V''(K) \mathbb{E}_t[\beta_{t,T}^{-1} G(w_{T,T})(w_{T,T} - K)_+] dK \\ &\quad + \int_{-\infty}^0 V''(K) \mathbb{E}_t[\beta_{t,T}^{-1} G(w_{T,T}) \\ &\quad \times (K - w_{T,T})_+] dK \end{aligned} \quad (17)$$

If we assume that rates cannot be negative, the second integral vanishes. In the current environment, it is not clear that rates cannot go slightly negative, so we should also compute the second part, which has value if we use, for example, a Gaussian model. Assuming $G(w_{T,T})$ as an approximation for $A(T)$ we have $\mathbb{E}_t[\beta_{t,T}^{-1}G(w_{T,T})] \approx A(t)$ and using equation (4), we get $\mathbb{E}_t[\beta_{t,T}^{-1}w_{T,T}G(w_{T,T})] \approx A(t)w_{t,T}$. We start the first integration from $K_0^\alpha = k - \varepsilon$ and set $K_i^\alpha = K_0^\alpha - i\Delta K$, with $\alpha_i^{\text{floor}} = -2h'(K_i^\alpha) + (k - K_i^\alpha)h''(K_i^\alpha)$ until we reach $K_{N_1}^\alpha = 0$. Concerning the receiver swaption, we start from $K_0^\beta = 0$ and set $K_j^\beta = K_0^\beta - j\Delta K$ with $\beta_j^{\text{floor}} = -2h'(K_j^\beta) + (k - K_j^\beta)h''(K_j^\beta)$ until we reach an arbitrary lower bound for the receivers $K_{N_2}^\beta = K_{\min} (< 0)$. Then we have

$$\begin{aligned} & CMS_{\text{floorlet},T}(K) \\ & \approx (V(0) + V'(0)w_{t,T})A(t) \\ & + \sum_{i=0}^{N_1} \alpha_i^{\text{floor}} CSS_{t,T}(K_i^\alpha, \zeta = 1) \\ & + \sum_{j=0}^{N_2} \beta_j^{\text{floor}} CSS_{t,T}(K_j^\beta, \zeta = -1) \quad (18) \end{aligned}$$

To compute the CMS rate, we see from expression (6) that if we assume interest rates to be positive as in the Black–Scholes case, we can value the CMS rate directly from a CMS caplet struck at $k = 0$ as $\mathbb{E}_t[(w_{T,T})_+] = \mathbb{E}_t[w_{T,T}]$:

$$CMS_t = \frac{1}{P(t,T)} CMS_{\text{caplet},T}(k = 0) \quad (19)$$

Alternatively, we use the call–put parity. Note that the choice of the strike k in the caplet and floorlet is arbitrary (a common practice is to set the strike to the forward swap rate $k = w_{t,T}$):

$$\begin{aligned} & CMS_t \\ & = k + \frac{CMS_{\text{caplet},T}(k) - CMS_{\text{floorlet},T}(k)}{P(t,T)} \quad (20) \end{aligned}$$

Alternative pricing methods have been proposed in the literature such as higher order Taylor expansions or chaos expansions (see [2]) or linear swap rate model approximation (see [6–8]).

Computation of the CMS rate in the standard case

In the standard case, fixing is at T , and payment takes place at $T + \theta$, so we have $P(t, T + \theta) \cdot CMS_t^{\text{std}} = \mathbb{E}_t\left[\frac{\beta_{0,t}}{\beta_{0,T+\theta}}w_{T,T}\right]$. We see that we can no longer apply the change of probability as in equation (4) because of the presence of $\beta_{0,T+\theta}$ instead of $\beta_{0,T}$: we artificially introduce $\beta_{0,T}$ in order to use equation (4). We have

$$\begin{aligned} \mathbb{E}_t\left[\frac{\beta_{0,t}}{\beta_{0,T+\theta}}w_{T,T}\right] &= \mathbb{E}_t\left[\frac{\beta_{0,t}}{\beta_{0,T}}\frac{\beta_{0,T}}{\beta_{0,T+\theta}}w_{T,T}\right] \\ &= \mathbb{E}_t\left[\frac{\beta_{0,t}}{\beta_{0,T}}w_{T,T}B(T, T + \theta)\right] \\ &= P(t, T)\mathbb{E}_t^T[w_{T,T}B(T, T + \theta)] \quad (21) \end{aligned}$$

There are at least two methods for computing the above quantity.

The first method consists in replacing $B(T, T + \theta)$ with a function involving $w_{T,T}$ plus a spread s . For example, we could write $B(T, T + \theta) = \frac{1}{1 + \theta(w_{T,T} + s)}$ so that $\mathbb{E}_t^T[w_{T,T}B(T, T + \theta)] = \mathbb{E}_t^T[f(w_{T,T})]$ with $f(w) = \frac{w}{1 + \theta(w + s)}$. We then use the replication formula (12) by replacing $V(w)$ with $V(w) = \frac{f(w)}{G(w)}$.

Alternatively, we can replace $B(T, T + \theta)$ with $B(T, T + \theta) = \frac{1}{1 + \theta L(T, T, T + \theta)}$. The approximation:

$$\frac{1}{1 + \theta L(T, T, T + \theta)} \approx 1 - \theta L(T, T, T + \theta) \quad (22)$$

together with a correlation assumption $\rho = \text{corr}^{Q^T}(L(T, T, T + \theta), w_{T,T})$ between the forward LIBOR and the forward swap rate under measure Q^T give us

$$\begin{aligned} & \mathbb{E}_t^T[w_{T,T}B(T, T + \theta)] \\ & \approx CMS_t - \theta CMS_t \mathbb{E}_t^T[L(T, T, T + \theta)] \\ & \quad - \theta \rho \sqrt{\text{Var}^{Q^T}(w_{T,T})} \sqrt{\text{Var}^{Q^T}(L(T, T, T + \theta))} \quad (23) \end{aligned}$$

where CMS_t is the nonstandard CMS rate computed previously and $\mathbb{E}_t^T[L(T, T, T + \theta)]$ is a convexity-adjusted forward LIBOR (see the next section for

more details). The variances under Q^T of the forward swap and LIBOR rates can be both valued *via* replication or approximated with the variances under their “natural” measures, respectively, Q^A and $Q^{T+\theta}$. Under Black–Scholes assumptions, they are simply $w_{t,T}^2(e^{(T-t)\sigma_w^2(T)} - 1)$ and $L(t, T, T + \theta)^2(e^{(T-t)\sigma_L^2(T)} - 1)$, respectively.

On the one hand, the second method involves more computations, themselves involving approximations, and uses an input ρ not explicitly given by the market. However, it better illustrates the “negative” convexity in the standard CMS case, that is, the presence of a negative second term in the formula.

In-arrears Swaps

In this section, we deal with LIBOR swaps, the CMS case having been described previously. As defined in the section Notations, *in-arrears* swaps are interest rate swaps that differ from vanilla swaps by the absence of a payment lag in the floating leg. More precisely, in a vanilla swap, the floating rate $L(T_{i-1}, T_{i-1}, T_i)$ fixes at T_{i-1} and is paid at T_i , that is, payment at T_i is $\theta_i L(T_{i-1}, T_{i-1}, T_i)$ while in an in-arrears swap, the floating rate is fixed and paid the same date at the end of the period, that is, the payment at T_i is $\theta_{i+1} L(T_i, T_i, T_{i+1})$. As we will see, this change also creates a positive convexity adjustment in the valuation of the floating leg. The valuation of in-arrears swaps can be done by at least two different methods: the traditional Black–Scholes adjustment and the replication technique. For the replication method, the underlying replication instruments are caplets/floorlets as opposed to swaptions for the CMS cash flows described in the previous section.

The time t -value of the in-arrears payment at T_i is

$$\begin{aligned} & \mathbb{E}_t \left[\frac{\beta_{0,t}}{\beta_{0,T_i}} \theta_{i+1} L(T_i, T_i, T_{i+1}) \right] \\ &= \theta_{i+1} P(t, T_i) \mathbb{E}_t^{T_i} [L(T_i, T_i, T_{i+1})] \end{aligned} \quad (24)$$

As the forward LIBOR rate $L(\cdot, T_i, T_{i+1})$ is a $Q^{T_{i+1}}$ -martingale, we have

$$\mathbb{E}_t^{T_{i+1}} [L(T_i, T_i, T_{i+1})] = L(t, T_i, T_{i+1}) \quad (25)$$

but

$$\mathbb{E}_t^{T_i} [L(T_i, T_i, T_{i+1})] \neq L(t, T_i, T_{i+1}) \quad (26)$$

To use the martingale property of $L(\cdot, T_i, T_{i+1})$ under $Q^{T_{i+1}}$, we discount $\theta_{i+1} L(T_i, T_i, T_{i+1})$ from date T_{i+1} . Obviously, the cash flow must be capitalized between T_i to T_{i+1} to compensate for this “late” discounting:

$$\begin{aligned} & \mathbb{E}_t \left[\frac{\beta_{0,t}}{\beta_{0,T_i}} \theta_{i+1} L(T_i, T_i, T_{i+1}) \right] \\ &= \mathbb{E}_t \left[\frac{\beta_{0,t}}{\beta_{0,T_{i+1}}} \theta_{i+1} L(T_i, T_i, T_{i+1}) \frac{\beta_{0,T_{i+1}}}{\beta_{0,T_i}} \right] \\ &= \mathbb{E}_t \left[\frac{\beta_{0,t}}{\beta_{0,T_{i+1}}} \theta_{i+1} L(T_i, T_i, T_{i+1}) \right. \\ & \quad \left. \times (1 + \theta_{i+1} L(T_i, T_i, T_{i+1})) \right] \\ &= P(t, T_{i+1}) \theta_{i+1} L(t, T_i, T_{i+1}) \\ & \quad + P(t, T_{i+1}) \theta_{i+1}^2 \mathbb{E}_t^{T_{i+1}} [L^2(T_i, T_i, T_{i+1})] \end{aligned} \quad (27)$$

Now we can compute $\mathbb{E}_t^{T_{i+1}} [L^2(T_i, T_i, T_{i+1})]$ either using a standard Black–Scholes adjustment or *via* replication on caplets. In the first case, assuming that the at-the-money implied volatility of the forward rate $L(\cdot, T_i, T_{i+1})$ is σ_i , we find $\mathbb{E}_t^{T_{i+1}} [L^2(T_i, T_i, T_{i+1})] = L(t, T_i, T_{i+1})^2 e^{\sigma_i^2(T_i-t)}$. In the second case, we can use formula (12) to find that

$$x^2 = 2 \int_0^{+\infty} (x - K)_+ dK \quad (28)$$

so that using the caplet definition given in the section Notations

$$\begin{aligned} & \mathbb{E}_t^{T_{i+1}} [L^2(T_i, T_i, T_{i+1})] \\ &= 2 \int_0^{+\infty} \mathbb{E}_t^{T_{i+1}} [(L(T_i, T_i, T_{i+1}) - K)_+] dK \\ &= \frac{2}{P(t, T_{i+1})} \int_0^{+\infty} \text{Caplet}_{t, T_{i+1}}(K) dK \end{aligned} \quad (29)$$

To be tractable, this integral must be discretized and capped as previously described in the CMS case.

Averaging Swaps

As nonstandard fixed-for-floating interest rate swaps, *averaging* swaps are somewhat less common than

CMS and in-arrears swaps. The floating leg of such a swap pays at each coupon date T_i the average of the LIBOR rates observed over a predetermined period. Let us consider a set of dates $T_{i(k)}$ with $i(k) < i$. The cash flow at date T_i is given by

$$\frac{1}{N} \sum_{k=1}^N \theta_{i(k)+1} L(T_{i(k)}, T_{i(k)}, T_{i(k)+1}) \quad (30)$$

So assuming that under the “terminal measure” Q_{T_i} , we know the drift $\mu_{i(k)}$,

$$\frac{dL(t, T_{i(k)}, T_{i(k)+1})}{L(t, T_{i(k)}, T_{i(k)+1})} = \mu_{i(k)}(t) dt + \sigma_{i(k)} dW_t^{T_i} \quad (31)$$

the averaging formula has a value at time t given by

$$\begin{aligned} & \mathbb{E}_t \left[\beta_{t, T_i}^{-1} \frac{1}{N} \sum_{k=1}^N \theta_{i(k)+1} L(T_{i(k)}, T_{i(k)}, T_{i(k)+1}) \right] \\ &= P(t, T_i) \sum_{k=1}^N \theta_{i(k)+1} L(t, T_{i(k)}, T_{i(k)+1}) \\ & \quad \times \exp \left(\int_t^{T_{i(k)}} \mu_{i(k)}(u) du \right) \end{aligned} \quad (32)$$

The convexity adjustments $\exp \left(\int_t^{T_{i(k)}} \mu_{i(k)}(u) du \right)$ can either be computed by Black–Scholes formula, or by replication. Alternatively, a term structure model such as a LIBOR market model (see **LIBOR Market Model**) or a Cheyette model (see **Markovian Term Structure Models**) could be used to obtain the required drifts. (See also **Change of Numeraire**; **Forward and Swap Measures**; **Convexity Adjustments**.)

References

- [1] Amblard, G. & Lebuchoux, J. (2000). *The Relationship Between CMS Options and the Smile*, Risk Magazine.
- [2] Benhamou, E. (2000). *Pricing Convexity Adjustment with Wiener Chaos*. Working paper available from www.ssrn.com.

- [3] Carr, P., Lewis, K. & Mandan, D. (2000). *On the Nature of Options*. working paper 2000, available from <http://www.math.nyu.edu/research/carrp/papers>.
- [4] Coleman, T. (1995). *Convexity Adjustments for CMS and Libor in Arrears Basis Swaps*. Working paper.
- [5] El Karoui, N., Geman, H. & Rochet, J.C. (1995). Changes of numeraire, changes of probability measures and option pricing, *Journal of Applied Probability* **32**, 443–458.
- [6] Hunt, J.P. & Kennedy, J.F. (2000). *On Convexity Corrections*. Working Paper, ABN-AMRO Bank and Warwick University.
- [7] Hunt, J.P. & Kennedy J.F. (2000). *Financial Derivatives in Theory and Practice*, Wiley, Chichester.
- [8] Pelsser, A. (2000). *Efficient Methods for Valuing Interest Rate Derivatives*, Springer Finance, Heidelberg.
- [9] Pugachewsky, D. (2001). Forward CMS rate adjustment, *RISK* March, 125–128.

Further Reading

- Amin, K. & Jarrow, R. (1992). Pricing options on risky assets in a stochastic interest rate economy, *Mathematical Finance* **2**, 217–237.
- Black, F & Scholes M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**(1), 637–654.
- Flesaker, B. (1993). Arbitrage free pricing of interest rate futures and forward contracts, *Journal of Futures Markets* **13**, 77–91.
- Hull, J. (2007). *Options, Futures and other Derivatives*, Prentice Hall India.
- Jamshidian, F. (1993). Option and Future evaluation with deterministic volatilities, *Mathematical Finance* **3**, 149–159.
- Musiela, M & Rutkowski, M. (1998). *Martingale Methods in Financial Modelling*, Springer Verlag.

Related Articles

Black–Scholes Formula; Convexity Adjustments; Forward and Swap Measures; LIBOR Rate; Markovian Term Structure Models.

LAURENT VEILEX

Bermudan Swaptions and Callable Libor Exotics

In this article we give an overview of callable interest rate products, a product class that includes Bermudan swaptions as well as callable Libor exotics (CLE). The main contractual features of these securities are discussed, and an outline of valuation algorithms is provided. CLEs are among the most complicated fixed income exotics derivatives traded, combining state-dependent coupons with other features such as a Bermudan-style callability option. As a consequence, these securities are particularly challenging to price and risk-manage.

Market Overview

CLEs most often start their life as callable structured notes. In a callable structured note, a *note issuer* receives the principal from an *investor* and pays a coupon, which is typically a function of interest rates, in return. In addition, the issuer retains the right to *call*, or *cancel*, the note on particular days, often coupon fixing days, after some initial *lock-out* (or *noncall*) period. When a note is called by the issuer, the principal is returned to the investor and the issuer stops paying the coupons.

As the principal received by the issuer is invested and typically yields a Libor rate (plus or minus a spread), a cancelable note is equivalent to a *cancelable swap*, that is, a swap of structured coupons against Libor rate payments that can be canceled on specified dates. Alternatively, it can be viewed as a noncancelable swap plus an option to *enter*, on a given set of dates, into a reverse swap. A CLE is typically defined to be such a *Bermudan-style option* to enter an exotic swap.

An investor granting the issuer the right to cancel the structured note is essentially selling a Bermudan-style right to enter an exotic swap. In return, the issuer compensates by making the coupon more attractive, which is often the motivation for adding the callability feature in the first place.

Types of Structured Coupons

CLEs are defined by the structured coupon they are paying, plus other features such as callability. In this

section we review some common types of coupons that one can encounter in a CLE. Coupons are usually structured to let the investors take speculative views on the interest rate market or to provide protection against adverse scenarios. For example, a floored payoff can be chosen if an investor desires protection against a fall in interest rates.

A structured coupon can be any function of observed interest rates (such as Libor or constant maturity swap (CMS); *see* **LIBOR Rate; Constant Maturity Swap**) on or before its fixing date. Many types are available. They can roughly be split into four categories: Libor-based, CMS-based, spread-based, and range-accruals.

To provide some detail, we assume in the following sections that a tenor structure is defined:

$$0 = T_0 < T_1 < \dots < T_N, \quad \tau_n = T_{n+1} - T_n \quad (1)$$

Typically, the rate observations, or fixings, for the n th coupon are made at T_n , and the coupon pays at T_{n+1} , $n = 1, \dots, N - 1$.

Libor-based Exotic Swaps

A Libor-based structured coupon is a function of a single Libor rate. Let us denote the n th coupon as a function of the n th Libor rate observed on its fixing date by

$$C_n = C_n(L_n(T_n)) \quad (2)$$

There is a large variety of structured coupons $C_n(\cdot)$ that can be used. For example^a

- *Standard swap*: For fixed coupon k ,

$$C_n(x) = k \quad (3)$$

- *Capped and floored floaters*: For strike s , gearing g , cap c , and floor f ,

$$C_n(x) = \max(\min(g \times x - s, c), f) \quad (4)$$

- *Capped and floored inverse floaters*: For spread s , gearing g , cap c , and floor f ,

$$C_n(x) = \max(\min(s - g \times x, c), f) \quad (5)$$

- *Digitals*: For strike s and coupon k ,

$$C_n(x) = k \times 1_{\{x > s\}} \quad (6)$$

- “Flip-flops” or “tip-tops”: For strike s and two coupons, k_1 and k_2 , “low” and “high” coupons, the coupon is

$$C_n(x) = \begin{cases} k_1, & x \leq s \\ k_2, & x > s \end{cases} \quad (7)$$

Different coupon types can be added together to create new types of structured coupons.

CMS-based Coupons

The same payoffs can be applied to CMS rates. Structured coupons are then deterministic functions of CMS rates. If an m -period rate is used, and we denote by $S_{n,m}(\cdot)$ the forward swap rate that fixes at T_n and covers m periods, then a structured coupon for period n is defined by

$$C_n = C_n(S_{n,m}(T_n)) \quad (8)$$

with $C_n(x)$ as defined in the previous section.

Spread-based Coupons

Spread-based structured coupons differ from Libor- or CMS-based ones in that they involve more than one market rate, Libor or CMS, in the calculation of structured coupons. The most common example is a CMS spread coupon. Let $S_{n,a}(\cdot)$ and $S_{n,b}(\cdot)$ be two collections of CMS rates, fixing on T_n , $n \geq 0$, and covering a and b periods, respectively. A CMS spread coupon with gearing g , spread s , cap c , and floor f is then defined by

$$C_n = \max(\min(g \times (S_{n,a}(T_n) - S_{n,b}(T_n)) + s, c), f) \quad (9)$$

A more general example can be obtained by using one of the payoff functions $C_n(x)$ defined in section Libor-based Exotic Swaps when applied to the spread $x = S_{n,a}(T_n) - S_{n,b}(T_n)$. In particular, digital and flip-flop CMS spread swaps are popular.

Spread-based exotic swaps typically cannot be decomposed into “standard” instruments, such as standard swaps, caps, and so on. Therefore, they, as a rule, cannot be valued by replication arguments, and a model is required.

More than two rates can be used in the definition of a coupon. Relatively recently, the so-called “curve

caps” became popular. In a curve cap, a coupon rate (Libor or CMS) is capped at a level that is given by a function (typically a spread) of two (other) rates. Often, the coupon rate is also floored at (another) function of two (yet other!) rates. Thus, the definition of a curve cap coupon can involve up to five rates. If the coupon rate is itself a spread, six rates define the coupon.

Range-accruals

A range-accrual structured coupon is defined as a fixed (or, sometimes, floating; see below) rate that only accrues when a reference rate is within a certain range. Let $X(t)$ be such a reference rate, and let l be the low bound and u be the upper bound. Let k be a fixed rate. A range-accrual coupon pays

$$C_n = k \times \frac{\#\{t \in [T_n, T_{n+1}] : X(t) \in [l, u]\}}{\#\{t \in [T_n, T_{n+1}]\}} \quad (10)$$

where $\#\{\cdot\}$ is used to denote the number of days that a given criterion is satisfied.

The fixed coupon k can be replaced with, for example, a Libor rate or, much less commonly, a CMS rate or any other structured coupon. Similarly, the reference rate $X(t)$ can be any market-observable rate such as a Libor rate fixing at t , a CMS rate fixing at t , or even a CMS spread.

A fixed-rate range-accrual coupon can be decomposed into simpler coupons, because

$$\begin{aligned} & \#\{t \in [T_n, T_{n+1}] : X(t) \in [l, u]\} \\ &= \sum_{t \in [T_n, T_{n+1}]} 1_{\{X(t) \in [l, u]\}} \end{aligned} \quad (11)$$

The sum on the right-hand side is over all business days in the period $[T_n, T_{n+1}]$. Thus, the range-accrual coupon can be seen as a collection of digitals on the reference rate. A similar decomposition can be applied to floating range-accruals.

Bermudan Swaptions and Callable Libor Exotics

Definition

A *Bermudan swaption* is a Bermudan-style option to enter a fixed-for-floating swap, that is, a swap with

the “structured” coupon paying fixed rate, $C_n(x) = k$. Bermudan swaptions on payer (pay-fixed, receive-floating) swaps are called *payer Bermuda swaptions*; similar conventions hold for receiver swaps. Bermudan swaptions are actively traded in both the US and European markets, with well-developed interdealer market in the United States for shorter-dated Bermudan swaptions.

If the underlying swap is an exotic swap, that is, a swap paying a structured coupon as in the previous section, then the Bermuda-style option to enter it is called a *CLE*.

For a CLE on an exotic swap with structured coupons $\{C_n\}_{n=0}^{N-1}$, we denote the value of the exotic swap that one can exercise on date T_n , the so-called *exercise value*, by

$$E_n(t) = \beta(t) \sum_{i=n}^{N-1} \tau_i \mathbf{E}_t(\beta^{-1}(T_{i+1}) \times (C_i - L_i(T_i))),$$

$$t \leq T_n \quad (12)$$

where β is a numeraire and E_t denotes time t expectation in the probability measure associated with β (see **Forward and Swap Measures**).

Types of Callable Libor Exotics

Any exotic swap can be used as an underlying for a CLE. The “taxonomy” of CLEs follows closely that of structured coupons; see the section Types of Structured Coupons. As we already mentioned, the simplest type of a CLE is a Bermudan swaption, with the underlying being a fixed-for-floating swap. We can generally distinguish four types of CLEs, Libor-based, CMS-based, spread-based, and callable range-accruals. If the underlying is an inverse floating swap, the CLE is called a *callable inverse floater*, and so on.

CLEs Accreting at Coupon Rate

Typically the notional of the underlying swap of a CLE is the same throughout the life of the deal, but occasionally this is not the case. A notional can vary deterministically by, for example, increasing or decreasing by the same amount or at a certain rate each period. Such deterministic accretion rarely adds extra complications from a modeling prospective. Sometimes, however, a contract specifies that the

notional of the swap increases, or accretes, at the structured coupon rate. As the structured coupon is not known in advance, the accretion rate is not deterministic. For such CLEs, definition (12) has to be amended. For this, let q_i be the notional to be applied to the coupon paid at the end of the period $[T_i, T_{i+1}]$. q_i is obtained from the notional over the previous period q_{i-1} by multiplying it by the structured coupon over the previous period. Formally,

$$E_n(t) = \beta(t) \sum_{i=n}^{N-1} \tau_i \mathbf{E}_t(\beta^{-1}(T_{i+1}) \times q_i \times (C_i - L_i(T_i))),$$

$$q_i = q_{i-1} \times (1 + \tau_{i-1} C_{i-1}) \quad (13)$$

The initial notional q_0 is contractually specified.

Snowballs

In a *snowball*, the structure coupon is not just a function of the interest rate, but of the previous coupon as well. The most common snowball is of inverse floating type. In particular, the n th coupon C_n is defined by

$$C_n = (C_{n-1} + s_n - g_n \times L_n(T_n))^+ \quad (14)$$

for $n = 1, \dots, N-1$ (with C_0 usually being a simple fixed-rate coupon). Here $\{s_n\}$ and $\{g_n\}$ are contractually specified deterministic sequences of spreads and gearings. With this particular type of coupon, a snowball is sometimes called a *callable inverse ratchet*.

Many variants on the snowball idea have appeared recently, all variations on the theme of using a previous coupon in the definition of the current one. For example, a *snowrange* is typically defined by

$$C_n = C_{n-1} \times \frac{\#\{t \in [T_n, T_{n+1}] : X(t) \in [l, u]\}}{\#\{t \in [T_n, T_{n+1}]\}} \quad (15)$$

Multitranches

As we discussed previously, the more optionality an investor can sell to the issuer, the better coupon he/she can receive. The option to call the note is already present in a callable structured note. Another

option that is sometimes embedded is the right for the issuer to increase the size of the note, or put more of the same note to the investor, whether he/she wants it or not. The name of this feature, a *multitranche* callable structured note, comes from the fact that these possible notional increases are formalized as tranches of the same note that the issuer has the right to put to the investor.

The times when the issuer has the right to increase the notional of the note typically come before the times when the issues can cancel the note altogether. Callability usually applies jointly to all “tranches” of the note.

Valuation of Callable Libor Exotics

The valuation of multicable securities, such as Bermudan or American options, requires a solution of the optimal exercise problem, a type of a dynamic programming problem (*see Binomial Tree*). If we denote by $H_n(t)$ the value, at time t , of a CLE that has only the dates $\{T_{n+1}, \dots, T_{N-1}\}$ as exercise opportunities, then the value of a CLE, the value $H_0(0)$, is defined recursively by

$$\begin{aligned} H_{n-1}(T_{n-1}) &= \beta(T_{n-1}) \mathbf{E}_{T_{n-1}} (\beta(T_n)^{-1} \\ &\quad \times \max\{H_n(T_n), E_n(T_n)\}), \\ n &= N-1, \dots, 1, \\ H_{N-1} &\equiv 0 \end{aligned} \tag{16}$$

Bermudan swaptions are the most liquid interest rate exotics. As their values are mostly derived from up- and down-moves of the yield curve, they are often valued in low-dimensional short-rate models such as developed in **Term Structure Models**, using lattice or PDE numerical methods (*see* articles in **Partial Differential Equations**). The value of the “switch” option built into a Bermudan swaption, that is, the option to change which swap is entered into, can be captured by a model parameter responsible for rate autocorrelations, such as mean reversion in a

one-factor Gaussian model, *see* [1]. To value other types of CLEs, a more involved setup is usually required. Complicated dependence of coupons on one or more underlying rates, combined with the need to account for callability, typically necessitates an application of a realistic, multifactor model of interest rates such as the Libor market model (*see LIBOR Market Model*). Owing to the high dimension of such a model, Monte Carlo methods become a necessity (*see LIBOR Market Models: Simulation*) and must be complemented by methods for estimating exercise boundaries of Bermudan-style options in Monte Carlo simulation (*see Bermudan Options; Exercise Boundary Optimization Methods; Early Exercise Options: Upper Bounds*). Additional details are available in [2].

End Notes

^aNote that in all definitions we ignore day counting fractions.

References

- [1] Andersen, L.B. & Andreasen, J. (2001). Factor dependence of Bermudan swaption prices: fact or fiction? *Journal of Financial Economics* **62**, 3–37.
- [2] Piterbarg V.V. (2005). Pricing and hedging callable Libor exotics in forward Libor models. *Journal of Computational Finance* **8**(2), 65–117.

Related Articles

Bermudan Options; Binomial Tree; Early Exercise Options: Upper Bounds; Exercise Boundary Optimization Methods; LIBOR Market Model; LIBOR Market Models: Simulation.

LEIF B.G. ANDERSEN & VLADIMIR V.
PITERBARG

Trigger Swaps

We discuss two classes of exotic interest rate derivatives: trigger swaps and targeted accrual redemption notes (TARNs). Subsequently, we outline possible valuation methodologies and assess the impact on the quality of risk estimates.

A trigger swap features an underlying swap. Cash flows of this underlying swap only start to be exchanged from the moment that a certain trigger condition is met. An example trigger condition is that some trigger index should be within a prespecified range. These are barrier type options, *see* **Barrier Options**.

The above-described trigger swap is of *knock-in* type, as opposed to a *knock-out* feature with which a swap is cancelled when the trigger index ends up in a prespecified range. There is a simple relation between an underlying swap and its knock-in and knock-out variants:

$$\begin{aligned}\text{value of swap} &= \text{value of knock-in swap} \\ &\quad + \text{value of knock-out swap}\end{aligned}$$

The above relation shows that the ability to value knock-out swaps also yields the ability to value knock-in swaps. Hence in the remainder of the article, we restrict the discussion to knock-out swaps.

A TARN is a trigger swap where the trigger index is the cumulative sum of structured coupons paid, typically of knock-out type. In this case, the barrier is referred to as a *lifetime cap*.

Example term sheets of a knock-in swap and a TARN are displayed in Tables 1 and 2, respectively. Both term sheets feature structured coupons that are exchanged for IBOR plus spread—this is more or less the general rule. There are several variations often applied to the basic TARN design:

1. The form in Table 2 is referred to as *partial final coupon* in order to exactly attain the lifetime cap.
2. The latter is in contrast to a *full-final coupon* where an excess over the lifetime cap is allowed.
3. If the sum of coupons is below target at maturity, then a *lifetime floor* dictates that a make up coupon be paid to attain the target.

Valuation

We focus the discussion primarily on valuation methods, such as a Markovian grid or Monte Carlo simulation. The presentation is more or less independent of model choice, unless explicitly mentioned otherwise.

It is well known that simulation-based risk estimates for derivatives with discontinuous payoff are less efficient than for those with continuous payoff. In contrast, a Markovian grid naturally provides high-quality risk estimates, as the backward induction expectation operator has a smoothing effect—as long as interpolation is dealt with properly. Knock-out features can be easily valued on a grid as the value of a knock-out swap is simply 0 for those state nodes for which the trigger index is within the knock-out range. Path-dependent aspects such as the cumulative coupon trigger index of TARNs may be incorporated by adding a dimension to the Markovian grid that keeps track of the cumulative coupon.

Although a Markovian grid has excellent credentials, the alternative of simulation may also successfully be applied. Its drawback—as mentioned above—lies in less than efficient estimates of risk sensitivities. The remainder of the article outlines four techniques to improve risk estimates for trigger swaps in a simulation framework: (i) large shifts, (ii) importance sampling (IS), (iii) conditioning, and (iv) smoothing.

All of the described methods may also be found in the article overview paper of Piterbarg (2004, [1]).

Large Shift Sizes

Ordinary simulation, without further enhancements, remains a viable technique of obtaining risk estimates. In the general case, risk sensitivities are calculated *via* finite differences, for which we must specify a shift size. For continuous payoffs, it is more efficient to use “smaller” shift sizes (say of the order 10^{-8}). However, for discontinuous payoffs, this is not the case as small shift sizes mostly ignore the risk component due to the digital option embedded in the knock-out structure. With discontinuous payoffs, it is more efficient to use somewhat larger shift sizes, say of the order of 1 basis point.

Importance Sampling (IS)

IS is typically used to reduce the variance of the simulation estimate. (For an overview of variance

Table 1 Example knock-in swap

Product	CMS spread knock-out
Currency	CCY
Maturity	X years
Receive	Float funding plus margin
Pay	$Y \times (\text{CCY CMS10} - \text{CMS2})$ Annually, floor at 0% If on any day the spread $\text{CCY CMS10} - \text{CMS2} \geq B\%$ then from the next coupon period onwards a fixed rate of $Z\%$ knocks in until maturity of the trade

Table 2 Example TARN

Product	Guaranteed inverse floater swap
Currency	CCY
Maturity	X years
Receive	Float funding plus margin
Pay	$Y - Z \times \text{CCY 6M IBOR in arrears}$ Semi-annually, floor at 0% If on any day the sum of structured coupons $\geq B\%$ then pay $B\% - [\text{sum of coupons}]$ and the trade cancels thereafter

reduction techniques, *see* **Variance Reduction**.) For knock-outs; however, the simulation value estimate is usually of desirable quality. It is, however, a side effect of IS that we are interested in. With IS, we sample conditional on knocking out or in. Realizations of these conditional samples then need to be multiplied by the likelihood ratio, which is a measure for the probability of the conditional event. These likelihood ratios smoothen the valuation and cause the discontinuity to disappear. Hence, we obtain efficient risk. See Glasserman and Staum [2].

Conditioning

Conditioning is specifically geared toward TARNs. With TARNs, each cash flow is contingent on a sum of index rates being within range. We focus on the valuation of one TARN cash flow, since we may value the whole TARN if we can value all cash flows. We halt the simulation when the penultimate index rate is determined. Subsequently, the remaining cash flow is a digital option with underlying rate the final index rate that determines whether or not we obtain the TARN cash flow. This digital option may

be valued analytically with a Black-type formula—the latter provides smoothing, from which we obtain more efficient risk estimates.

Smoothing

For knock-outs, we observe a trigger index x . If the trigger is outside of the survival range $[L, U]$, then the trade knocks out. We assume $L \leq U$. We consider the survival ratio A_i for each cash flow i . The survival ratio A_i is either 0 or 1. It evolves from cash flow to cash flow as follows:

$$A_{i+1} = A_i \times (1 - R_i^{\text{KO}}) \quad (1)$$

The quantity R_i^{KO} is the knock-out ratio. For the remainder of the presentation, we omit the cash flow index i for convenience. The knock-out ratio R^{KO} is either 0 or 1, and is given by:

$$R^{\text{KO}} = 1\{x < L\} + 1\{x > U\} \quad (2)$$

A smoothed version of the function in equation (2) is easily devised. An example is given in Figure 1. The smoothed function may be parameterized. For “small” parameters, the smoothed version is hardly distinguishable from the original discontinuous function. The use of “large” parameters leads to an intolerable bias in the derivative value. Experience and testing may provide guidance in selecting a parameter that bares acceptable bias yet sufficiently improves risk sensitivity estimates.

We test smoothing for a trade such as given in Table 1. We use 10^{-6} bps (“small”) shift for smooth and 0.5 bps (“large”) for nonsmooth. The terminology

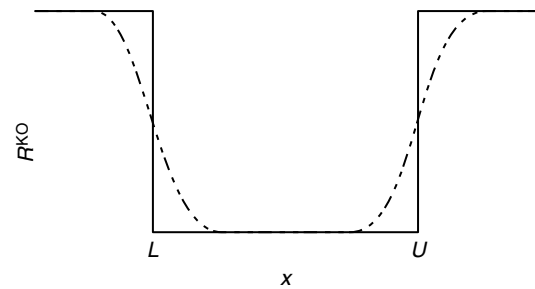


Figure 1 Knock-out ratio *versus* trigger index x , with lower and upper knock-out barriers L and U . Discontinuous, according to contract (full line) and smoothed (dotted line)

Table 3 P&L prediction results

	-10 bps	-1 bps	0 bps	1 bps	10 bps
Nonsmooth, large shifts					
NPV	1 005 251	994 113	992 594	990 804	981 537
P&L	-1266	-1519		-1790	-1106
Pred	-1597	-2042		-1871	-1772
Gap	331	523		82	667
Smooth, small shifts					
NPV	1 002 928	991 884	990 656	989 429	978 420
P&L	-1227	-1228		-1227	-1224
Pred	-1226	-1228		-1227	-1223
Gap	-1	0		0	-1

“Pred” means predicted P&L. Market data and other data used for this test may be obtained from the author upon request

is as follows: NPV, trade value. P&L, realized profit and loss per basis point up-shift. Predicted P&L: the average of the delta at two points: at the associated shift size and for the unperturbed case. Gap: predictability gap, P&L – Predicted P&L. The results are displayed in Table 3.

References

- [1] Piterbarg, V.V. (2004). TARNs: models, valuation, risk sensitivities, *Wilmott* **14**, 62–71.
- [2] Glasserman, P. & Staum, J. (2001). Conditioning on one-step survival for barrier option simulations, *Operations Research* **49**, 923–937.

RAOUL PIETERSZ

CMS Spread Products

According to recent estimates, the volume of traded contracts involving euro-denominated options on the spread between two *constant maturity swap* (CMS) rates (see **Constant Maturity Swap**) amounted to €240 billions in 2007 in the interbank market alone. This makes CMS spread options the most rapidly growing interest-rate derivatives market.

The underlying of any option on CMS spread is the CMS rate. By definition, the CMS rate that fixes and settles at a generic time T (associated to a swap of maturity T_n and starting at T) is equal to the swap rate of the associated swap. At any time prior to T , the value of that rate is then formally given by $CMS(t) = E_t^T(S(T, T_n))$, by absence of arbitrage. Here, the expectation $E_t^T(\cdot)$ is intended at time t with respect to the T —*forward measure* Q^T (see **Forward and Swap Measures**), where the zero coupon bond $B(t, T)$ is the associated *numéraire*. The variable leg of a CMS swap pays a stream of CMS rates at any settlement date. In turn, the payout $f_T = f(CMS_1(T), CMS_2(T); K)$ of a European call (respectively, put) option on a CMS spread expiring at time T reads as $f_T = (CMS_1(T) - CMS_2(T) - K)^+$, (respectively $f_T = (K - CMS_1(T) + CMS_2(T))^+$), where we have defined, as usual, $(\cdot)^+ = \text{Max}(\cdot, 0)$.

Generally speaking, CMS spread options are simple and liquid financial instruments that allow to take a view on the future shape of the yield curve (or immunizing a portfolio against it). In the most commonly traded combination, we have that $CMS_{1,2}(T) = CMS_{10Y,2Y}(T)$ (i.e., the two CMS rates are associated with a 10Y and 2Y swap, respectively). The buyer of a call (respectively, put) CMS spread option will then benefit from a future curve steepening (respectively, flattening) scenario. More complex option strategies involving CMS spreads are actively traded in the over-the-counter (OTC) market, as well. They include, to name a few, digitals, barrier options, and Bermudan-style derivatives.

Option Pricing without Smiles

The fair value at time t of the most generic call option is formally given by $C(t, T; K) = B(t, T)E_t^T(\alpha_1 CMS_1(T) - \alpha_2 CMS_2(T) - K)^+$, by arbitrage,

with α_1 and α_2 being constants. This expression is obviously reminiscent of an option on the spread between two assets and reduces to a simple *exchange option* (see **Exchange Options**) when $K = 0$. In the simplest case, one assumes that each CMS rate follows a simple arithmetic or geometric Brownian motion under the relevant martingale measure. In the former case, a closed-form formula for $C(t) = C(t, T; K)$ can be given [3], while in the latter the price can be only expressed in integral form unless $K = 0$, in which case a closed-form formula can be exhibited [9]. Some authors propose to use the first approach as an approximation for the second one [3, 11] for a generic $K \neq 0$. One must be, however, warned against these oversimplifications as market bid/offer spreads are relatively tight. Further risk sensitivities are very different in the two settings with profound implications as long as portfolio replication quality is concerned.

Differently from the single asset case, the process difference of two asset prices is now allowed to take negative values. Therefore, the arithmetic Brownian motion framework is generally considered as the simplest viable approach. Since, by definition, the CMS rate is a Q^T -martingale, we assume that the two rates $CMS_{1,2} = X_{1,2}$ evolve according to the following Gaussian processes under Q^T : $dX_{1,2}(t) = \sigma_{1,2} dW_{1,2}^T(t)$ with constant volatility $\sigma_{1,2}$ and where $d\langle W_1^T, W_2^T \rangle_t = \rho dt$ for some constant correlation coefficient ρ . In this case, it is easy to verify that the price of the option $C(t)$ is given by the modified *Bachelier formula* (see **Bachelier, Louis (1870–1946)**) $C(t) = B(t, T)[\sigma\sqrt{\tau}n(d(F_t, \tau)) + (F_t - K)N(d(F_t, \tau))]$ where $\tau = T - t$, $F_t = \alpha_1 X_1(t) - \alpha_2 X_2(t)$, $\sigma^2 = \alpha_1^2 \sigma_1^2 - 2\rho\alpha_1\alpha_2\sigma_1\sigma_2 + \alpha_2^2 \sigma_2^2$ and $d(F_t, \tau) = (F_t - K)/(\sigma\sqrt{\tau})$. Here, $n(\cdot)$ and $N(\cdot)$ stand for the standard Gaussian density and cumulative distribution function, respectively. In the lognormal case, one has to resort to a quasi-closed form formula (see [3] for a review).

The advantage of the above formula, similarly to Black–Scholes (BS) model (see **Black–Scholes Formula**) for European options on single assets, is its simplicity. However, while in the BS case inverting the market price provides a unique *implied volatility*, here the situation is more complex. There are now three (as opposed to one) free parameters of the theory, that is, ρ , σ_1 , and σ_2 . In a perfectly liquid market, one could, in principle, infer σ_1 and σ_2 by inverting the Bachelier formula for the two respective

options on $CMS_{1,2}$, and then use the correlation as the unique free parameter of the theory. Interestingly, this indicates that buying or selling spread options is, in principle, equivalent to trading *implied spread correlation*.

Unfortunately, the above approach relies on the assumption that CMS rates dynamics are well modeled by an arithmetic Brownian motion. In practice, this is not the case. The main reason has to do with the presence of the volatility smile rather than with the request of positivity of CMS rates. As it is well known, a CMS rate settling at time T , and associated to a swap of length τ , can be statically replicated through a linear combination of European swaptions (of different strike) expiring at T to enter into a swap of length τ . The sum is actually infinite, that is, it is an integral over all possible swaptions for that given maturity [1, 6]. Because it is well known that implied swaption volatilities are different at different strikes (i.e., a volatility smile is present) it means that the swaption underlying—the forward swap rate—cannot follow a simple Gaussian process in the relevant martingale measure. Consequently, the CMS rate, viewed as a linear combination of swaptions, must evolve accordingly. Needless to say that using a model for spread options where the underlying process is inconsistent with the market available information on plain-vanilla instruments has important consequences on the quality of the risk management [4, 5].

Option Pricing with Smiles

There are essentially three possible ways to quote CMS spread options so as to ensure partial or full consistency with the underlying (CMS) implied dynamics.

Stochastic volatility models are very popular among academics and practitioners as they provide a simple and often effective mechanism of static generation as well as dynamic smile evolution [10]. The first approach consists of assuming that each CMS rate in the spread follows a diffusion with its own stochastic volatility. The Stochastic Alpha Beta Rho (SABR) model, for instance, has become the market standard for European options on interest rates [7]. By coupling two SABR diffusions, one can easily calibrate each parameter set on the respective market-implied CMS smile. The method has, however, two

major drawbacks. First, neither known formula nor simple approximation exists on options for multivariate SABR models. Second, there are six independent correlations to specify and several among them are not directly observable (e.g., the correlation between the first CMS rate and the volatility of the second one). In addition, it is easy to verify that some of those parameters are fully degenerate with respect to the price of a spread option of given strike.

The second approach resorts to using arbitrage-free dynamic models for the whole yield curve dynamics, in the Heath-Jarrow-Morton (HJM) sense [8], *see Heath–Jarrow–Morton Approach*. Dynamics of the spread between any two CMS rates is then inferred from dynamics of the whole curve. This second method allows pricing and risk managing all spread options on different pairs (e.g., 10Y–2Y, 10Y–5Y, 30Y–10Y, etc.) within a unique modeling setup rather than treating them as separate problems. This offers the great advantage of measuring and aggregating correlation exposures across all pairs at once and correlation risk diversification can be achieved. Also, exotic derivatives can be priced in this framework. On the negative side, it is very difficult to reproduce the implied smile of each CMS rate unless very complex models are introduced (e.g., a multifactor HJM model with possibly multivariate stochastic volatility).

Finally, a third possibility consists of disentangling the marginal behavior and the dependence structure between the two CMS rates. One can infer the marginal probability density from plain-vanilla swaptions, that is, match their respective individual smiles, and then “recombine” them *via a copula-based method* [2], and references therein) to get the bivariate distribution function. In [2], a simple numerical trick to reduce the dimensionality of the necessary integration routines is also described. The great advantage of this approach is its simplicity and the guarantee that, by construction, the price of the spread option is, at a given time, consistent with the current swaption market. On the negative side, the approach is purely static, since no simple method exists to assign a dynamics on a bivariate process such that the associated density is consistent with the chosen copula function at any time. In addition, the choice of the copula itself is, to a large extent, arbitrary.

Implied Correlation and Normal Spread Volatility

Similar to the BS case, practitioners often prefer to measure and compare spread option prices through homogeneous quantities. For simple options, people use implied volatility. For spread options, the natural equivalent is the concept of implied correlation.

Assume that a spread option is being priced through a Gaussian copula-based method. Put simply, this amounts to inferring the two CMS marginal densities from each respective swaption market and then coupling them *via* a Gaussian copula function. It should be mentioned that a Gaussian copula is parameterized by a single correlation ρ and that a spread option price is monotonically decreasing as a function of ρ . Therefore, given the market price of a generic call spread option $C(t, T; K)$ struck at K , and given the two marginal CMS underlying densities, there exists a unique $\rho(K)$ such that the market price is matched by a copula method with correlation $\rho(K)$. This unique number is termed *implied copula correlation*. As for simple options, the function $\rho(K)$ displays a significant dependence on the strike. This is the correlation smile phenomenon (Figure 1).

Interestingly, it is possible to analyze the situation from a different, albeit similar, angle. In a previous

section, we showed that the simplest way to price options on CMS spread consists of coupling two simple Gaussian processes. The resulting closed-form formula is of Bachelier type with a modified normal volatility given by $\sigma^2 = \alpha_1^2 \sigma_1^2 - 2\rho\alpha_1\alpha_2\sigma_1\sigma_2 + \alpha_2^2 \sigma_2^2$. Given the option price $C(t, T; K)$, one can then invert the Bachelier-like formula to get a unique *implied normal spread volatility* $\sigma(K)$. Once more, function $\sigma(K)$ displays a smile. This alternative approach is still very popular among some practitioners.

It must be noticed, however, that the two above smile generation methods are not equivalent. In fact, only the first one is fully consistent with the underlying swaption smile observed in the market. In addition, the former approach concentrates on correlation, while the latter on the normal spread volatility that corresponds to the covariance of the joint process. Therefore, the first method is better suited if one considers volatility and correlation markets as evolving separately so that correlation movements are partly unrelated to price changes for swaptions. On the other side, the second method assumes that correlation and volatility markets are essentially indistinguishable to the extent that only the product of volatility and correlation (i.e., the

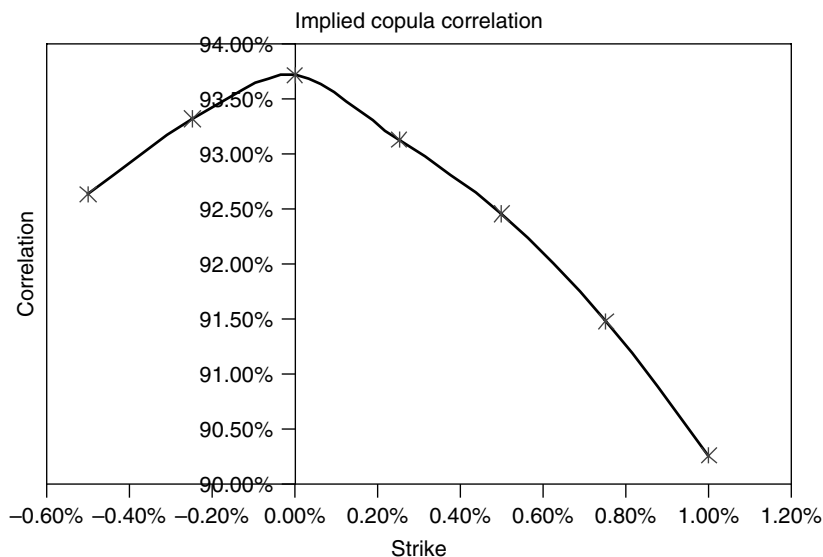


Figure 1 The figure displays the typical pattern of the implied copula correlation smile associated to a contract on the 10Y–2Y CMS spread. Volatility is associated to a 7×10 cap on CMS spread, starting 7 years from today and maturing 10 years from today. [Source: BNP Paribas]

covariance) is the relevant quantity as far as risk management is concerned.

CMS Spread Derivatives

In recent times, a whole range of financial products has been introduced where the underlying is a spread between two CMS rates. In the broker market, option on CMS spreads are usually quoted as a cap/floor comprising a given number of caplets/floorlets with quarterly frequency. This is completely equivalent to caps/floors written on LIBOR rates. Plain options on spreads are also commonly traded in the interbank market (mainly on maturities less than three years). These are referred to as *single-look* (or *one-look*) options on CMS spread.

Investors can benefit by taking a view on the future evolution of the difference between two swap rates, that is, by taking a view on the shape of the yield curve. This can be best captured by structuring financial derivatives written on the spread. Here, we limit ourselves to briefly mention a few important examples.

A digital option on a CMS spread gives the purchaser the right to receive one unit of notional should the spread exceed the strike price at expiry and nothing otherwise. For instance, an investor who believes that the level of the current 10Y swap rate will always be larger than the 2Y, one can buy a digital (call) option on the 10Y–2Y CMS spread with strike set to 0. Similarly, a range accrual option on the same spread gives the holder the right to receive one unit of notional times the number of days the 10Y swap exceeds the 2Y swap between two given dates. Both digital and range accrual options are relatively simple instruments. More complex derivatives written on CMS spreads have also appeared in the market and enjoyed vast success: Bermudan options, target redemption notes (TARNs), knockout options as well as multiple underlying derivatives, to name a few. A simple example of multiple underlying derivatives is a digital option that pays a variable coupon (e.g., a LIBOR rate) if the spread between two CMS rates fixes above a given strike.

It is worth reminding that the pricing and risk management of all structured derivatives on CMS spreads are, by construction, heavily dependent on the correlation smile. Further, for Bermudan and

path-dependent derivatives, they also depend on the yield curve/forward volatility assumptions that are associated to a given model for interest rates dynamics.

Acknowledgments

Author Olivier Scaillet thanks the Swiss NSF for financial support through the NCCR Finrisk.

References

- [1] Amblard, G. & Lebuchoux, J. (1999). *Smile-Consistent Pricing of CMS Contracts*.
- [2] Berrahoui, M. (2004). Pricing CMS spread options and digital CMS spread options with smile, *Wilmott Magazine*, 63–69.
- [3] Carmona, R. & Durrleman, V. (2003). Pricing and hedging spread options, *SIAM Review* **45**, 627–685.
- [4] Galluccio, S. & Di Graziano, G. (2007). On model selection and its impact on the hedging of financial derivatives, in *Advances in Risk Management*, G. Gregoriou, ed, Palgrave MacMillan.
- [5] Galluccio, S. & Di Graziano G. (2007). *Model Misspecification for General Processes: Theory and Practice*.
- [6] Hagan, P. (2003). Convexity conundrums: pricing CMS swaps, caps and floors, *Wilmott Magazine*, 38–44.
- [7] Hagan, P., Lesniewski, A., Kumar, D. & Woodward, D. (2002). Managing smile risk, *Wilmott Magazine*, 84–108.
- [8] Heath, D., Jarrow, R. & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation, *Econometrica* **60**, 77–105.
- [9] Margrabe, W. (1978). The value of an option to exchange one asset for another, *Journal of Finance* **33**, 177–186.
- [10] Musiela, M. & Rutkowski, M. (2004). *Martingale Methods in Financial Modelling*, 2nd Edition, Springer-Verlag, Berlin.
- [11] Poitras, G. (1998). Spread options, exchange options and arithmetic Brownian motion, *Journal of Futures Markets*, **18**, 487–517.

Related Articles

Bermudan Swaptions and Callable Libor Exotics; LIBOR Rate; Swap Market Models.

STEFANO GALLUCCIO & OLIVIER SCAILLET

Yield Curve Construction

The objective of an interest rate model is to describe the random movement of a curve of zero-coupon bond prices through time, starting from a known initial condition. In reality, however, only a few short-dated zero-coupon bonds are directly quoted in the market at any given time, a long stretch from the assumption of many models that an initial curve of zero-coupon bond prices is observable for a continuum of maturities. Fortunately, a number of liquid securities depend, in relatively straightforward fashion, on zero-coupon bonds, opening up for the possibility of uncovering zero-coupon bond prices from prices of such securities. Still, as only a finite set of securities are quoted in the market, constructing a continuous curve of zero-coupon bond prices will require us to complement market observation with an interpolation rule, based perhaps on direct assumptions about functional form or perhaps on a regularity norm to be optimized on. A somewhat specialized area of research, discount curve construction relies on techniques from a number of fields, including statistics and computer graphics. We discuss the basic topics in detail and refer the reader to appropriate sources for advanced applications. In the same spirit, we pay scant attention to the subtle intricacies of actual swap and bond market conventions.

Discount Curves

Let $P(t, T)$ be the time t price of a zero-coupon bond maturing at time T . Going forward, we use the abbreviated notation $P(T) = P(0, T)$ where $P : [0, T] \rightarrow (0, 1]$ is a continuous, monotonically declining *discount curve*. T denotes the maximum maturity considered, typically given as the longest maturity in the set of securities the curve is built to match. Let there be N such securities—the *benchmark set*—with observable prices $V_1, \dots, V_N$. We fundamentally assume that the time 0 price $V_i = V_i(0)$ of security i can be written as a linear combination of zero-coupon bond prices at different maturities,

$$V_i = \sum_{j=1}^M c_{ij} P(t_j), \quad i = 1, \dots, N \quad (1)$$

where $0 < t_1 < t_2 < \dots < t_M \leq T$ is a given finite set of dates, in practice obtained by merging together

the cash-flow dates of each of the N benchmark securities. Securities that satisfy relationship (1) include coupon bonds, forward rate agreements (FRAs), as well as fixed-floating interest rate swaps. For instance, consider a newly issued unit-notional fixed-floating swap, paying a coupon of $c\tau$ at times $\tau, 2\tau, 3\tau, \dots, n\tau$. If no spread is paid on the floating rate, the time 0 total swap value V_S to the fixed-rate payer is

$$V_S = 1 - P(n\tau) - \sum_{j=1}^n c\tau P(j\tau) \Rightarrow$$

$$1 - V_S = P(n\tau) + \sum_{j=1}^n c\tau P(j\tau) \quad (2)$$

which is in the form (1) once we interpret $V_i = 1 - V_S$.

The choice of the securities to be included in the benchmark set depends on the market under consideration. For instance, to construct a treasury bond curve, it is natural to choose a set of treasury bonds and T-bills. On the other hand, if we are interested in constructing a discount curve applicable for bonds issued by a particular firm, we would naturally use bonds and loans used by the firm in question. In many applications, the most important yield curve is the *Libor curve*, constructed out of market quotes for Libor deposits, swaps, and Eurodollar futures. In the construction of this curve, most firms would use a few certificates of deposit for the first 3 months of the curve, followed by a strip of Eurodollar futures^a (with maturities staggered 3 months apart) out to 3 or 4 years. Par swaps are then used for the rest of the curve, often going out to 30 years or more.

Matrix Formulation & Transformations

Define the M -dimensional discount bond vector

$$\mathbf{P} = (P(t_1), \dots, P(t_M))^T \quad (3)$$

and let $\mathbf{V} = (V_1, \dots, V_N)^T$ be the vector of observable security prices. Also let $\mathbf{c} = \{c_{ij}\}$ be an $N \times M$ dimensional matrix containing all the cash flows produced by the chosen set of securities. $\mathbf{c}$ would typically be quite sparse.

In a friction-free market without arbitrage, the fundamental relation

$$\mathbf{V} = \mathbf{cP} \quad (4)$$

must be satisfied, giving us a starting point to find $\mathbf{P}$. In practice, however, we normally have $M > N$, in which case equation (4) is insufficient to uniquely determine $\mathbf{P}$. The problem of curve construction essentially boils down to supplementing equation (4) with enough additional assumptions to allow us to extract $\mathbf{P}$ and to determine $P(T)$ for values of T not in the cash-flow timing set $\{t_j\}_{j=1}^M$.

As it is normally easier to devise an interpolation scheme on a curve that is reasonably flat (rather than exponentially decaying), it is common to perform the curve-fitting exercise on *zero-coupon yields*, rather than directly on discount bond prices^b. Specifically, we introduce a continuous yield function $y : [0, T] \rightarrow \mathbb{R}_+$ given by

$$e^{-y(T)T} = P(T) \Rightarrow y(T) = -T^{-1} \ln P(T) \quad (5)$$

such that in equation (4) $\mathbf{P} = (e^{-y(t_1)T_1}, \dots, e^{-y(t_M)T_M})$. The mapping $T \mapsto y(T)$ is known as the *yield curve*; it is related to the discount curve by the simple transformation (5). Of related interest is also the *instantaneous forward curve* $f(T)$, given by

$$P(T) = e^{-\int_0^T f(u)du} \Rightarrow f(T) = y(T) + \frac{dy(T)}{dT}T \quad (6)$$

For more transformations of the discount curve—and a discussion of relative merits in curve construction—see [1, 14]. For simplicity, in this article we work primarily with $y(T)$ unless otherwise noted.

Construction Principles

We have at least three options for solving equation (4).

1. We can introduce new and unspanned securities such that $N = M$ and equation (4) allows for exactly one solution.
2. We can use a parameterization of the yield curve with precisely N parameters, using the N equations in equation (4) to recover these parameters.
3. We can search the space of all solutions to equation (4) and choose the one that is “optimal” according to a given criterion.

Let us provide some comments to these three ideas. First, in option 1, introduction of new securities might not truly be possible—such securities may simply not exist—but sometimes interpolation rules applied to the given benchmark set may allow us to provide reasonable values for an additional set of “fictitious” securities. Although it can occasionally be useful in preprocessing to pad an overly sparse benchmark set, this idea will often require some quite *ad hoc* decisions about the specifics of the fictitious securities, and excessive use may ultimately lead to odd-looking curves and suboptimal hedge reports. When an interpolation rule is to be used, it is typically better to apply it to fundamental quantities such as zero-coupon yields or forward rates, thereby maintaining a higher degree of control over the resulting yield curve.

In option 2 above, parametric functional forms such as that in [9] are sometimes used, but it is far more common to work with a spline representation with N user-selected knots (typically at the maturity dates of the benchmark securities), with the level of the yield curve at these knots constituting the N unknowns to be solved for. We discuss the details of this approach in the section Yield Curve Fitting with N -knot Splines, using a number of different spline types. We assume some knowledge of spline theory here—classical references are [2, 11].

Option 3 is covered in the section Nonparametric Optimal Yield Curve Fitting and constitutes the most sophisticated approach. It can often be stated in completely nonparametric terms, with the yield curve emerging naturally as the solution to an optimization problem. If carefully stated, this approach can be set up to also handle the situation where the system of equations (4) is (near-) singular, in the sense that either no solutions exist or all solutions are irregular and nonsmooth.

Yield Curve Fitting with N -Knot Splines

In this section, we discuss a number of yield Curve algorithms based on polynomial and exponential (tension) splines of various degrees of differentiability. Throughout, we assume that we can select and arrange our benchmark set of securities to guarantee that the maturities of the benchmark securities satisfy

$$T_i > T_{i-1}, \quad i = 2, 3, \dots, N \quad (7)$$

where the inequality is strict. Equation (7) constitutes a “spanning” condition and allows us to select the N deal maturities as distinct knots in our splines.

Bootstrapping

A simple—and widely used—approach involves assuming that the yield curve $y(T)$ is a continuous piecewise linear spline. This assumption allows for a straightforward solution procedure, known as *bootstrapping*. Specifically, we proceed according to the algorithm:

1. Let $y(t_j)$ be known for $t_j \leq T_{i-1}$, such that prices for benchmark securities $1, \dots, i-1$ are matched.
2. Make a guess for $y(T_i)$; linearly interpolate to find $y(t_j)$, $T_{i-1} < t_j < T_i$.
3. Compute V_i from the known values of $P(t_j)$, $t_j \leq T_i$.
4. If V_i equals the market value, stop. Otherwise return to equation (4).
5. If $i < N$, set $i = i + 1$ and repeat.

The updating of guesses when iterating over steps 4 and 6 can be handled by a standard one-dimensional root-search algorithm (e.g., the Newton–Raphson or secant methods).

With $y(T)$ being piecewise linear, the forward curve $f(T)$ (see equation 6) takes on a discontinuous saw-tooth shape, as shown in Figure 1. It may be tempting to replace the assumption of continuous piecewise linear yields with an assumption of continuous piecewise linear forwards, but such an interpolation rule turns out to be numerically unstable and prone to oscillations. However, we may

instead assume that the forward curve is piecewise flat—or, equivalently, that $\log P(T)$ is piecewise linear—which again allows for stable application of the bootstrapping principle. For reference, the forward curve resulting from this idea is also shown in Figure 1.

Catmull–Rom Splines

To ensure that the forward curve stays continuous, we need a yield curve that is at least once differentiable, that is, in C^1 . A common assumption involves setting $y(T)$ equal to a once differentiable Hermite cubic spline:

$$y(T) = a_{3,i}(T - T_i)^3 + a_{2,i}(T - T_i)^2 + a_{1,i}(T - T_i) + a_{0,i}, \quad T \in [T_i, T_{i+1}] \quad (8)$$

for a series of constants $a_{3,i}$, $a_{2,i}$, $a_{1,i}$, $a_{0,i}$ to be determined from exogenously given values of $y(T_i)$, $y(T_{i+1})$, $y'(T_i)$, and $y'(T_{i+1})$. In practice, the first derivatives $y'(T_i) = dy(T_i)/dT$ are most often specified as finite difference coefficients $y'(T_i) = (y(T_{i+1}) - y(T_{i-1})) / (T_{i+1} - T_{i-1})$, giving rise to the so-called *Catmull–Rom spline* [3].

Solving for the Catmull–Rom spline that satisfies equation (4) involves an iterative search for the unknown levels $y(T_1), \dots, y(T_N)$, with each iteration involving a construction of the spline and a computation of $\mathbf{P} = (e^{-y(t_1)t_1}, \dots, e^{-y(t_M)t_M})$. Any standard multidimensional root-search algorithm can be applied here. We notice that the Catmull–Rom spline links values of $y(T)$, $T \in (T_i, T_{i+1})$ to only four knots, namely, $y(T_{i-1})$, $y(T_i)$, $y(T_{i+1})$, $y(T_{i+2})$,

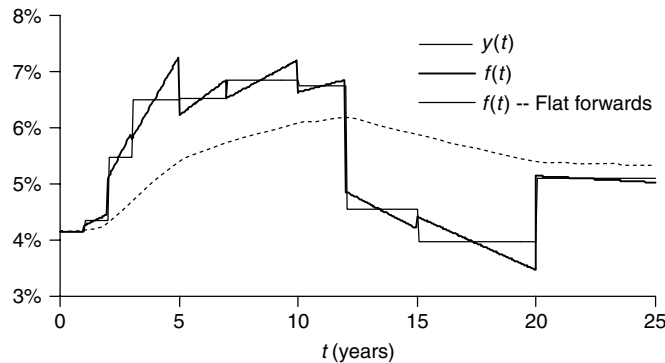


Figure 1 Yield and forward curve (linear yield bootstrap)

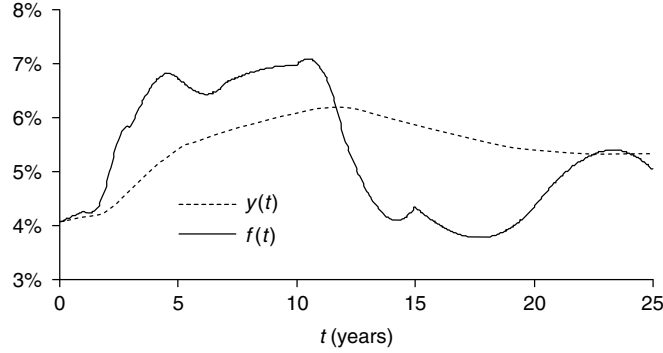


Figure 2 Yield and forward curve (Catmull–Rom spline)

which simplifies the causality structure in the model and allows for application of “near-bootstrap” methods. Figure 2 shows typical yield and forward curves generated by the Catmull–Rom spline approach, using the same benchmark set as was used to construct Figure 1.

We can easily extend the procedure above beyond Catmull–Rom splines to more complicated C^1 cubic splines in the Hermite class; for instance, it is relatively straightforward to add *tension* to the Catmull–Rom spline. See [7] for details.

C^2 Cubic Splines

While the spline method introduced in the previous section often produces acceptable yield curves, the method is heuristic in nature and ultimately does not produce a smooth forward curve. To improve on the latter, one alternative is to remain in the realm of cubic splines, but now insist that the curve is twice differentiable everywhere on $[T_1, T_N]$. The resulting spline equations are

$$\begin{aligned}
 y(T) = & \frac{(T_{i+1} - T)^3}{6h_i} y_i'' + \frac{(T - T_i)^3}{6h_i} y_{i+1}'' \\
 & + (T_{i+1} - T) \left(\frac{y_i}{h_i} - \frac{h_i}{6} y_i'' \right) + (T - T_i) \\
 & \times \left(\frac{y_{i+1}}{h_i} - \frac{h_i}{6} y_{i+1}'' \right), \quad T \in [T_i, T_{i+1}]
 \end{aligned} \tag{9}$$

where $y_i'' = d^2 y(T_i)/dT^2$, $y_i = y(T_i)$, and $h_i = T_{i+1} - T_i$. Continuity of the second derivative across

the $\{T_i\}$ knots requires that y_i'' and y_i , $i = 1, \dots, N$, are connected through a tri-diagonal linear system of equations; see [2, 10] for the well-known details. Full specification of this system of equations requires exogenous characterization of behavior at the boundaries T_1 and T_N . The most common—and often best—specification is that of the *natural spline*, where we set $y_1'' = y_N'' = 0$.

As for the Catmull–Rom spline, solving the C^2 cubic spline yield curve that satisfies equation (4) involves a numerical search for the unknown levels^c $y_1, \dots, y_N$. The fitting problem is typically good-natured, and virtually all standard root-search packages can tackle it successfully. References [1, 14] use a simple Gauss–Newton scheme, whereas [6] applies a fixed-point-type iteration. Both these simple suggestions are most likely outperformed by the backtracking Newton method or the Broyden method, both of which are described in Press *et al.* [10]. An example of the forward curve arising from this approach can be seen in Figure 3 (the case $\sigma = 0$).

While the C^2 cubic spline discussed here has attractive smoothness, it is not necessarily an ideal representation of the yield curve. As discussed in [1, 6], among others, twice differentiable cubic spline yield curves are often subject to oscillatory behavior, spurious inflection points, poor extrapolation behavior, and nonlocal behavior when prices in the benchmark set are perturbed. In particular, perturbation of a single benchmark price can cause a slow-decaying “ringing” effect on the C^2 cubic yield curve, with the effect of the perturbation of the benchmark instrument price spilling into the entire yield curve. This behavior is a nuisance in risk-management applications, and is much less present in curves constructed by

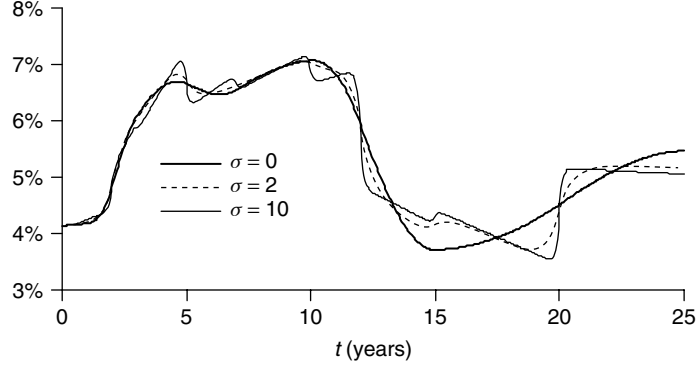


Figure 3 Forward curve (tension spline)

bootstrapping or by the Hermite spline approach.^d A pragmatic, but inherently inconsistent, approach is to use a C^2 cubic spline for pricing purposes only, but use bootstrapping when stability to perturbations is required. A more sophisticated approach is discussed below.

C^2 Tension Splines

Hermite cubic splines are less prone to nonlocal perturbation behavior than C^2 cubic splines, but accomplish this in a somewhat *ad hoc* fashion by giving up one degree of differentiability. Rather than taking such an extreme step, one wonders whether there may be a way to retain the C^2 feature of the cubic spline, yet still allow control of the curve locality and “stiffness”. As it turns out, an attractive remedy to the shortcomings of the pure C^2 cubic spline is to insert some *tension* in the spline, that is, to apply a tensile force to the end points of the spline. Details about this idea can be found in [12]; when applied to the yield Curve setting, the construction involves a modification of the cubic equation (9) for $y(T)$ to

$$\begin{aligned}
 y(T) = & \left(\frac{\sinh(\sigma(T_{i+1} - T))}{\sinh(\sigma h_i)} - \frac{T_{i+1} - T}{h_i} \right) \frac{y_i''}{\sigma^2} \\
 & + \left(\frac{\sinh(\sigma(T - T_i))}{\sinh(\sigma h_i)} - \frac{T - T_i}{h_i} \right) \frac{y_{i+1}''}{\sigma^2} \\
 & + y_i \frac{T_{i+1} - T}{h_i} + y_{i+1} \frac{T - T_i}{h_i}, \\
 & T \in [T_i, T_{i+1}]
 \end{aligned} \tag{10}$$

where $\sigma \geq 0$ is the *tension factor*, and where we recall the definition $h_i = T_{i+1} - T_i$.

Among the properties of the tension splines are the facts that setting $\sigma = 0$ will recover the ordinary C^2 cubic spline, whereas letting $\sigma \rightarrow \infty$ will make the tension spline uniformly approach a linear spline (i.e., the spline we used in the section Bootstrapping). Loosely, we can thus think of a tension spline as a twice differentiable hybrid between a cubic spline and a linear spline. Equally loosely, as we increase σ , inflections and ringing in the cubic spline are gradually “stretched” out of the curve, accompanied by rising (absolute values of) second derivatives at the knots.

More details on tension splines in yield Curve application can be found in [1], which also contains a discussion of computationally efficient local spline bases and the usage of T -dependent tension factors for additional curve control. For our purposes here, it suffices to note that equation (10) is structurally similar to equation (9), and also allows for a tri-diagonal matrix equation linking y_i'' and y_i , $i = 1, \dots, N$. The solution procedure for the yield curve is therefore the same as in the section C^2 Cubic Splines. Figure 3 illustrates the effect of varying the tension factor on the shape of the instantaneous forward curve $f(t)$; notice how increasing the tension parameter gradually moves us from smooth cubic spline behavior to bootstrap behavior. Examples of how the tension parameter dampens ringing in the forward curve after input perturbations can be found in [1].

Nonparametric Methods

The techniques we outlined so far generally suffice for the construction of a discount curve from a “clean” set of non-duplicate benchmark securities, including the carefully selected set of liquid staggered-maturity deposits, futures, and swaps, that most banks assemble for the purpose of constructing a Libor yield curve. In some settings, however, the benchmark set may be significantly less well structured, involving illiquid securities with little order in their cash-flow timing and considerable noise in their prices. This situation may, say, arise when one attempts to construct a yield curve from corporate bonds.

When the input benchmark set is noisy, a straight solution of (4) may be erratic or may not exist. To overcome this, and to reflect that noise in the input data may make us content to solve (4) only to within certain error bounds, we now proceed to replace this equation by minimization of a penalized least-squares norm. Specifically, define the space $\mathcal{A} = C^2[t_1, t_M]$ of all twice differentiable functions $[t_1, t_M] \rightarrow \mathbb{R}$ and introduce the M -dimensional discount vector

$$\mathbf{P}(y) = (e^{-y(t_1)t_1}, \dots, e^{-y(t_M)t_M}) \quad (11)$$

Also, let $\mathbf{W}$ be a diagonal $N \times N$ weighting matrix. Then, as our best estimate $\hat{y}$ of the yield curve we may use

$$\hat{y} = \arg \min_{y \in \mathcal{A}} \mathcal{I}(y) \quad (12)$$

$$\begin{aligned} \mathcal{I}(y) \equiv & \frac{1}{N} (\mathbf{V} - \mathbf{cP}(y))^\top \mathbf{W}^2 (\mathbf{V} - \mathbf{cP}(y)) \\ & + \lambda \left(\int_{t_1}^{t_M} [y''(t)^2 + \sigma^2 y'(t)^2] dt \right) \end{aligned} \quad (13)$$

where λ and σ are positive constants. The norm $\mathcal{I}(y)$ consists of three separate terms:

- A least-squares penalty term

$$\begin{aligned} & \frac{1}{N} (\mathbf{V} - \mathbf{cP}(y))^\top \mathbf{W}^2 (\mathbf{V} - \mathbf{cP}(y)) \\ & = \frac{1}{N} \sum_{i=1}^N W_i^2 \left(V_i - \sum_{j=1}^M c_{ij} e^{-y(t_j)t_j} \right)^2 \end{aligned} \quad (14)$$

where W_i is the i th diagonal element of $\mathbf{W}$. This term is an outright precision-of-fit norm and measures the degree to which the constructed discount curve can replicate input security prices. The weight-matrix $\mathbf{W}$ can be used to express the relative importance of the various securities in the benchmark set, or to turn price errors into yield errors.

- A weighted smoothness term $\lambda \int_{t_1}^{t_M} y''(t)^2 dt$, penalizing high second-order gradients of y to avoid kinks and discontinuities.
- A weighted curve-length term $\lambda \sigma^2 \int_{t_1}^{t_M} y'(t)^2 dt$, penalizing oscillations and excess convexity/concavity.

To construct the yield curve, we have replaced the nonlinear root-search problems encountered in the section Yield Curve Fitting with N -knot Splines with the functional optimization problem (12). Fortunately, the latter approach can be linked to that of the former by the following result, which can be shown by variational methods.

Proposition 1 *The curve $\hat{y}$ that satisfies equation (12) is a natural exponential tension spline with tension factor σ and knots at all cash-flow dates $t_1, t_2, \dots, t_M$.*

Proposition 1 establishes that the curve we are looking for is a tension spline with tension factor σ , but does not, in itself, allow us to identify the optimal spline directly, beyond the fact that (i) it is a natural spline with boundary conditions $y''(t_1) = y''(t_M) = 0$; (ii) it has knots at all t_i . Identification of the correct tension spline involves solving for unknown levels $y(t_1), y(t_2), \dots, y(t_M)$ to directly optimize equation (13). This optimization problem can be solved by standard methods, for example, by use of the Levenberg–Marquardt method, or the Gauss–Newton method in [1, 14].

Remark 1 If we let $\sigma = 0$, the solution to the optimization problem becomes a *cubic smoothing spline*; see [14] for more details on this case.

Choice of Smoothing Parameter

The parameter λ may be specified exogenously by the user, as a way to provide a trade-off between pricing accuracy and curve regularity. In practice, however, a good magnitude of λ may sometimes be

hard to ascertain by inspection, and a procedure to estimate λ directly from the data is often useful. One possibility is to use a cross-validation approach, either outright or through the more efficient generalized cross-validation (GCV) criterion in [4]. Some results along these lines can be found in [14], for instance. A more pragmatic approach is to specify a target value for the least-squares term in equation (13), iterating on λ until the target is met; in general, we would expect the least-squares error term to increase monotonically with λ . Most trading desks should have little difficulty specifying a meaningful least-squares target error directly from, say, observed bid–offer spreads. We note that the target error value may be set to zero, if a perfect fit to benchmark securities is required.

Special Topics

While our discussion of curve construction algorithms generally relied on the notion that the forward curve should ideally be *smooth*, there may be circumstances where we want to make exceptions. For instance, it may be reasonable to expect instantaneous forwards to jump on or around meetings of monetary authorities, such as the Federal Reserve in the United States. In addition, other “special” situations may exist that might warrant introduction of discontinuities into the forward curve. A well-known example is the turn-of-year (TOY) effect where short-dated loan premiums spike for loans between the last business day of the year and the first business day of the next year. One common way of incorporating TOY-type effects is to exogenously specify an *overlay curve* $\epsilon_f(t)$ on the instantaneous forward curve. Specifically, the forward curve $f(t)$ is written as

$$f(t) = \epsilon_f(t) + f^*(t) \quad (15)$$

where $\epsilon_f(t)$ is user-specified—and most likely contains discontinuities around special events dates—and $f^*(t)$ is unknown. The yield curve construction problem is then subsequently applied to the construction of $f^*(t)$, using algorithms such as those discussed earlier.

We should note that the curve construction algorithms outlined in this article are meant only for single-currency applications. In a setting with multiple currencies, care must be taken to ensure that discount curves in different currencies are properly

related, in the sense that they together replicate the prices for foreign exchange forward agreements, as well as cross-currency floating–floating basis swaps. In general, the ability to match currency markets requires that the discount curve and the forward rate curve (e.g., the Libor curve) be separated into two entities separated by a cross-currency spread; see [5]. It is normally straightforward to embed a single-currency yield curve solver into a cross-currency setting, for instance, by means of an iterative adjustment to the price vector $\mathbf{V}$ in equation (4). Similar techniques can be used to accommodate the so-called *tenor* basis, that is, the fact that different Libor tenors (e.g., 3-month versus 6-month) in practice do not swap flat against each other.

Acknowledgments

The authors are grateful for the suggestions of Brian Ostrow, Igor Polonsky, David Price, and Branko Radosavljevic.

End Notes

^aOwing to their daily mark-to-market provision, Eurodollar futures contracts do not allow for a pricing expression of the form (1), so a preprocessing step is normally employed to convert the futures rate quote to a forward rate quote. See **Eurodollar Futures and Options**.

^bSee, for example, [13] for a discussion of the pitfalls associated with curve interpolators that work directly on the discount function $P(T)$ (as in [8]).

^cA more contemporary approach replaces this search with a search for coefficients in a local spline basis. Andersen [1] contains more details on this.

^dIntuitively, this is because linear and Hermite splines link values of $y(T)$ to only a few (2 and 4, respectively) of the values $y(T_i)$, $i = 1, \dots, N$. The C^2 cubic spline, on the other hand, links $y(T)$ to *all* $y(T_i)$, $i = 1, \dots, N$.

References

- [1] Andersen, L. (2006). Discount curve construction with tension splines, *Review of Derivatives Research* **10**(3), 227–267.
- [2] de Boor, C. (2001). *A Practical Guide to Splines (revised edition)*, Springer Verlag, New York.
- [3] Catmull, E. & Rom, R. (1974). A class of local interpolating spline, in *Computer Aided Geometric Design*, R.E. Barnhill & R.F. Riesenfeld, eds, Academic Press, New York.

- [4] Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematik* **31**, 377–403.
- [5] Fruchard, E., Zammouri, C. & Willems, E. (1995). Basis for change, *RISK Magazine* October, 70–75.
- [6] Hagan, P. & West, G. (2006). Interpolation methods for yield curve construction, *Applied Mathematical Finance* **3**(2), 89–129.
- [7] Kochanek, D. & Bartels, R. (1984). Interpolating splines with local tension, continuity, and bias control, *ACM SIGGRAPH* **18**(3), 33–41.
- [8] McCulloch, J.H. (1975). The tax-adjusted yield curve, *Journal of Finance* **30**, 811–830.
- [9] Nelson, C.R. & Siegel, A.F. (1987). Parsimonious modeling of yield curves, *Journal of Business* **60**, 473–489.
- [10] Press, W., Teukolsky, S., Vetterling, W. & Flannery, B. (1992). *Numerical Recipes in C*, Cambridge University Press.
- [11] Schoenberg, I. (1973). *Cardinal Spline Interpolation*. SIAM CBMS-NSF Regional Conference Series in Applied Mathematics 12.
- [12] Schweikert, D.G. (1966). An interpolating curve using a spline in tension, *Journal of Mathematics and Physics* **45**, 312–317.
- [13] Shea, G.S. (1984). Pitfalls in smoothing interest rate terms structure data: equilibrium models and spline approximation, *Journal of Financial and Quantitative Analysis* **19**, 253–269.
- [14] Tanggaard, C. (1997). Nonparametric smoothing of yield curves, *Review of Quantitative Finance and Accounting* **9**, 251–267.

Related Articles

Eurodollar Futures and Options; Hedging of Interest Rate Derivatives; LIBOR Rate.

LEIF B.G. ANDERSEN & VLADIMIR V.
PITERBARG

Stochastic Volatility Interest Rate Models

Stochastic volatility has been widely used to model implied volatility smiles for European caps and swaptions. We discuss models of the yield curve that incorporate stochastic volatility, defined as randomness in the volatility of the bond prices that is not spanned by movements in the yield curve. We argue that it is difficult to specify short rate models that exhibit unspanned stochastic volatility and that a more natural choice for construction of stochastic volatility models is the Heath, Jarrow, and Morton (HJM) or Libor Market Model framework. We then consider the specification of stochastic volatility interest rate models and survey some of the stochastic volatility models found in the literature.

“Unspanned” Stochastic Volatility

Stochastic volatility interest-rate models are models that prescribe moves in the volatility of rates that cannot directly be inferred from the shape or the level of the yield curve. Let $P(t, T)$ be the time t price of a zero-coupon bond maturing at time T and let the time t continuously compounded forward rate for deposit over the interval $[T, T + dT]$ be given by

$$f(t, T) = -\frac{\partial \ln P(t, T)}{\partial T}, t \leq T \quad (1)$$

Assume that interest rates evolve continuously. As shown by Heath *et al.* [9], absence of arbitrage implies that the forward rates have to evolve according to (see **Heath–Jarrow–Morton Approach** for the HJM model)

$$df(t, T) = \sigma(t, T) \cdot \left(\int_t^T \sigma(t, s) ds \right) dt + \sigma(t, T) \cdot dW(t) \quad (2)$$

where W is a vector Brownian motion under the risk-neutral measure, and $\{\sigma(t, T)\}_{t \leq T}$ some family of vector processes.

A “true” stochastic volatility model has the property that there exists some stochastic process z and at least one maturity U , so that

$$\frac{\partial \sigma(t, U)}{\partial z(t)} \neq 0 \quad (3a)$$

and

$$\frac{\partial f(t, T)}{\partial z(t)} = 0 \quad (3b)$$

for all T .

In other words, “true” or “unspanned” stochastic volatility is when there is uncertainty in the volatility of the rates, which cannot be fully hedged by taking positions in bonds. There is considerable empirical evidence of unspanned stochastic volatility in interest rates and interest rate options markets (see, e.g., Casassus *et al.* [5]).

It is very difficult to specify a traditional short-rate model that can be categorized as a true stochastic volatility model. This is so because stochastic short-rate volatility will tend to show up as a second factor that the bond prices will depend on. Consider, for example, the model by Fong and Vasicek [7]:

$$\begin{aligned} dr(t) &= \kappa(\theta - r(t)) dt + \sqrt{v(t)} dW_1(t) \\ dv(t) &= \beta(\alpha - v(t)) dt + \varepsilon \sqrt{v(t)} dW_2(t) \\ dW_1(t) \cdot dW_2(t) &= \rho dt \end{aligned} \quad (4)$$

where $\kappa, \theta, \beta, \alpha, \varepsilon, \rho$ are constants, and W_1 and W_2 Brownian motions under the risk-neutral measure. In this model, we have

$$\begin{aligned} P(t, T) &= E_t \left[e^{-\int_t^T r(u) du} \right] \\ &= E \left[e^{-\int_t^T r(u) du} | r(t), v(t) \right] \\ &\equiv P(t, T; r(t), v(t)) \end{aligned} \quad (5)$$

So the bond price becomes a function of two stochastic variables. Hence, we can invert the system and infer the level of both the short rate and the short rate volatility from any two points on the yield curve. Thus, the model is not a true stochastic volatility model. This is also the case for the Longstaff and Schwartz [13] model and other early attempts to produce stochastic volatility yield curve model.

In fact, it is also the case for attempts to formulate a stochastic volatility yield curve model in the context of the Markov functional approach by Hunt *et al.* [12] (see **Markov Functional Models** for the Markov functional models).

So, as observed by Andreasen *et al.* [4], the most straightforward way of formulating a stochastic volatility yield curve model is to directly use the

HJM approach, or equivalently the Libor market model approach, and directly specify the stochastic nature of the bond or forward rate volatility structure (see **LIBOR Market Model** for the Libor market model). In the HJM modeling approach, we see that it is easy to specify a volatility structure satisfying equation (3a,b). We could, for example, set $\sigma(t, T) = c \cdot \sqrt{z(t)}$ for some constant c and some Markov process z .

Intuitively, if the volatility is nondeterministic, then the minimal number of state variables in a HJM model is two, so with the addition of stochastic volatility the number of state variables in a “true” stochastic volatility HJM model is at least three. In fact, Dufresne and Goldstein [6] provide a partial differential equation (PDE)-based argument to justify that the minimal number of state variables for a “true” stochastic volatility interest rate model is three.

Model Specifications

In the following, we present some examples of stochastic volatility interest-rate models. A Libor market model is based on a discrete time grid $0 = t_0 < t_1 < \dots$. Let

$$L_k(t) = \frac{1}{t_{k+1} - t_k} \left(\frac{P(t, t_k)}{P(t, t_{k+1})} - 1 \right) \quad (6)$$

be the forward Libor rate over the period $[t_k, t_{k+1}]$. Under absence of arbitrage, we have

$$\begin{aligned} dL_k(t) = & \vartheta_k(t) \cdot \sum_{j=i+1}^k \frac{\delta_j \vartheta_j(t)}{1 + \delta_j L_j(t)} dt \\ & + \vartheta_k(t) \cdot d\tilde{W}(t), t_{i-1} \leq t < t_i \end{aligned} \quad (7)$$

for a discrete set of vector processes $\{\vartheta_k(t)\}_{t \leq t_k}$ and $\tilde{W}$ being a vector Brownian motion under the martingale measure with $B(t) = \left(\prod_{j=0}^{n-1} P(t_j, t_{j+1}) \right) \times P(t, t_{n+1})$, $t_n \leq t < t_{n+1}$ as numeraire.

Andersen and Brotherton-Ratcliffe [3] consider an uncorrelated stochastic volatility extended constant elasticity of variance (CEV) Libor market model

$$\begin{aligned} \|\vartheta_k(t)\| &= \sqrt{z(t)} \lambda_k(t) L_k(t)^\beta \\ dz(t) &= \theta(1 - z(t)) dt + \varepsilon \sqrt{z(t)} dZ(t) \\ dL_k(t) \cdot dz(t) &= 0 \end{aligned} \quad (8)$$

where Z is a Brownian motion, λ_k is a time-dependent function, and θ, ε are constants. Andersen and Brotherton-Ratcliffe suggest asymptotic expansions for solving for European swaption prices based on approximation of the swap rate dynamics.

Piterbarg [16] replaces the CEV assumption for the forward rate volatility with a linear one:

$$\begin{aligned} \|\vartheta_k(t)\| &= \sqrt{z(t)} \lambda_k(t) [\beta_k(t) L_k(t) \\ &\quad + (1 - \beta_k(t)) L_k(0)] \\ dz(t) &= \theta(1 - z(t)) dt + \varepsilon \sqrt{z(t)} dZ(t) \\ dL_k(t) \cdot dz(t) &= 0 \end{aligned} \quad (9)$$

Piterbarg shows that using time- and tenor-dependent skew coefficients ($\beta_k(t)$) improves the simultaneous fit to implied cap and swaption skews and smiles. Piterbarg solves for European swaption prices using Markovian projection techniques applied to the process for the swap rate.

Andersen and Andreasen [2] present a one-factor Markov HJM model with uncorrelated stochastic volatility:

$$\begin{aligned} P(t, T) &= \frac{P(0, T)}{P(0, t)} e^{-G(t, T)x(t) - \frac{1}{2}G(t, T)^2 y(t)} \\ G(t, T) &= \frac{1 - e^{-\kappa(T-t)}}{\kappa} \\ dx(t) &= (-\kappa x(t) + y(t)) dt + \eta(t) dW(t) \\ dy(t) &= (\eta(t)^2 - 2\kappa y(t)) dt \\ \eta(t) &= \lambda(t) \sqrt{z(t)} [\beta R(t, s, T) \\ &\quad + (1 - \beta) R(0, s, T)] \\ dz(t) &= \theta(1 - z(t)) dt + \varepsilon \sqrt{z(t)} dZ(t) \\ dW(t) \cdot dZ(t) &= 0 \end{aligned} \quad (10)$$

where W, Z are Brownian motions under the risk-neutral measure and

$$R(t, s, T) = -\frac{1}{T - s} \ln \frac{P(t, T)}{P(t, s)} \quad (11)$$

is a continuously compounded zero-coupon forward rate that is linked to the choice of calibration instruments. Piterbarg’s Markovian projection techniques are used for calibration of the model. Owing to the limited number of state variables (equation 3), the model allows for finite difference solution and efficient simulations.

Andreasen [1] presents a multifactor Markov HJM model that extends equation (10) and allows for time- and tenor-dependent skew as in equation (9). For a selected set of (continuously compounded) forward rates with tenors, the dynamics are similar to the forward rate dynamics in equation (9) (*see also Markovian Term Structure Models*).

The models discussed so far are all based on zero correlation between the interest rates and the stochastic volatility process. The reason for this choice is mainly technical: if it were not the case, then the stochastic volatility process would be different for different annuity measures (*see Forward and Swap Measures*) and it would often include complicated terms in its drift. This would, in turn, make approximation of the swaption prices and thereby calibration more complicated.

The reason for the choice of square-root process for the stochastic volatility is also tractability. The square-root process admits computation of exponential moments in closed form and this can either be used for approximation of at-the-money option prices as in Piterbarg [16] or for direct computation of option prices *via* numerical inversion of Fourier transforms. An example of the latter is the model in equation (10) with level independent but correlated volatility:

$$\begin{aligned}\eta(t) &= \lambda(t)\sqrt{z(t)} \\ dW(t) \cdot dZ(t) &= \rho(t) dt\end{aligned}\quad (12)$$

where ρ is a time-dependent function. For this model, the processes for the maturity U forward bond price $P(t, U)/P(t, T)$ and the stochastic volatility factor, z , are

$$\begin{aligned}\frac{d(P(t, U)/P(t, T))}{(P(t, U)/P(t, T))} &= \lambda(t)\sqrt{z(t)}(G(t, T) \\ &\quad - G(t, U)) dW^T(t) \\ dz(t) &= \theta(1 - z(t)) dt + \varepsilon\sqrt{z(t)} dZ^T(t) \\ &\quad - \rho\varepsilon z(t)\lambda(t)G(t, T) dt\end{aligned}\quad (13)$$

where W^T, Z^T are correlated Brownian motions under the maturity T forward measure. So, essentially, the processes in equation (13) are similar to the single-asset stochastic volatility model of Heston [11] but with time-dependent parameters. So, as shown by Dufresne and Goldstein [6], this means that the prices of caplets and options on zero-coupon bonds can be

found by numerical inversion of Fourier transforms in the same way as for the Heston model. Swaption prices can, in turn, be found by approximating swaptions as options on zero-coupon bonds by duration matching as suggested by Munk [14].

As an alternative to the square-root process for the stochastic volatility, Rebonato [15] and Henry-Labordere [10] consider the SABR-based stochastic volatility Libor market models with correlated stochastic volatility of the form (*see [8] for the SABR model*).

$$\begin{aligned}dL_k(t) &= z\lambda_k(t)L_k(t)^{\beta_k} dW_k(t) + O(dt), \\ k &= 1, \dots, n \\ dz(t) &= v z(t) dW_{n+1}(t) \\ dW_i(t) \cdot dW_j(t) &= \rho_{ij}(t) dt, \\ i, j &= 1, \dots, n+1\end{aligned}\quad (14)$$

where v is a constant and $\{\rho_{ij}(t)\}$ is time dependent. Henry-Labordere provides asymptotic expansion results for the prices of caplets and swaptions based on hyperbolic geometry methods.

References

- [1] Andreasen, J. (2005). Back to the future, *Risk* **18**(9), 72–78.
- [2] Andersen, L. & Andreasen, J. (2002). Volatile volatilities, *Risk* **15**(12), 65–71.
- [3] Andersen, L. & Brotherton-Ratcliffe, R. (2005). Extended libor market models with stochastic volatility, *Journal of Computational Finance* **9**, 1–40.
- [4] Andreasen, J., Dufresne, P. & Shi, W. (1994). *An arbitrage term Structure model of interest rates with stochastic volatility*.
- [5] Casassus, J., Dufresne, P. & Goldstein, R. (2005). Unspanned stochastic volatility and fixed income derivatives pricing, *Journal of Banking and Finance* **29**, 2723–2749.
- [6] Dufresne, P. & Goldstein, R. (2002). Do bonds span the fixed income markets? Theory and evidence for unspanned stochastic volatility, *Journal of Finance* **57**, 1685–1730.
- [7] Fong, H. & Vasicek, O. (1991). Fixed income volatility management, *The Journal of Portfolio Management*, 41–46.
- [8] Hagan, P., Kumar, D., Lesniewski, A. & Woodward, D. (2002). Managing smile risk, *Wilmott Magazine* **2**(7), 84–108.
- [9] Heath, D., Jarrow, R. & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation, *Econometrica* **60**, 77–106.

- [10] Henry-Labordere, P. (2007). Combining the SABR and LMM models, *Risk* **20**(10), 102–107.
- [11] Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**, 327–344.
- [12] Hunt, P., Kennedy, J. & Pelsser, A. (1998). Fit and run, *Risk* **10**(11), 65–67.
- [13] Longstaff, F. & Schwartz, E. (1992). Interest rate volatility and the term structure: a two-factor general equilibrium model, *Journal of Finance* **47**, 1259–1282.
- [14] Munk, C. (1999). Stochastic duration and fast coupon bond option pricing in multi-factor models, *Review of Derivatives Research* **3**, 157–181.
- [15] Rebonato, R. (2001). The stochastic volatility LIBOR market model, *Risk* **13**(10), 105–110.
- [16] Piterbarg, V. (2003). *A Stochastic Volatility Forward Libor Model with a Term Structure of Volatility Smiles*. Bank of America working paper. Available from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=472061.

Related Articles

Heath–Jarrow–Morton Approach; Heston Model; LIBOR Market Model; Markovian Term Structure Models.

JESPER ANDREASEN

Heath–Jarrow–Morton Approach

Basics

We consider a financial market model living on a filtered probability space $(\Omega, \mathcal{F}, F, P)$, where $F = \{\mathcal{F}_t\}_{t \geq 0}$ and P is the objective probability measure. The basis is assumed to carry a standard m -dimensional P -Wiener process $\bar{W}$.

Our main object of study is the zero coupon bond market, and we denote the price by $p(t, T)$, at t , of a zero coupon bond maturing at T .

We define, as usual, the *instantaneous forward rate with maturity T , contracted at t* , by

$$f(t, T) = -\frac{\partial \log p(t, T)}{\partial T} \quad (1)$$

The instantaneous *short rate at time t* is defined by $r(t) = f(t, t)$.

The HJM Framework

We now turn to the specification of the Heath–Jarrow–Morton (HJM) framework (see [13]). We start by specifying everything under a given objective measure P .

Assumption 1 We assume that, for every fixed $T > 0$, the forward rate $f(\cdot, T)$ has a stochastic differential, which, under the objective measure P , is given by

$$df(t, T) = \alpha(t, T) dt + \sigma(t, T) d\bar{W}_t \quad (2)$$

$$f(0, T) = f_{\text{in}}(0, T) \quad (3)$$

where, for each fixed T , $\alpha(\cdot, T)$ and $\sigma(\cdot, T)$ are adapted processes. The curve f_{in} is the initially observed forward rate curve.

It is important to note that the HJM approach to model the evolution of interest rates is not a proposal of a specific *model*, like, for example, the Cox–Ingersoll–Ross model (see **Cox–Ingersoll–Ross (CIR) Model**). It is, instead, a *framework* that is used for analyzing interest-rate models. In fact, every interest-rate model can be equivalently formulated in forward rate terms. To turn the HJM

framework into a *model*, we have to specify the volatility and drift structure, that is, we have to specify σ and α .

There are two main advantages of the HJM approach: First, the forward rate volatility structure $\sigma(t, T)$ is an *input* to the model, whereas in a factor model such as a short rate model (see **Term Structure Models**), it would be an output. Second, by using the observed forward rate curve as an initial condition, we automatically obtain a perfect fit to the observed yield curve.

The first main result shows how the bond price dynamics are determined by the forward rate dynamics.

Proposition 1 *If the forward rate dynamics are given by equation (2) then the induced bond price dynamics are given by*

$$\begin{aligned} dp(t, T) = & p(t, T) \{ r(t) + A(t, T) \\ & + \frac{1}{2} \|S(t, T)\|^2 \} dt \\ & + p(t, T) S(t, T) d\bar{W}_t \end{aligned} \quad (4)$$

where $\|\cdot\|$ denotes the Euclidean norm, and

$$\begin{cases} A(t, T) = -\int_t^T \alpha(t, s) ds \\ S(t, T) = -\int_t^T \sigma(t, s) ds \end{cases} \quad (5)$$

Absence of Arbitrage

Using proposition 1 above, an application of the Girsanov Theorem gives us the following basic result concerning absence of arbitrage.

Theorem 1 (HJM Drift Condition). *Assume that the family of forward rates is given by equation (2). Then the induced bond market is arbitrage free if and only if there exists a d -dimensional column-vector process $\lambda(t) = [\lambda_1(t), \dots, \lambda_d(t)]^*$ with the property that for all $T \geq 0$ and for all $t \leq T$, we have*

$$\alpha(t, T) = \sigma(t, T) \int_t^T \sigma(t, s)^* ds - \sigma(t, T) \lambda(t) \quad (6)$$

In these formulas $*$ denotes transpose.

Martingale Modeling

In many cases, the specification of the forward rate dynamics is done directly under a martingale

measure Q as

$$\begin{aligned} df(t, T) &= \alpha(t, T) dt + \sigma(t, T) dW_t \\ f(0, T) &= f_{\text{in}}(0, T) \end{aligned} \quad (7)$$

where W is a (d -dimensional) Q -Wiener process. In this setting, absence of arbitrage is no longer an issue, but we have to give conditions that guarantee that all the induced bond price processes have the correct martingale dynamics, that is, short rate as their local rate of return. This directly follows from the earlier result by setting $\lambda = 0$.

Proposition 2 (HJM Drift Condition). *Under the martingale measure Q , the processes α and σ must satisfy the following relation, for every t and every $T \geq t$.*

$$\alpha(t, T) = \sigma(t, T) \int_t^T \sigma(t, s)^* ds \quad (8)$$

Thus when specifying the forward rate dynamics, (under Q) we may freely specify the volatility. The drift is then uniquely determined. In practical applications, one thus has to specify the number d of Wiener processes as well as the volatility structure.

It is common to assume a deterministic volatility structure, and then try to estimate this as well as the number d by principal component analysis. The deterministic volatility is very tractable since it will lead to Gaussian forward rates (see **Gaussian Interest-Rate Models**) and lognormal bond prices. Analytical formulas for bond options are easily available.

The HJM approach has been extended to include a driving marked point process in [4], a random measure [3], a Levy process [8], a Gaussian random field [15], and a Levy field [1].

The Musiela Parameterization

In many applications, it is more natural to use time to maturity, rather than time of maturity, to parameterize bonds and forward rates, and this approach was first described in [6] and [16]. If we denote time to maturity by x , then we have $x = T - t$, and in terms of x , the forward rates are defined as follows.

Definition 1 *For all $x \geq 0$ the forward rates $r_t(x)$ are defined by the relation*

$$r_t(x) = f(t, t + x) \quad (9)$$

and we denote by r_t the forward rate curve $x \mapsto r_t(x)$ at time t . We can thus view r as a process taking values in some Hilbert space $\mathcal{H}$ of forward rate curves.

Suppose now that we have the standard HJM-type model for the forward rates under a martingale measure Q

$$df(t, T) = \alpha(t, T) dt + \sigma_0(t, T) dW_t \quad (10)$$

where σ_0 denotes the HJM volatility structure. The question is to find the Q -dynamics for $r(t, x)$, and we have the following result.

Proposition 3 (Musiela parameterization). *Assume that the forward rate dynamics under Q are given by (10). Then*

$$dr_t(x) = \left\{ \frac{\partial}{\partial x} r_t(x) + D(t, x) \right\} dt + \sigma(t, x) dW_t \quad (11)$$

where

$$\begin{aligned} \sigma(t, x) &= \sigma_0(t, t + x) \\ D(t, x) &= \sigma(t, x) \int_0^x \sigma(t, s)^* ds \end{aligned} \quad (12)$$

If the volatility σ is of the simple deterministic form $\sigma_t(x) = \sigma(x)$, then the Musiela equation mentioned above takes the form

$$dr_t = \{\mathbf{F}r_t + D\} dt + \sigma dW_t$$

where $D(x) = \sigma(x) \int_0^x \sigma(s) ds$ and the operator $\mathbf{F}$ is given by $\partial/\partial x$. In this case, the forward rate equation is an infinite dimensional linear stochastic differential equation (SDE) on $\mathcal{H}$ with formal solution

$$r_t = e^{\mathbf{F}t} r_0 + \int_0^t e^{\mathbf{F}(t-s)} D ds + \int_0^t e^{\mathbf{F}(t-s)} \sigma dW_s \quad (13)$$

where the semigroup $e^{\mathbf{F}t}$ is left translation, that is, $e^{\mathbf{F}t} f(x) = f(t + x)$.

Geometric Interest Rate Theory

Assume that the volatility process σ is of the Markovian form $\sigma_t(x) = \sigma(r_t, x)$. Then the Musiela equation for the r process is an infinite dimensional SDE and we write it compactly as

$$dr_t = \mu(r_t) dt + \sigma(r_t) dW(t) \quad (14)$$

where

$$\begin{aligned} \mu(r, x) &= \frac{\partial}{\partial x} r(x) + D(r, x) \\ D(r, x) &= \sigma(r, x) \int_0^x \sigma(r, s) ds \end{aligned} \quad (15)$$

We can thus view μ and σ as vector fields on $\mathcal{H}$, and we now formulate a couple of natural problems:

1. Consider a given parameterized family $\mathcal{G}$ of forward rate curves, such as the Nelson–Siegel family, where forward rates are parameterized as

$$G(z, x) = z_1 + z_2 e^{-z_3 x} + z_4 x e^{-z_3 x} \quad (16)$$

where $z_1, \dots, z_4$ are the parameters. The question is now under which conditions this family is *consistent* with the dynamics of the interest-rate model mentioned above? Here consistency is interpreted in the sense that, given an initial forward rate curve in $\mathcal{G}$, the interest-rate model will (with probability 1) produce forward rate curves belonging to the given family $\mathcal{G}$.

2. When does the given, inherently infinite dimensional, interest-rate model admit a *finite dimensional Markovian state space realization*, that is, when can the r process be realized by a system of the form

$$\begin{aligned} dZ_t &= a(Z_t) dt + b(Z_t) dW_t \\ r_t(x) &= G(Z_t, x) \end{aligned} \quad (17)$$

where Z (interpreted as the state vector process) is a finite dimensional diffusion, $a(z)$, $b(z)$ and $G(z, x)$ are deterministic functions and W is the same Wiener process as in equation (11).

Consistency

A finitely parameterized family of forward rate curves is a real-valued function of the form $G(z, x)$ where

z lies in some open subset of R^k , that is, for each fixed parameter vector z , we have the forward rate curve $x \mapsto G(z, x)$. A typical example is the Nelson–Siegel forward curve family in equation (16). The mapping G can also be viewed as a mapping $G : \mathcal{Z} \rightarrow \mathcal{H}$, and we now define the *forward curve manifold* $\mathcal{G}$ as the set of all forward rate curves produced by this family, that is., $\mathcal{G} = \text{Im}(G)$. The main result concerning consistency is as follows (see [2]).

Theorem 2 (Consistency). *The forward curve manifold $\mathcal{G}$ is consistent with the forward rate process if and only if,*

$$\begin{aligned} G'_x(z) + D(r) - \frac{1}{2} \sigma'_r(r) \sigma(r) &\in \text{Im}[G'_z(z)] \\ \sigma(r) &\in \text{Im}[G'_z(z)] \end{aligned} \quad (18)$$

hold for all $z \in \mathcal{Z}$ with $r = G(z)$, where D is defined in equation (15)

Here, G'_z and G'_x denote the Frechet derivative of G with respect to z and x , respectively.

The invariance problem was originally posed and studied in [2] and then extended and studied in great depth in [9] and [10]. In particular, it is shown in [9] that no nondegenerate arbitrage-free forward rate model is consistent with the Nelson–Siegel family.

Finite Dimensional Markovian Realizations

The existence of finite dimensional Markovian realizations (FDR) was first studied in [7] and [17] where sufficient conditions were given for particular choices of volatility structures (see **Markovian Term Structure Models**) for a detailed discussion of these special cases and more references. General necessary and sufficient results were first obtained in [5] and extended in [11]. For an arbitrary SDE in Hilbert space (with the forward rate SDE as a special case) of the form

$$dr_t = \mu(r_t) dt + \sigma(r_t) dW_t \quad (19)$$

the main general result is as follows.

Theorem 3 *The infinite dimensional SDE above admits an FDR if and only if the Lie algebra generated by the vector fields $\mu - 1/2 \sigma' \sigma$ and σ (where σ' denotes the Frechet derivative) has finite dimension*

(evaluated pointwise) in a neighborhood of the initial point r_0 .

All known examples in the literature are easy consequences of this general result, which can also be extended to stochastic volatility.

A special case of a finite dimensional realization is when a HJM model generates a Markovian short rate process. This corresponds to the case of a two-dimensional realization with running time and short rate as Z-factors. For the case of a short rate dependent volatility, it was shown in [14] that this occurs if and only if the model is affine. This result has a remarkable extension in [12], where it is shown that all models admitting finite dimensional realizations are, in fact, affine.

References

- [1] Albeverio, S., Litvynov, A. & Mahnig, A. (2004). A model of the term structure of interest rates based on Levy fields, *Stochastic Processes and Their Applications* **114**(2), 251–263.
- [2] Björk, T. & Christensen, B. (1999). Interest rate dynamics and consistent forward rate curves, *Mathematical Finance* **9**(4), 323–348.
- [3] Björk, T., Di Masi, G., Kabanov, Y. & Runggaldier, W. (1997). Towards a general theory of bond markets, *Finance and Stochastics* **1**, 141–174.
- [4] Björk, T., Kabanov, Y. & Runggaldier, W. (1995). Bond market structure in the presence of a marked point process, *Mathematical Finance* **7**(2), 211–239.
- [5] Björk, T. & Svensson, L. (2001). On the existence of finite dimensional realizations for nonlinear forward rate models, *Mathematical Finance* **11**(2), 205–243.
- [6] Brace, A. & Musiela, M. (1994). A multifactor Gauss Markov implementation of Heath, Jarrow, and Morton, *Mathematical Finance* **4**, 259–283.
- [7] Cheyette, O. (1996). *Markov Representation of the Heath–Jarrow–Morton Model*, BARRA, Preprint.
- [8] Eberlein, E. & Raible, S. (1999). Term structure models driven by general Levy processes, *Mathematical Finance* **9**(1), 31–53.
- [9] Filipović, D. (1999). A note on the Nelson–Siegel family, *Mathematical Finance* **9**(4), 349–359.
- [10] Filipović, D. (2001). *Consistency Problems for Heath–Jarrow–Morton Interest Rate Models*, Springer Lecture Notes in Mathematics, Springer Verlag, Vol. 1760.
- [11] Filipović, D. & Teichmann, J. (2003). Existence of invariant manifolds for stochastic equations in infinite dimension, *Journal of Functional Analysis* **197**, 398–432.
- [12] Filipović, D. & Teichmann, J. (2004). On the geometry of the term structure of interest rates, *Proceedings of the Royal Society* **460**, 129–167.
- [13] Heath, D., Jarrow, R. & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation, *Econometrica* **60**, 77–105.
- [14] Jeffrey, A. (1995). Single factor Heath–Jarrow–Morton term structure models based on Markov spot interest rate dynamics, *Journal of Financial and Quantitative Analysis* **30**, 619–642.
- [15] Kennedy, D. (1994). The term structure of interest rates as a Gaussian random field, *Mathematica Finance* **4**, 247–258.
- [16] Musiela, M. (1993). *Stochastic PDE:s and term structure models*, Preprint.
- [17] Ritchken, P. & Sankarasubramanian, L. (1995). Volatility structures of forward rates and the dynamics of the term structure, *Mathematical Finance* **5**(1), 55–72.

TOMAS BJÖRK

Forward and Swap Measures

Forward and swap measures are instances of general numeraire measures and are useful in interest-rate modeling and derivatives valuation. They take a zero-coupon bond and an annuity as the numeraire, respectively. Accordingly, the forward price of any asset and the instantaneous and Libor forward rates are martingales under the forward measure. Likewise, the forward swap rate is a martingale under the swap measure.

Assuming deterministic forward Libor or swap rate volatility leads to industry-standard Black–Scholes type pricing formulae for caplets and swaptions, respectively. The forward measure has other interesting applications in option pricing and in the Libor market model.

The forward measure was implicit in Merton’s [7] extension of the Black–Scholes model to stochastic interest rates. Early development and application of the concept appeared in [4] and [2]. The swap measure was discussed heuristically in [8] and formalized in [5]. Forward and swap measures are instances of the change of numeraire, as described heuristically in [6] and formalized in [3].

Numeraire Measures

We take as given a stochastic basis $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ and a family $(A^i)_{i \in I}$ of semimartingales. Each A^i is viewed as the observable price process of a traded zero-dividend asset. We assume that this family is *arbitrage free* in that there exists a positive semimartingale $S = (S_t)$ such that $S_- > 0$, $S_0 = 1$, and SA^i is a martingale for every $i \in I$. Such a process S is called a *state price density* (or sometimes a *state price deflator*, cf. Duffie [1]), and we fix one such process. (In a complete market, it is unique.)

For our purposes here, we define an *asset* as a *semimartingale* C such that there exist a finite subset $J \subset I$ and bounded predictable processes δ^j satisfying $C = \sum_{j \in J} \delta^j A^j$ and $dC = \sum_{j \in J} \delta^j dA^j$. As such, an asset is the price process of a dynamic self-financing portfolio, for example, a (static) linear combination of A^i . The δ^j are called the *deltas* or the *hedge ratios*.

It can be shown that SC is a martingale for any asset C . This implies *the law of one price*: if two assets have almost surely the same prices at some time T , then they will have identical prices at all times $t < T$.

Let N be a *numeraire*, that is, a positive asset. For each $T > 0$ define the measure $\mathbb{P}^{N,T}$ on $\mathcal{F}_T$ via its Radon–Nikodym derivative by

$$\frac{d\mathbb{P}^{N,T}}{d\mathbb{P}} = \frac{S_T N_T}{N_0} \quad (1)$$

This is an equivalent probability measure and is called the associated *equivalent martingale measure* or *numeraire measure*. Since SN is a martingale, for any $s < T$, the restriction of $\mathbb{P}^{N,T}$ to $\mathcal{F}_s$ equals $\mathbb{P}^{N,s}$. (In an incomplete market, $\mathbb{P}^{N,T}$ depends on the choice of S .)

By an easy and well-known consequence of the Bayes’ rule, given a process C , the process SC is a $\mathbb{P}$ -martingale on $[0, T]$ if and only if C/N is a $\mathbb{P}^{N,T}$ -martingale on $[0, T]$. In particular, this holds for all assets C , yielding for $t \leq T$ the pricing formula,

$$C_t = N_t \mathbb{E}^{\mathbb{P}^{N,T}} \left[\frac{C_T}{N_T} \mid \mathcal{F}_t \right] \quad (2)$$

A useful technique is the *change of numeraire*. Suppose B is another numeraire and M is a $\mathbb{P}^{N,T}$ -martingale. Note that both $F := N/B$ and MF are $\mathbb{P}^{B,T}$ -martingales. Hence, using Itô’s product rule, $\int F_- dM + [M, F]$ is a $\mathbb{P}^{B,T}$ -local martingale. (Here, $[M, F]$ denotes the quadratic covariation of M and F .) Thus, dividing by F_- , so is $M + \int d[M, F]/F_-$. In particular, if F is continuous, then the $\mathbb{P}^{B,T}$ -drift of M equals $-d[M, \log F]$.

The Forward Measure

For a fixed maturity T , let us assume there exists an asset P^T such that $P_T^T = 1$. Such an asset, called the *T-maturity zero-coupon bond*, is necessarily positive and unique on $[0, T]$ by the law of one price. Its associated numeraire measure on $\mathcal{F}_T$ is called the *T-forward measure* and denoted $\mathbb{P}^T$. Its expectation operator is denoted $\mathbb{E}^T$. Since $P_T^T = 1$, by equation (1)

$$\frac{d\mathbb{P}^T}{d\mathbb{P}} = \frac{S_T}{P_0^T} \quad (3)$$

By equation (2), the T -forward price C/P^T of any asset C is a $\mathbb{P}^T$ -martingale on $[0, T]$ and

$$C_t = P_t^T \mathbb{E}^T[C_T | \mathcal{F}_t] \quad (4)$$

The price of a European thus equals the discounted expected value of its payoff.

Another important property is that *forward interest rates* are martingales under the forward measure. The simple (or Libor) T -forward rate $L^{T,\varepsilon}$ of length $\varepsilon > 0$ is defined by

$$L^{T,\varepsilon}_t := \frac{P^{T-\varepsilon}_t - P_t^T}{\varepsilon P_t^T} \quad (5)$$

Assuming that $P^{T-\varepsilon}$ is an asset, $P^{T-\varepsilon}/P^T$, and with it, $L^{T,\varepsilon}$ is a $\mathbb{P}^T$ -martingale.

As ε approaches zero, $L^{T,\varepsilon}_t$ approaches the instantaneous forward rate f_t^T defined by

$$f_t^T := -\frac{\partial \log(P_t^T)}{\partial T} \quad (6)$$

As such, in the limit, the instantaneous forward rate process f^T is also a $\mathbb{P}^T$ -martingale.

Option Pricing and Hedging

Consider a T -expiry option on an asset A (e.g. a stock or a bond) with time T -payoff $g(A_T)$, where $g(x)$ is a Borel function of linear growth. For example, $g(x) = \max(x - K, 0)$ for a call option struck at K . We wish to construct an asset C satisfying $C_T = g(A_T)$. From equation (4), we know that the only possible candidate is $C = P^T F$, where

$$F_t := \mathbb{E}^T[g(A_T) | \mathcal{F}_t] \quad (7)$$

This works if we assume the process $X := A/P^T$ is continuous and $F_t = f(t, X_t)$ for some $C^{1,2}$ function $f(t, x)$. The desired deltas (hedge ratios) are then simply given by

$$\delta_t^A := \frac{\partial f}{\partial x}(X_t, t), \quad \delta^P := F - \delta^A X \quad (8)$$

Indeed, $C := P^T F = \delta^A A + \delta^P P^T$ obviously. Moreover, since both X and F are $\mathbb{P}^T$ -martingales, Itô's formula implies $dF = \delta^A dX$. An application of Itô's product rule known as the *numeraire-invariance theorem* cf. [1]), then shows that $dC = \delta^A dA + \delta^P dP^T$.

The above Markovian assumption that $F_t = f(t, X_t)$ for some function $f(t, x)$ is generally satisfied when X is a positive diffusion, specifically when $d[X]_t = X_t^2 \sigma^2(t, X_t) dt$ for some positive continuous bounded function $\sigma(t, x)$. (Here $[X]$ denotes the quadratic variation of X .) This is equivalent to $dX_t = X_t \sigma(t, X_t) dW_t$ for some $\mathbb{P}^T$ -Brownian motion W . In this case, $f(t, x)$ is basically obtained as $\mathbb{E}^T[g(X_T) | X_t = x]$. By Itô's formula, $f(t, x)$ satisfies

$$\frac{\partial f}{\partial t} + \frac{1}{2} x^2 \sigma^2(t, x) \frac{\partial^2 f}{\partial x^2} = 0 \quad (f(x, T) = g(x)) \quad (9)$$

Closed-form Solutions

The classical case assumes as in [7] that the forward price volatility $\sigma(t, X_t)$ is deterministic, that is, independent of X . Then X is a $\mathbb{P}^T$ log-Gaussian martingale, and hence conditioned on time t , $X_T = A_T$ is $\mathbb{P}^T$ -lognormally distributed with mean X_t and log-variance $\int_t^T \sigma^2(s) ds$. As such, for, say, a call option with payoff function $g(x) = \max(x - K, 0)$, equation (7) readily yields

$$C(t) = K P_t^T C_{BS} \left(\frac{A_t}{P_t^T K}, \int_t^T \sigma^2(s) ds \right) \quad (10)$$

where, denoting the standard normal distribution function by $N(\cdot)$,

$$C_{BS}(x, v) := x N \left(\frac{\log(x)}{\sqrt{v}} + \frac{\sqrt{v}}{2} \right) - N \left(\frac{\log(x)}{\sqrt{v}} - \frac{\sqrt{v}}{2} \right) \quad (11)$$

Specific examples are the *Vasicek* or more general *Gaussian interest-rate models* (see **Gaussian Interest-Rate Models**) where the deterministic (forward) zero-coupon bond price volatilities are determined endogenously in terms of mean reversion and other model parameters. For zero-coupon bond options, equation (7) can be computed in the *Cox–Ingersoll–Ross model* (see **Cox–Ingersoll–Ross (CIR) Model**) and the *quadratic Gaussian model* (see **Quadratic Gaussian Models**) in terms of the noncentral chi-squared distribution function. This is derived by showing that the spot interest rate $r_T := f_T^T$ is noncentral chi-squared distributed under the forward measure $\mathbb{P}^T$ (e.g., [4]).

Cap Pricing

A cap is a portfolio of consecutive caplets, and a caplet of maturity T , length ε , and strike rate K is an option with payoff $\varepsilon \max(L_{T-\varepsilon}^{T,\varepsilon} - K, 0)$ at time T . A caplet is actually equivalent to a zero-coupon bond put option, so a bond option model as in the previous section is applicable. However, more directly, by the pricing formula given by equation (4) the caplet price C_t is given by

$$C_t = \varepsilon P_t^T \mathbb{E}^T[\max(L_{T-\varepsilon}^{T,\varepsilon} - K, 0) | \mathcal{F}_t] \quad (12)$$

By the section The Forward Measure, the forward Libor process $L^{T,\varepsilon}$ is a martingale under the forward measure $\mathbb{P}^T$. Hence, if its volatility $\sigma(t)$ is deterministic, it is log-Gaussian and we get, as in, the section Closed-form Solutions,

$$C_t = \varepsilon K P_t^T \text{CBS} \left(\frac{L_t^{T,\varepsilon}}{K}, \int_t^T \sigma^2(t) dt \right) \quad (13)$$

The Forward Measure by Changing the Risk-neutral Measure

Let $r_t := f_t^t$ denote the spot interest rate. One often starts with the “money market” asset $\exp(\int_0^t r_t dt)$ as the numeraire and uses its equivalent martingale measure, often called the *risk-neutral measure* and denoted by $\mathbb{Q}$. Accordingly, $\exp(-\int_0^t r_t dt)C$ is a $\mathbb{Q}$ -martingale for any asset C . One can then change the numeraire to the T -maturity bond P^T and obtain the T -forward measure $\mathbb{P}^T$ by the formula

$$\frac{d\mathbb{P}^T}{d\mathbb{Q}} = \frac{1}{P_0^T} \exp \left(- \int_0^T r_t dt \right) \quad (14)$$

Since the forward rate process f^T is a $\mathbb{P}^T$ -martingale, it follows from the section Numeraire Measures that when P^T is continuous, the $\mathbb{Q}$ -drift of f^T equals $-d[\log P^T, f^T]$.

Libor Market Model SDE in the Forward Measure

Consider a sequence of dates $0 < T_1 \dots < T_{n+1}$, for example, equidistant semiannually. Given “daycount fractions” $\varepsilon_i \approx T_{i+1} - T_i$, the forward Libor rates L_t^i

are defined as in the section The Forward Measure by

$$\varepsilon_i L_t^i = \frac{P_t^{T_i}}{P_t^{T_{i+1}}} - 1 \quad (t \leq T_i, i = 1, \dots, n) \quad (15)$$

Evidently, L^i is a $\mathbb{P}^{T_{i+1}}$ -martingale on $[0, T_i]$.

In some applications such as valuation by Monte Carlo simulation, it is necessary to determine the dynamics of all the forward Libor processes L^i under the same measure. One appropriate measure is the *spot-Libor measure*, a simple-compounding analog of the risk-neutral measure that takes as numeraire a “rolling zero-coupon bond” (cf. [5]). Another convenient measure is the *terminal measure*, that is, the T_{n+1} -forward measure $\mathbb{P}^{T_{n+1}}$. Let $W^1, \dots, W^n$ be $\mathbb{P}^{T_{n+1}}$ -Brownian motions with correlations ρ^{ij} , that is, $d[W^i, W^j]_t = \rho^{ij} dt$. Assume

$$dL_t^i = \mu_t^i dt + \sigma_t^i dW_t^i \quad (16)$$

for some predictable processes μ^i and σ^i . For example, in the deterministic-volatility Libor market model, $\sigma_t^i = \sigma_i(t)L_t^i$ for some deterministic functions $\sigma_i(t)$. Since L^i is a $\mathbb{P}^{T_{i+1}}$ -martingale, it follows from the section Numeraire Measures that $\mu^i dt = -d[L^i, \log F]$, where

$$F := \frac{P^{T_{i+1}}}{P^{T_{n+1}}} = \prod_{j=i+1}^n (1 + \varepsilon_j L^j) \quad (17)$$

Therefore, the drift of the forward Libor rate in the terminal measure is given by

$$\begin{aligned} \mu^i dt &= - \sum_{j=i+1}^n \frac{\varepsilon_j d[L^i, L^j]}{1 + \varepsilon_j L^j} \\ &= - \sum_{j=i+1}^n \frac{\varepsilon_j \sigma^i \sigma^j \rho^{ij}}{1 + \varepsilon_j L^j} dt \end{aligned} \quad (18)$$

The Swap Measure

Let T_i and ε_i be as in the previous section. For each $1 \leq i < j$, the swap measure $\mathbb{P}^{ij}$ is defined on $\mathcal{F}_{T_{i+1}}$ as the equivalent martingale measure associated with the *annuity numeraire*

$$A^{ij} := \varepsilon_i P^{T_{i+1}} + \dots + \varepsilon_{j-1} P^{T_j} \quad (19)$$

The forward swap rate S_t^{ij} with start date T_i and end date T_j is defined for $t \leq T_i$ by

$$S_t^{ij} := \frac{P_t^{T_i} - P_t^{T_j}}{A_t^{ij}} \quad (20)$$

It follows from the section Numeraire Measures that S^{ij} is a martingale under the swap measure $\mathbb{P}^{ij}$.

The main application of the swap measure is to European swaptions, that is, options to enter an interest-rate swap at a fixed strike rate. Specifically, a payer swaption with start date T_i , end date T_j , expiration $T \leq T_j$ and strike rate K has the payoff C_T at time T given by

$$C_T = A_T^{ij} \max(S_T^{ij} - K, 0) \quad (21)$$

(When $j = i + 1$, a payer swaption is just a caplet). Arguments similar to those in the section Option Pricing and Hedging show that the swaption is replicable under general diffusion assumptions, for example, when S^{ij} has deterministic volatility or, more generally, when it is a diffusion process under the swap measure $\mathbb{P}^{ij}$. The swaption price process C is then uniquely characterized by C/A^{ij} being a $\mathbb{P}^{ij}$ martingale, implying by equation (21) that

$$C_t = A_t^{ij} \mathbb{E}^{\mathbb{P}^{ij}} [\max(S_T^{ij} - K, 0) | \mathcal{F}_t] \quad (22)$$

When S^{ij} has a deterministic volatility $\sigma_{ij}(t)$ (i.e., $d[S^{ij}]_t = \sigma_{ij}^2(t)(S_t^{ij})^2 dt$), this yields

$$C_t = K A_t^{ij} \text{C}_{\text{BS}} \left(\frac{S_t^{ij}}{K}, \int_t^T \sigma_{ij}^2(t) dt \right) \quad (23)$$

The market uses this formula to quote swaptions, namely a constant volatility σ_{ij} is quoted from which one computes the swaption price by equation (23). Receiver swaptions are treated similarly.

The valuation of Bermudan options is more complex. Here, the swaption can be exercised at any

time $T_i, \dots, T_{j-1}$. One approach, known as the *coterminous swap market model*, assumes that the forward swap rates $S^{ij}, \dots, S^{j-1,j}$ all have deterministic volatilities. According to equation (23), the model is then automatically calibrated to all the European swaptions with start dates $T_i, \dots, T_{j-1}$ and the same end date T_j , thus ruling out obvious arbitrage opportunities.

Constructs similar to the swap measure have been applied to credit default swaptions.

References

- [1] Duffie, D. (2001). *Dynamic Asset Pricing Theory*, 3rd Edition, Princeton University Press.
- [2] Geman, H. (1989). *The Importance of the Forward Neutral Probability in a Stochastic Approach of Interest Rates*, ESSEC working paper.
- [3] Geman, H., El-Karoui, N. & Rochet, J.C. (1995). Change of numeraire, change of probability measure, and option pricing, *Journal of Applied Probability* **32**, 443–458.
- [4] Jamshidian, F. (1987). *Pricing of Contingent Claims in the One-Factor Term Structure Model*, working paper, appeared in *Vasicek and Beyond*, Risk Publications, (1996).
- [5] Jamshidian, F. (1997). Libor and swap market model and measures. *Finance and Stochastics* **1**, 293–330.
- [6] Margrabe, W. (1978). The value of an option to exchange one asset for another, *Journal of Finance* **33**, 177–186.
- [7] Merton, R. (1973). Theory of rational option pricing, *Bell Journal of Economics* **4**(1), 141–183.
- [8] Neuberger, A. (1990). *Pricing Swap Options Using the Forward Swap Market*, IFA Preprint.

Related Articles

Caps and Floors; Change of Numeraire; Exchange Options; Itô's Formula; LIBOR Market Model; LIBOR Rate; Martingales; Term Structure Models; Swap Market Models.

FARSHID JAMSHIDIAN

Term Structure Models

Term structure models describe the behavior of interest rates as a function of time and term (time to maturity). As a function of time, rates behave as stochastic processes (see Figure 1). As a function of term, interest rates on a given date form a *yield curve* (see Figure 2).

Interest rates of different maturities behave as a joint stochastic process. Not all joint processes, however, can describe interest-rate behavior in an efficient market. For instance, suppose that a term structure model postulates that rates of all maturities change in time by equal amounts, that is, that yield curves move by parallel shifts (which, empirically, appears to be a reasonable first-order approximation). It can be shown that in this case a portfolio consisting of a long bond and a short bond would always outperform a medium-term bond with the same Macaulay duration (*see Bond*). In an efficient market, supply and demand would drive the price of the medium maturity bond down and the prices of the long and short bonds up. As this would cause the yield on the medium bond to increase and the yields on the long and short bonds to decrease, the yield curves would not stay parallel. This model therefore cannot describe interest-rate behavior.

In order that riskless arbitrage opportunities are absent, the joint process of interest-rate behavior must satisfy some conditions. Determining these conditions and finding processes that satisfy them is the purpose of term structure models.

The joint stochastic process will be driven by a number of sources of uncertainty. For continuous processes, the sources of uncertainty are often specified as Wiener processes. If the evolution of the yield curve can be represented by Markovian state variables, these variables are called *factors*.

Let $B(t, T)$ be the price at time t of a default-free zero-coupon bond maturing at time T with unit maturity value. *Yield to maturity* $R(t, s)$ at time t with term s is defined as the continuously compounded rate of return on the bond,

$$R(t, s) = -\frac{1}{s} \log B(t, t + s) \quad (1)$$

The instantaneous interest rate will be called the *short rate*,

$$r(t) = \lim_{s \rightarrow 0} R(t, s) \quad (2)$$

An asset accumulating interest at the short rate will be called the *money market account*,

$$\beta(t) = \exp \left(\int_0^t r(\tau) d\tau \right) \quad (3)$$

Forward rates $f(t, T)$ are defined by the equation

$$B(t, T) = \exp \left(- \int_t^T f(t, \tau) d\tau \right) \quad (4)$$

One-factor Models

A general theory of one-factor term structure models was given by Vasicek [9]. He assumed the following:

1. The short rate follows a continuous Markov process.
2. The price $B(t, T)$ of a bond is determined by the assessment at time t of the segment $\{r(\tau), t \leq \tau \leq T\}$ of the short-rate process over the term of the bond.
3. The market is efficient; that is, there are no transaction costs, information is available to all investors simultaneously, and every investor acts rationally (prefers more wealth to less, and uses all available information).

Assumption 3 implies that investors have homogeneous expectations and that no profitable riskless arbitrage is possible.

By assumption 1, the development of the short rate over an interval (t, T) , $t \leq T$, given its values prior to time t , depends only on the current value $r(t)$. Assumption 2 then implies that the price $B(t, T)$ is a function of $r(t)$. Thus, the value of the short rate is the only state variable for the whole term structure.

Let the dynamics of the short rate be given by

$$dr(t) = \zeta(r, t) dt + \varphi(r, t) dW(t) \quad (5)$$

where $W(t)$ is a Wiener process. Denote the mean and variance of the instantaneous rate of return on the bond with price $B(t, T)$ by $\mu(t, T)$ and $\sigma^2(t, T)$, respectively,

$$\frac{dB(t, T)}{B(t, T)} = \mu(t, T) dt - \sigma(t, T) dW(t) \quad (6)$$

Consider an investor who at time t issues an amount w_1 of a bond with maturity date T_1 , and

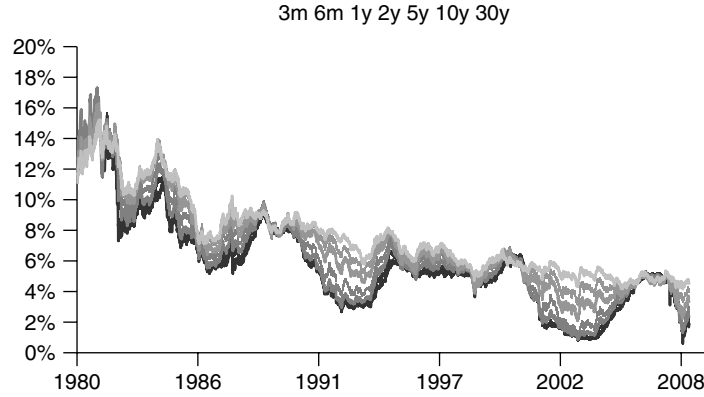


Figure 1 US Treasury Yields

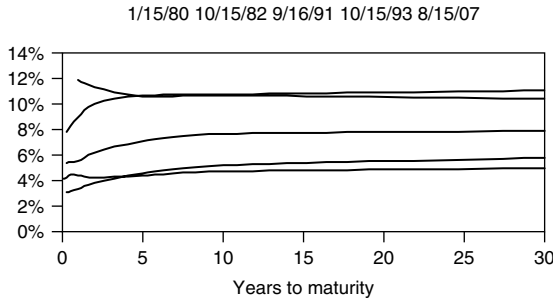


Figure 2 US Treasury Yield Curves

simultaneously buys an amount w_2 of a bond with maturity date T_2 . Suppose the amounts w_1 and w_2 are chosen to be proportional to $\sigma(t, T_2)$ and $\sigma(t, T_1)$, respectively. Then the position is instantaneously riskless, and should realize the short rate of return $r(t)$. It follows that the ratio $(\mu(t, T) - r(t))/\sigma(t, T)$ is independent of T . Its common value $\lambda(t)$ is called the *market price of risk*, as it specifies the increase in the expected rate of return on a bond per an additional unit of risk. We thus have

$$\mu(t, T) = r(t) + \lambda(t)\sigma(t, T) \quad (7)$$

Applying Ito's lemma to the price $B(t, T) = B(t, T, r)$ and comparing with equation (6) yields

$$\frac{\partial B}{\partial t} + (\zeta + \varphi\lambda)\frac{\partial B}{\partial r} + \frac{1}{2}\varphi^2\frac{\partial^2 B}{\partial r^2} - rB = 0 \quad (8)$$

The bond price is subject to the boundary condition $B(T, T) = 1$.

The solution to equation (8) is given by

$$B(t, T) = E_t \exp \left(- \int_t^T r(\tau) d\tau - \frac{1}{2} \int_t^T \lambda^2(\tau) d\tau + \int_t^T \lambda(\tau) dW(\tau) \right) \quad (9)$$

This equation, called the *fundamental bond pricing equation* (the *Vasicek equation*), fully describes the term structure and its behavior.

Model Examples

Various specific cases have been proposed in the literature. Vasicek [9] gives an example of a term structure model in which the short rate follows a mean reverting random walk (the Ornstein–Uhlenbeck process, see **Ornstein–Uhlenbeck Processes**)

$$dr = \alpha(\theta - r) dt + \varphi dW \quad (10)$$

and the market price of risk $\lambda(t, r) = \lambda$ is constant. In that case, the expectation in equation (9) can be evaluated explicitly to give

$$B(t, T) = \exp \left(\frac{1}{\alpha} (1 - e^{-\alpha(T-t)}) (R(\infty) - r) - (T - t)R(\infty) - \frac{\varphi^2}{4\alpha^3} (1 - e^{-\alpha(T-t)})^2 \right) \quad (11)$$

where

$$R(\infty) = \theta + \lambda\varphi/\alpha - \frac{1}{2}\varphi^2/\alpha^2 \quad (12)$$

Interest rates are Gaussian. The advantage of this specification is its tractability. A drawback is that interest rates can become negative.

Cox *et al.* [3] derive a model in which

$$dr = \alpha(\theta - r) dt + \varphi\sqrt{r} dW \quad (13)$$

and the market price of risk $\lambda(t, r) = \eta\sqrt{r}$. In this case, the bond prices can also be explicitly given (*see Cox–Ingersoll–Ross (CIR) Model*). They have the form

$$B(t, T) = A(t, T) \exp(-D(t, T)r(t)) \quad (14)$$

Interest rates are always nonnegative.

Hull and White [8] extended these two models by allowing the parameters in equations (10) and (13), as well as the market price of risk, to be time dependent. This has the advantage that the model can be made consistent with initial data. For instance, by making θ a function of time, the model can be made to exactly fit the initial term structure of interest rates (which is not possible with time-homogeneous models). Similarly, making the volatility φ a function of time allows calibration of the model to the term structure of swaption volatilities. Hull and White give closed-form solutions for bond prices for what they call the *extended Vasicek* and the *extended Cox–Ingersoll–Ross* models. These cases belong to the class of models that Duffie and Kan [4] call the *affine term structure models* (*see Affine Models*), in which bond prices have the form (14).

Black *et al.* [1] and Black and Karasinski [2] give a model with

$$d \log r = \alpha(t)(\log \theta(t) - \log r) dt + \varphi(t) dW \quad (15)$$

In this model, bond prices cannot be given in closed-form formulas, but can be calculated numerically. Interest rates are lognormal. Lognormal models have regularity issues, for example, they produce infinite Eurodollar future prices (*see Eurodollar Futures and Options*).

The term structure theory generalizes easily to multiple factors and multiple sources of uncertainty. In fact, the bond pricing equation (9) is universally valid for any arbitrage-free term structure model. If W, φ, λ are vectors, their products are interpreted as inner products.

Contingent Claim Pricing

One of the main tasks of term structure models in applications is pricing of *interest-rate-contingent claims* (interest-rate derivatives). This could be approached in several ways. For one-factor models it can be shown, by means of an arbitrage argument similar to that above for bonds, that the price $P(t)$ of any interest-rate derivative satisfies the partial differential equation (8). The valuation of the derivative is then accomplished by solving that equation subject to boundary conditions that describe the derivative asset payouts. If a closed-form solution cannot be given, the equation can be solved numerically in a tree or a finite difference lattice. A more general method is to realize that such a solution has the form

$$P(t) = E_t P(s) \exp \left(- \int_t^s r(\tau) d\tau - \frac{1}{2} \int_t^s \lambda^2(\tau) d\tau + \int_t^s \lambda(\tau) dW(\tau) \right) \quad (16)$$

This equation holds even in the cases where there are no Markovian state variables. To calculate the expectation in equation (16), however, is typically more difficult than solving a partial differential equation.

The modern theory of derivative asset pricing (*see Harrison and Kreps [5]*) introduces a change of probability measure as the basic pricing tool. There exists an equivalent probability measure P^* , called the *risk-neutral measure*, such that the value of any asset expressed in units of the money market account $\beta(t)$ follows a martingale under that measure,

$$\frac{P(t)}{\beta(t)} = E_t^* \frac{P(s)}{\beta(s)} \quad (17)$$

The process

$$W^*(t) = W(t) - \int_0^t \lambda(\tau) d\tau \quad (18)$$

is a Wiener process under the risk-neutral probability measure P^* .

If current bond prices are considered given, interest-rate derivatives can be priced without knowing the market price of risk $\lambda(t)$ by writing the dynamics of interest rates directly in terms of the process $W^*(t)$. From equations (6) and (7), bond

prices are subject to

$$\frac{dB(t, T)}{B(t, T)} = r(t) dt - \sigma(t, T) dW^*(t) \quad (19)$$

Integrating equation (19) with respect to t and differentiating with respect to T yields

$$\begin{aligned} f(t, T) - f(0, T) \\ = \int_0^t \varphi(\tau, T) \sigma(\tau, T) d\tau + \int_0^t \varphi(\tau, T) dW^*(\tau) \end{aligned} \quad (20)$$

where $\varphi(t, T)$ is the volatility of the forward rate $f(t, T)$ and

$$\sigma(\tau, T) = \int_{\tau}^T \varphi(\tau, s) ds \quad (21)$$

Thus, knowledge of the initial term structure $f(0, T)$, $T \geq 0$ and of the forward-rate volatilities is sufficient for pricing interest-rate-contingent claims. This was proposed in essence by Ho and Lee [7] and later formalized by Heath *et al.* [6] (see **Heath–Jarrow–Morton Approach**).

References

- [1] Black, F., Derman, E. & Toy, W. (1990). A one-factor model of interest rates and its application to Treasury bond options, *Financial Analysts Journal* January-February, 33–39.
- [2] Black, F. & Karasinski, P. (1991). Bond and option pricing when interest rates are lognormal, *Financial Analysts Journal* July-August, 52–59.
- [3] Cox, J., Ingersoll, J. Jr. & Ross, S. (1985). A theory of the term structure of interest rates, *Econometrica* **53**, 385–407.
- [4] Duffie, D. & Kan, R. (1996). A yield-factor model of interest rates, *Mathematical Finance* **6**, 379–406.
- [5] Harrison, J.M. & Kreps, D.M. (1979). Martingales and arbitrage in multiperiod security markets, *Journal of Economic Theory* **20**, 381–408.
- [6] Heath, D., Jarrow, R. & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation, *Econometrica* **60**, 77–105.
- [7] Ho, T.S.Y. & Lee, S.-B. (1986). Term structure movements and pricing interest rate contingent claims, *Journal of Finance* **41**, 1011–1028.
- [8] Hull, J. & White, A. (1990). Pricing interest-rate derivative securities, *Review of Financial Studies* **3**, 573–592.
- [9] Vasicek, O.A. (1977). An equilibrium characterization of the term structure, *Journal of Financial Economics* **5**, 177–188.

Related Articles

Affine Models; Bond; Caps and Floors; Cox–Ingersoll–Ross (CIR) Model; Heath–Jarrow–Morton Approach.

OLDRICH ALFONS VASICEK

Cox–Ingersoll–Ross (CIR) Model

The Cox–Ingersoll–Ross (CIR) model, one of the most well-known short-rate models (see **Term Structure Models**), was proposed in 1985 by Cox, Ingersoll, and Ross (see **Ross, Stephen**). In their pioneering work [3], they use an equilibrium approach to derive an explicit formula for the interest rate as a function of the wealth and the state of technology. On the basis of economic arguments, they specify their general framework in [4] and obtain the following dynamics for the short rate (r_t , $t \geq 0$) under the objective probability measure $\mathbb{P}^o$ with a risk premium factor Λ :

$$t \geq 0, dr_t = [\kappa\theta - (\kappa + \Lambda)r_t] dt + \sigma\sqrt{r_t} dW_t^o \quad (1)$$

Here, the process (W_t^o , $t \geq 0$) is a standard Brownian motion, and the real parameters κ , θ , σ satisfy $\kappa\theta \geq 0$ and $\sigma > 0$. To deal with pricing, we, however, consider in the sequel the dynamics under the risk-neutral measure $\mathbb{P}$ (see **Hedging**). The usual assumption is to take a risk premia function equal to $\Lambda\sqrt{r}/\sigma$ (see **Risk Premia**). This choice allows to keep a similar dynamics under $\mathbb{P}$:

$$t \geq 0, dr_t = \kappa(\theta - r_t) dt + \sigma\sqrt{r_t} dW_t \quad (2)$$

$W_t = W_t^o - \int_0^t (\Lambda\sqrt{r_s}/\sigma) ds$ being a Brownian motion under $\mathbb{P}$. Indeed, using $\int_0^t \sqrt{r_s} dW_s^o = [r_t - r_0 - \kappa\theta t + \int_0^t (\kappa + \Lambda)r_s ds]/\sigma$, we can check from equation (3) that $\mathbb{E}^o[\exp[\int_0^t (\Lambda\sqrt{r_s}/\sigma) dW_s^o - \frac{1}{2} \int_0^t (\Lambda/\sigma)^2 r_s ds]] = 1$, so $\mathbb{P}$ is indeed equivalent to $\mathbb{P}^o$. The parameters have a clear interpretation. First, σ determines the volatility of the interest rate. Moreover, for the common practical choice $\kappa, \theta > 0$ that we assume in the following, the short rate has a mean reversion toward θ with a speed driven by κ . It is known that equation (2) has a (pathwise) unique non-negative solution for any starting value $r_0 \geq 0$. Furthermore, the short rate remains positive at any time as long as $r_0 > 0$ and $2\kappa\theta \geq \sigma^2$. Nonnegativity is, of course, a nice feature when modeling interest rates. This is the main qualitative difference between CIR and Vasicek models (see **Term Structure Models**)

that have otherwise similar properties for pricing derivatives.

Analytical Results

Beyond the natural meaning of its parameters, the CIR model strength is to provide analytical formulas for the main financial quantities. It belongs to the class of affine models (see **Affine Models**), which means that the Laplace transform of the joint distribution ($r_t, \int_0^t r_s ds$) is known to be for $\lambda, \mu > 0$,

$$\mathbb{E} \left[e^{-\lambda r_t - \mu \int_0^t r_s ds} \right] = A_{\lambda, \mu}(t) \exp(-r_0 B_{\lambda, \mu}(t)) \quad (3)$$

We have the following formulas (see [10]):

$$A_{\lambda, \mu}(t) = \left[\frac{2\gamma_\mu e^{t \frac{\gamma_\mu + \kappa}{2}}}{\{\sigma^2 \lambda (e^{\gamma_\mu t} - 1) + \gamma_\mu\}} \right]^{\frac{2\kappa\theta}{\sigma^2}},$$

$$B_{\lambda, \mu}(t) = \frac{\lambda[\gamma_\mu + \kappa + e^{\gamma_\mu t}(\gamma_\mu - \kappa)] + 2\mu(e^{\gamma_\mu t} - 1)}{\sigma^2 \lambda (e^{\gamma_\mu t} - 1) + \gamma_\mu - \kappa + e^{\gamma_\mu t}(\gamma_\mu + \kappa)} \quad (4)$$

with $\gamma_\mu = \sqrt{\kappa^2 + 2\sigma^2\mu}$. An analogous formula to equation (3) holds for the Fourier transform, handling complex logarithms with care, and even for a wider range of complex values of λ and μ as long as the left-hand side is well defined. Thanks to equation (3), the law of r_t is known. Defining $c_t = \frac{4\kappa}{\sigma^2(1 - e^{-\kappa t})}$, $c_t r_t$ is distributed as a chi-square distribution with $\nu = \frac{4\kappa\theta}{\sigma^2}$ degrees of freedom and noncentrality parameter $d_t = c_t r_0 e^{-\kappa t}$. More explicitly, this means that r_t has the following density function for $r > 0$,

$$\sum_{i=0}^{\infty} \frac{e^{-\frac{d_t}{2}} \left(\frac{d_t}{2}\right)^i}{i!} \frac{c_t/2}{\Gamma(i + \frac{\nu}{2})} \left(\frac{c_t r}{2}\right)^{i-1+\frac{\nu}{2}} e^{-\frac{c_t r}{2}} \quad (5)$$

In particular, one can see that r_t converges to a steady-state distribution with density $\frac{(c_\infty/2)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} r^{\frac{\nu}{2}-1} e^{-\frac{c_\infty}{2} r}$ where $c_\infty = 4\kappa/\sigma^2$ when $t \rightarrow +\infty$. This is a Gamma distribution with mean θ and variance

$\theta\sigma^2/(2\kappa)$. It is the stationary law of the stochastic differential equation (2).

Derivative Pricing

Under a short-rate model, the initial price of a zero-coupon bond with maturity $T > 0$ (see definition in **Bond**) is given by $\mathbb{E}[e^{-\int_0^T r_s ds}]$. It is here analytically known and is equal to $A_{0,1}(T) \exp(-r_0 B_{0,1}(T))$. More generally, the price at time $t > 0$ of a zero-coupon bond with maturity T is given by

$$P(r_t, t, T) = A_{0,1}(T - t) \exp(-r_t B_{0,1}(T - t)) \quad (6)$$

The CIR model also provides closed form formulas for some option prices. For instance, let us consider a call option with strike K and maturity T , written on a zero-coupon bond with maturity $S \geq T$. Its initial price is given by

$$\mathbf{C} = \mathbb{E}[e^{-\int_0^T r_s ds} (P(r_T, T, S) - K)^+] \quad (7)$$

To calculate $\mathbf{C}$, we use another nice feature of the CIR model: the short-rate distribution is known under the forward measure. Let us recall that for a fixed maturity $T > 0$, the T -forward measure is defined by

$$\mathbb{P}^T(A) = \frac{\mathbb{E}\left[e^{-\int_0^T r_s ds} \mathbf{1}_A\right]}{P(r_0, 0, T)} \quad (8)$$

for any event A anterior to time T (see **Forward and Swap Measures**), which amounts to taking the zero-coupon bond as a numeraire. Under $\mathbb{P}^T$, $(r_t, 0 \leq t \leq T)$ solves the following SDE

$$dr_t = [\kappa\theta - (\kappa + \sigma^2 B_{0,1}(T - t))r_t] dt + \sigma\sqrt{r_t} dW_t^T \quad (9)$$

where $(W_t^T, 0 \leq t \leq T)$ is a Brownian motion under $\mathbb{P}^T$. This diffusion is again of the affine type and is tractable. In particular, for $t \in [0, T]$, the law of r_t under $\mathbb{P}^T$ is known: it is distributed as $1/(2\alpha_T(t))$ times a chi-square random variable with ν degrees of freedom and noncentrality parameter $\frac{8r_0\gamma_1^2 e^{\gamma_1 t}}{\sigma^4(e^{\gamma_1 t} - 1)^2 \alpha_T(t)}$, where $\alpha_T(t) = \frac{2\gamma_1}{\sigma^2(e^{\gamma_1 t} - 1)} +$

$\frac{\kappa + \gamma_1}{\sigma^2} + B_{0,1}(T - t)$ (see [2, 4, 8]). Then, from equations (6 and 7), we easily get the call price

$$\begin{aligned} \mathbf{C} &= P(r_0, 0, S) \mathbb{P}^S(r_T \leq r^*) \\ &\quad - K P(r_0, 0, T) \mathbb{P}^T(r_T \leq r^*) \end{aligned} \quad (10)$$

where $r^* = \ln(A_{0,1}(S - T)/K)/B_{0,1}(S - T)$. Obviously, we have a similar formula for put options. Nonetheless, options on zero-coupon bonds are not standard in practice, and products like caps, floors, and swaptions are mostly preferred to get hedge against interest-rate fluctuations (see **Caps and Floors**). A well-known relation is that the price of a floorlet (resp. caplet) between maturities T and S can be written as a simple function of a call (resp. put) option on the zero-coupon bond between T and S (see, e.g., [2]). This way, from equation (10), we derive closed form formulas for cap and floor prices also. Receiver (resp. payer) European swaptions can readily be seen as call (resp. put) option with a unit strike on a bond whose coupon rate corresponds to the swaption strike (see **Bond Options**). Thus, denoting the maturity by T , and with $T < S_1 < \dots < S_n$, the payment grid, its price has the following form $\mathbb{E}[e^{-\int_0^T r_s ds} (\sum_{i=1}^n \alpha_i P(r_T, T, S_i) - 1)^+]$ with $\alpha_i \geq 0$, $\sum_{i=1}^n \alpha_i > 1$. Thanks to the strike decomposition introduced by Jamshidian [7], it turns out to be a combination of call prices on zero-coupon bonds. Indeed, $P(r, T, S)$ being decreasing with respect to r , there is a unique $\rho > 0$ such that $\sum_{i=1}^n \alpha_i P(\rho, T, S_i) = 1$, and we have $(\sum_{i=1}^n \alpha_i [P(r_T, T, S_i) - P(\rho, T, S_i)])^+ = \sum_{i=1}^n \alpha_i (P(r_T, T, S_i) - P(\rho, T, S_i))^+$.

Calibration

Let us turn to the calibration of the CIR model to the market prices. We have analytical formulas for zero-coupon bond, cap, and floor prices. Swaption prices can also be computed very quickly, ρ being obtained, for example, by dichotomy. Therefore, the distance between the market prices and the theoretical ones obtained from the CIR model can be computed quickly and minimized with the use of any optimization algorithm. In this way, we can identify optimal parameters r_0 , κ , θ , and σ . In practice, it is likely to better fit swaption prices than cap and floor prices. Indeed, they do not only

describe the evolution of each single rate for different maturities but also the dependence between them. Unfortunately, these four parameters are anyway not enough in practice, if we want to accurately capture all the market data. For this reason, many works have been done to extend the CIR model. While doing this, the challenge is to preserve the nice analytical tractability of the CIR model. Without being exhaustive, we mention here the extended CIR model [9] where parameters κ , θ and σ are supposed to be time dependent, and the particular case where $\kappa(t)\theta(t)/\sigma^2(t)$ is constant that allow to preserve some closed formulas [8]. Other generalizations have been proposed, like adding a deterministic shift or another independent CIR process (see [2] for some numerical experiments). Most of these extensions are embedded in the general affine framework described in [5] (see **Term Structure Models**).

Monte Carlo Simulation

Finally, it is important to mention the simulation issues concerning square-root diffusions. This topic goes beyond the world of interest derivatives because these diffusions are widespread in finance, such as in the Heston model (see **Heston Model**). First, exact simulation is possible for the CIR model since we know how to simulate a noncentral chi-square distribution [6]. However, in some applications, it may be more convenient to use discretization schemes that often lead to smaller computation times. We face the difficulty that usual schemes such as Euler and Milstein generally fail. This is due to the square root in the CIR dynamics, which is non-Lipschitzian near the origin and is not defined for negative values. Tailored schemes should then be considered as in [1]. More information on the simulation of the CIR process can be found in **Simulation of Square-root Processes**.

Conclusion

The fundamental features of the CIR model—intuitive parameterization, nonnegativity, pricing formulas for main options—explain why it has been widely used for hedging risk on interest-rate derivatives. Nowadays, owing to the complexity of the fixed income market, more sophisticated models are required, such as Libor or Swap Market models (see

LIBOR Market Model and **Swap Market Models**). Nonetheless, the CIR model is often used as a building block for more complex models. Many extensions that rely on the same mathematical properties are widespread for interest-rate modeling (multifactor quadratic Gaussian models (see **Quadratic Gaussian Models**, affine models (see **Affine Models**), the Heston model for equity (see **Heston Model**), or the Duffie–Singleton model on credit derivatives (see **Duffie–Singleton Model**).

References

- [1] Alfonsi A. (2008). *High Order Discretization Schemes for the CIR Process: Application to Affine Term Structure and Heston Models*, available on <http://hal.archives-ouvertes.fr/>
- [2] Brigo D. & Mercurio, F. (2006). *Interest Rate Models: Theory and Practice*, 2nd Edition, Springer-Verlag.
- [3] Cox, J.C., Ingersoll, J.E. & Ross, S.A. (1985). An intertemporal general equilibrium model of asset prices, *Econometrica* **53**, 363–384.
- [4] Cox, J.C., Ingersoll, J.E. & Ross, S.A. (1985). A theory of the term structure of interest rates, *Econometrica* **53**, 385–407.
- [5] Duffie, D., Pan, J. & Singleton, S.A. (2000). Transform analysis and asset pricing for affine jump diffusions, *Econometrica* **68**, 1343–1376.
- [6] Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*, Series: Applications of Mathematics, Vol. 53, Springer.
- [7] Jamshidian, F. (1989). An exact bond option formula, *Journal of Finance* **44**, 205–209.
- [8] Jamshidian, F. (1995). A simple class of square-root interest-rate models, *Applied Mathematical Finance* **2**, 61–72.
- [9] Hull, J. & White, A. (1990). Pricing interest rate derivative securities, *The Review of Financial Studies* **3**, 573–592.
- [10] Lamberton, D. & Lapeyre, B. (1996). *An Introduction to Stochastic Calculus Applied to Finance*, Chapman & Hall.

Related Articles

Affine Models; Bond; Bond Options; Caps and Floors; Duffie–Singleton Model; Forward and Swap Measures; Hedging; Heston Model; LIBOR Market Model; Quadratic Gaussian Models; Risk Premia; Ross, Stephen; Simulation of Square-root Processes; Swap Market Models; Term Structure Models.

AURÉLIEN ALFONSI

Gaussian Interest-Rate Models

A major milestone in interest-rate modeling was the one-factor Gaussian mean-reverting model proposed by Vasicek [6] in 1977 for the short-term interest rate, using a time-homogeneous setup of the Ornstein–Uhlenbeck process (*see Term Structure Models*). Its analytical tractability and feasibility of effective numerical methods made the model popular among practitioners for many years. Multifactor Gaussian interest-rate models were subsequently developed as a natural extension of the Vasicek model and share these properties.

In 1990, Hull and White [3] generalized the Vasicek model to time-dependent parameters to permit fitting term structures of yields and volatilities. The instantaneous short rate $r(t)$ for the Hull–White one-factor (HW1F) model satisfies the evolution

$$dr(t) = (\vartheta(t) - a(t)r(t)) dt + \sigma(t) dW(t) \quad (1)$$

A volatility function $\sigma(t)$ and a mean reversion $a(t)$ are the main parameters of the model. The drift compensator $\vartheta(t)$ serves to fit the initial yield curve. Zero-bond and option pricing, as well as transition probabilities, are available analytically for the HW1F model, which makes it attractive not only for a calibration procedure but also for effective numerical implementations.

One of the main drawbacks of the HW1F model is the fact that yields of all maturities are functions of only one state-variable, $r(t)$, implying that all points on the yield curve are perfectly correlated. For financial contracts sensitive to the joint movements of multiple points on the yield curve, the HW1F model is therefore clearly inadequate. To decorrelate yields of different maturities, Hull and White [4] in 1995 proposed a two-dimensional generalization of the HW1F model, introducing a stochastic mean reversion $u(t)$ in the short rate,

$$dr(t) = (\vartheta(t) + u(t) - a r(t)) dt + \sigma dW(t) \quad (2)$$

$$du(t) = -a_u u(t) dt + \sigma_u dW_u(t) \quad (3)$$

Apart from previously defined short-rate parameters, the stochastic mean reversion $u(t)$ has its own time-independent volatility σ_u and mean reversion

a_u . Nontrivial correlation between the two Brownian motions, $\rho_u = E[dW(t) dW_u(t)]/dt$, and the two mean reversions, $a \neq a_u$, guarantee nontrivial correlations between yields of different maturities.

Duffie and Kan [2] generalized the model (2–3) to an arbitrary dimension as a special case of their affine model. Here, the short-rate process is presented as a sum of correlated Gaussian mean-reverting processes, $x_i(t)$, and an additional deterministic function, $\theta(t)$, used to match the original yield curve,

$$r(t) = \sum_{i=1}^N x_i(t) + \theta(t) \quad (4)$$

Each underlying process $x_i(t)$ obeys the Ornstein–Uhlenbeck equation,

$$dx_i(t) = -a_i(t)x_i(t) dt + \sigma_i(t) dW_i(t) \quad (5)$$

with correlated Brownian motions $E[dW_i(t) dW_j(t)] = C_{ij}(t) dt$. The model's stochastic differential equations (SDEs) are understood to be in the risk-neutral measure associated with the savings account numeraire $N(t) = e^{\int_0^t ds r(s)}$ and the expectation operator $E[\cdot \cdot \cdot]$.

The Hull–White two-factor model (2–3) and the $N = 2$ case of the symmetric form (4–5) are equivalent for $a \neq a_u$ provided that $a = a_1$, $a_u = a_2$, $\sigma^2 = \sigma_1^2 + \sigma_2^2 + 2\rho \sigma_1 \sigma_2$, $\sigma_u = (a_1 - a_2)\sigma_2$, $\rho_u = (\sigma_1 \rho + \sigma_2)/\sigma$, and $\vartheta(t) = a_1 \theta(t) + \theta'(t)$.

Analytical Properties

It is convenient to orthogonalize the Brownian motions in the model (5). For this, introduce vector-valued volatilities $\gamma_i(t)$ with elements $\gamma_{if}(t)$ for $f = 1, \dots, N$. Also consider vector-valued Brownian motion $dZ(t) = \{dZ_1, dZ_2, \dots, dZ_N\}$ with independent elements $E[dZ_f(t) dZ_{f'}(t)] = \delta_{ff'} dt$, where δ_{ij} is the Kronecker symbol. Now the underlying mean-reverting process (5) can be rewritten as

$$dx_i(t) = -a_i(t)x_i(t) dt + \gamma_i(t) \cdot dZ(t) \quad (6)$$

where the dot symbol denotes the dot product $\sum_{f=1}^N \gamma_{if}(t) dZ_f(t) = \gamma_i(t) \cdot dZ(t)$. To restore the initial dynamics (5), one should identify the scalar volatility σ_i with the module $|\gamma_i|$ and the correlation structure $C_{ij}(t) = E[dW_i(t) dW_j(t)]/dt$ with $\gamma_i(t) \cdot$

$\gamma_j(t)/|\gamma_i(t)| |\gamma_j(t)|$. In these notations, the initial Brownian motion can be expressed as $dW_i(t) = \gamma_i(t) \cdot dZ(t)/|\gamma_i(t)|$.

Key Formulas

This mean-reverting Ornstein–Uhlenbeck process (6) has the solution

$$x_i(\tau) = e^{-\int_t^\tau a_i(u) du} \left(x_i(t) + \int_t^\tau e^{\int_t^s a_i(u) du} \gamma_i(s) \cdot dZ(s) \right) \quad (7)$$

provided that its value at time t is fixed. Conditional moments of the x_i are easily shown to be

$$E[x_i(\tau) | x_i(t) = y_i] = e^{-\int_t^\tau a_i(u) du} y_i \quad (8)$$

$$\begin{aligned} \text{Cov}(x_i(\tau), x_j(\tau) | x_i(t) = y_i, x_j(t) = y_j) \\ = \int_t^\tau e^{-\int_t^s (a_i(u) + a_j(u)) du} \\ \times \gamma_i(s) \cdot \gamma_j(s) ds \end{aligned} \quad (9)$$

Let $P(t, T)$ be the time t price of a discount bond maturing at time T , and define instantaneous forward rates by $f(t, T) = -\partial \ln P(t, T)/\partial T$. In the model (4–6),

$$\begin{aligned} P(t, T) = E \left[e^{-\int_t^T r(\tau) d\tau} | F_t \right] = e^{-\int_t^T \theta(\tau) d\tau} \\ \times E \left[e^{-\sum_i \int_t^T x_i(\tau) d\tau} | F_t \right] \end{aligned} \quad (10)$$

which can be evaluated as the processes x_i are Gaussian. Indeed, it is easily established that

$$dP(t, T)/P(t, T) = r(t) dt - \sum_i \alpha_i(t, T) \gamma_i(t) \cdot dZ(t),$$

$$\alpha_i(t, T) \equiv \int_t^T ds e^{-\int_t^s a_i(u) du} \quad (11)$$

where we have used the fact that the drift of $dP(t, T)/P(t, T)$ must be $r(t)$ in the risk-neutral measure. Differentiating the SDE for $\ln P(t, T)$ over T , the forward rate dynamics emerge as

$$df(t, T) = \sum_i e^{-\int_t^T a_i(u) du} \gamma_i(t) \cdot dZ(t) + \mu(t, T) dt \quad (12)$$

where we have defined

$$\mu(t, T) = \sum_{i,j} e^{-\int_t^T a_i(u) du} \alpha_j(t, T) \gamma_i(t) \cdot \gamma_j(t) \quad (13)$$

Integrating the forward rate SDE leads to

$$\begin{aligned} f(t, T) = f(0, T) + \sum_i e^{-\int_t^T a_i(u) du} x_i(t) \\ + \int_0^t \mu(\tau, T) d\tau \end{aligned} \quad (14)$$

so forward rates are clearly Gaussian. For a general approach to the forward rates evolution, *see Heath–Jarrow–Morton Approach*. Integrating equation (14) yields the discount bond reconstitution formula

$$\begin{aligned} P(t, T) \\ = \frac{P(0, T)}{P(0, t)} e^{-\sum_i \alpha_i(t, T) x_i(t) - \frac{1}{2} \int_0^t (v(\tau, T) - v(\tau, t)) d\tau} \end{aligned} \quad (15)$$

where $v(t, T)$ is the instantaneous variance of $dP(t, T)/P(t, T)$, that is, $v(t, T) = \left| \sum_i \alpha_i(t, T) \gamma_i(t) \right|^2$. Note that equation (15) demonstrates that the entire discount curve at time t can be computed from the N Markov state-variables $x_i(t)$, $i = 1, \dots, N$. Also note that, from equation (14),

$$r(t) \equiv f(t, t) = f(0, t) + \sum_i x_i(t) + \int_0^t \mu(\tau, t) d\tau \quad (16)$$

which immediately establishes the drift $\theta(t)$ in formula (4) to be $\theta(t) = f(0, t) + \int_0^t \mu(\tau, t) d\tau$.

Finally, we present the covariance structure for the forward rates,

$$\text{Cov}(df(t, T), df(t, T')) = \frac{E[df(t, T) df(t, T')]}{dt}$$

$$= \sum_{i,j} e^{-\int_t^T a_i(u) du - \int_t^{T'} a_j(u) du} \gamma_i(t) \cdot \gamma_j(t) \quad (17)$$

from which one can obtain two important financial quantities: forward rate variance $V_F(t, T)$ and correlation between two forward rates $C_F(t, T, T')$, or

$$\begin{aligned} V_F(t, T) &= \frac{E[df(t, T) df(t, T)]}{dt} \\ &= \sum_{i,j} e^{-\int_t^T (a_i(u) + a_j(u)) du} \\ &\quad \times \gamma_i(t) \cdot \gamma_j(t) \end{aligned} \quad (18)$$

$$C_F(t, T, T') = \frac{\text{Cov}(df(t, T), df(t, T'))}{\sqrt{V_F(t, T) V_F(t, T')}} \quad (19)$$

European Option Prices

Consider a general European payer swaption with exercise date T_0 and payment dates $T_1, \dots, T_N$. For a fixed rate K , the swaption payoff is

$$\Pi(T_0) = \left(1 - P(T_0, T_N) - K \sum_{n=1}^N \delta_n P(T_0, T_n) \right)^+ \quad (20)$$

where δ_n is a day count fraction for a period starting at T_{n-1} and ending at T_n .

Defining discounted bonds as $R(t, T) = P(t, T)/P(t, T_0)$, the time zero value of the swaption can be written as

$$\begin{aligned} C &= P(0, T_0) E_0 \left[\left(\frac{\Pi(T_0)}{P(t, T_0)} \right)^+ \right] = P(0, T_0) E_0 \\ &\quad \times \left[\left(1 - R(T_0, T_N) - K \sum_{n=1}^N \delta_n R(T_0, T_n) \right)^+ \right] \end{aligned} \quad (21)$$

where $E_0[\dots]$ denotes expectation operator in the T_0 -forward measure, that is, the measure associated with a numeraire of $N_0(t) = P(t, T_0)/P(0, T_0)$ (see **Forward and Swap Measures**). As $R(t, T)$ must be a martingale in the T_0 -forward measure, equation (11) shows that

$$dR(t, T) = -R(t, T) \Sigma(t, T) \cdot dZ_0(t) \quad (22)$$

where $Z_0(t)$ is a Brownian motion in the T_0 -forward measure and $\Sigma(t, T)$ is a vector log-volatility

$$\Sigma(t, T) \equiv \sum_i e^{\int_0^t a_i(u) du} \gamma_i(t) \int_{T_0}^T ds e^{-\int_0^s a_i(u) du} \quad (23)$$

A solution for the discounted bonds can be easily written as

$$R(t, T_n) = \frac{P(0, T_n)}{P(0, T_0)} e^{-Y_n(t) - 1/2 V_n(t)} \quad (24)$$

where we have denoted Gaussian processes $Y_n(t) = \int_0^t \Sigma(\tau, T_n) \cdot dZ_0(\tau)$ and its variance, $V_n(t) = \int_0^t |\Sigma(\tau, T_n)|^2 d\tau$. Note that the processes Y_n can be presented as a linear combination of the driving processes x_i ,

$$\begin{aligned} Y_n(t) &= \sum_i \beta_i(T_n) x_i(t) \text{ for} \\ \beta_i(T_n) &= \int_{T_0}^{T_n} ds e^{-\int_0^s a_i(u) du} \end{aligned} \quad (25)$$

Substituting solution (24) into the price formula (21), we obtain

$$C = P(0, T_0) E_0 \left[\left(1 - \sum_{n=1}^N e^{A_n - Y_n(T_0)} \right)^+ \right] \quad (26)$$

where we have denoted $A_N = \ln((1 + K \delta_N) P(0, T_N)/P(0, T_0)) - 1/2 V_N(T_0)$ and $A_n = \ln(K \delta_n P(0, T_n)/P(0, T_0)) - 1/2 V_n(T_0)$ for $n = 1, \dots, N - 1$.

For the case where $N = 1$, that is when our swaption is really a caplet, equation (26) can be solved in closed form

$$\begin{aligned} C_1 &= P(0, T_0) \Phi \left(-\frac{A_1}{\sqrt{V_1(T_0)}} + \frac{\sqrt{V_1(T_0)}}{2} \right) \\ &\quad - (1 + K \delta_1) P(0, T_1) \\ &\quad \times \Phi \left(-\frac{A_1}{\sqrt{V_1(T_0)}} - \frac{\sqrt{V_1(T_0)}}{2} \right) \end{aligned} \quad (27)$$

where $\Phi(x)$ is the cumulative Gaussian distribution function and $A_1 = \ln((1 + K \delta_1) P(0, T_1)/P(0, T_0)) - 1/2 V_1(T_0)$.

For the case of a regular multiperiod swaption, we can rely on the trick in Jamshidian [5]. It is based on the observation that, for a continuous stochastic variable X taking values on the whole real axis, the following equation

$$\begin{aligned} E \left[\left(1 - \sum_{n=1}^N e^{A_n - B_n X} \right)^+ \right] \\ = E \left[\left(1 - \sum_{n=1}^N e^{A_n - B_n X} \right) 1_{X > x_0} \right] \end{aligned} \quad (28)$$

holds for barrier level x_0 satisfying $1 = \sum_{n=1}^N e^{A_n - B_n x_0}$, provided that coefficients $B_n \geq 0$ and at least one coefficient B is strictly positive. For Gaussian variable V with zero mean and variance v , we have

$$\begin{aligned} E \left[\left(1 - \sum_{n=1}^N e^{A_n - B_n X} \right)^+ \right] = \Phi \left(-\frac{x_0}{\sqrt{v}} \right) \\ - \sum_{n=1}^N e^{A_n + 1/2 B_n^2 v} \Phi \left(-\frac{x_0}{\sqrt{v}} - B_n \sqrt{v} \right) \end{aligned} \quad (29)$$

For the one-factor case, the Jamshidian trick applies directly to equation (26) and leads to a closed-form swaption pricing formula involving a simple root-search for the trigger level x_0 . For two-factor models, where $Y_n(T_0) = \beta_1(T_n) x_1(T_0) + \beta_2(T_n) x_2(T_0)$, defines Gaussian stochastic variables $X_1 = x_1(T_0)$ and $X_2 = x_2(T_0)$ with known covariance matrix by formula (9). To integrate the option price (26) over Gaussian X_1 and X_2 , one can compute analytically the conditional-to- X_1 average

$$E_0 \left[\left(1 - \sum_{n=1}^N e^{A_n - Y_n(T_0)} \right)^+ \middle| X_1 \right] \quad (30)$$

using the Jamshidian lemma, and then do the numerical integration over the Gaussian variable X_1 . The first step is indeed possible since the variable X_2 conditional to X_1 is normally distributed, and the exponents in the option formula (35) depend linearly on the variable X_2 as in expression (29). More details can be found in, for instance, [1].

If the exact option pricing algorithm is slow for concrete applications, or if the number of model factors is more than two, one can come up with a purely

analytical approximation^a in the so-called *swap measure* (see **Forward and Swap Measures**). Namely, introduce a swap level $L(t) = \sum_{n=1}^N \delta_n P(t, T_n)$ of and associate the swap measure with the numeraire $N_S(t) = L(t)/L(0)$ equipped with Brownian motion dZ_S and corresponding expectation operator $E_S[\cdot \cdot \cdot]$. Then, a swap rate,

$$S(t) = \frac{P(t, T_0) - P(t, T_N)}{\sum_{n=1}^N \delta_n P(t, T_n)} = \frac{P(t, T_0) - P(t, T_N)}{L(t)} \quad (31)$$

is a martingale in the swap measure. The option price written in the swap measure reduces to a simple expression, $C(T_0) = L(0) E_S[(S(T_0) - K)^+]$. The swap rate SDE can be easily written,

$$dS(t) = \sum_i \frac{\partial S(t)}{\partial x_i(t)} \gamma_i(t) \cdot dZ_S(t) \quad (32)$$

The partial swap derivatives $d_i(t, x(t)) \equiv \partial S(t) / \partial x_i(t)$ can be calculated using the zero-bond solution (15), $\partial P(t, T) / \partial x_i(t) = -\alpha_i(t, T) P(t, T)$.

Given that the forward rates are here Gaussian, it is natural to assume that the swap rates are as well, approximating derivatives with their value for the underlying rates at the origin, that is $d_i(t, x(t)) \simeq d_i(t, 0)$. This approximation resembles the freezing technique of low-variance processes used for swaption pricing in Libor market models by many authors (see **LIBOR Market Model**).

Define $v = \int_0^{T_0} \left| \sum_i d_i(t, 0) \gamma_i(t) \right|^2 dt$. Then the swaption price for the Gaussian approximation of the swap rate,

$$dS(t) \simeq \sum_i d_i(t, 0) \gamma_i(t) \cdot dZ_S(t) \quad (33)$$

can be easily written

$$\begin{aligned} C = L(0) E_S[(S(T_0) - K)^+] &\simeq L(0) \\ &\times \left((S(0) - K) \Phi \left(\frac{S(0) - K}{\sqrt{v}} \right) \right. \\ &\left. + \sqrt{v} \Phi' \left(\frac{S(0) - K}{\sqrt{v}} \right) \right) \end{aligned} \quad (34)$$

Numerical Methods

Given available transition probabilities for the model, one can apply lattice methods on the basis of the conditional expectation calculus *via* convolution with the Gaussian kernel for the pricing of general payouts. Of course, other lattice techniques—such as finite differences—can be successfully used as well.

When the payout is path-dependent or the dimension of the model is beyond, say, 3 or 4, lattice methods must be replaced by Monte Carlo simulation; path simulation in the Monte Carlo method is straightforward and involves making Gaussian draws with moments computed from the conditional expectations (8–9).

Properties

We cover in detail the two-factor case, $r(t) = x_1(t) + x_2(t) + \theta(t)$ frequently used in the financial industry. In practical applications, the model correlation between the two Brownian motions $\rho(t) = E[dW_1(t)dW_2(t)]/dt$ in the notation (5), or the cosine of the angle between the two volatility vectors $\rho(t) = \gamma_1(t) \cdot \gamma_2(t)/|\gamma_1(t)||\gamma_2(t)|$ in the notations (6), typically takes highly negative values, $\rho(t) \sim -0.9$. The two mean reversions are often radically different $a_1(t) \sim 0.5$ and $a_2(t) \sim 0.05$ with the volatilities $\sigma_1(t) = |\gamma_1(t)| \sim 0.005$ and $\sigma_2(t) = |\gamma_2(t)| \sim 0.01$

Volatility Hump and Correlations

For illustration, we consider in more detail the two-factor model with time-independent parameters. The forward rates variance (18) simplifies to

$$V_F(t, T) = e^{-2a_1(T-t)}\sigma_1^2 + 2e^{-(a_1+a_2)(T-t)} \times \sigma_1\sigma_2\rho + e^{-2a_2(T-t)}\sigma_2^2 \quad (35)$$

For positive correlation ρ , the variance is a monotonous function of $T - t$, although, for negative ρ , it can give the volatility hump observed on the market, see [1] for details.

In our two-dimensional model, the forward rate $f(t, T)$ has instantaneous volatility,

$$\sigma(t, T) = e^{-(T-t)a_1}\gamma_1 + e^{-(T-t)a_2}\gamma_2 \quad (36)$$

obtained from the general formula (12). The correlation between two forward rates $f(t, T)$ and $f(t, T')$ can be computed from equation (19) as

$$C_F(t, T, T') = \frac{\sigma(t, T) \cdot \sigma(t, T')}{|\sigma(t, T)||\sigma(t, T')|} \quad (37)$$

We notice that the correlation is one when $a_1 = a_2$, a result of the fact that the volatilities for $f(t, T)$ and $f(t, T')$ are colinear in this case.

Volatility Smile

Swap rates in the Gaussian model are, as we have seen, nearly Gaussian, irrespective of parameter choice. As such, there is essentially no way to control the volatility skew implied by the model, which is often more steeply downward-sloping than market-observed smiles. Consequently, some care must be taken when applying the model to smile-sensitive instruments (*see Markovian Term Structure Models*).

Calibration

Practitioners typically use two- and three-factor Gaussian models. For four factors and more, pricing of an exotic instrument may become too time consuming.

A standard approach to the two-factor model calibration includes the following steps. First, we fix the time-independent correlation to a highly negative value or calibrate it to average historical correlations between forward rates by inversion of the correlation formula (19). Second, we calibrate the time-dependent volatilities and time-independent mean reversion^b to European options. The calibration options are often taken at-the-money, which reflects the absence of control of the skew and smile. The time-dependent parameters are typically considered as step-wise constant between option exercise dates. One can also use specially parameterized volatility curves, for example, having a hump form. Another popular calibration technique maintains a fixed ratio between the time-dependent volatilities, $\sigma_1(t)/\sigma_2(t) = \text{const}$.

The calibration is typically done using a numerical global optimizer to fit the model option prices to the market prices. The model option prices are calculated analytically, see the section European Option Prices.

For fixed mean reversions, one can also use a bootstrap in option exercise dates for the step wise constant volatilities calibration.

Acknowledgments

The author is indebted to Leif Andersen, Vladimir Piterbarg, and Jesper Andreasen for numerous discussions he had with them and their help with references. He is also grateful to Leo Mizrahi, Maria Belyanina, and his NumeriX colleagues, especially, to Greg Whitten, Serguei Issakov, Nicolas Audet, Meng Lu, Serguei Mechkov, and Patti Harris, for their valuable comments on the article.

End Notes

^aThanks to Leif Andersen and Vladimir Piterbarg for suggesting this approach.

^bTime-dependence in mean reversion is often avoided as it implies nonstationary behavior of the shape of the volatility term structure.

References

- [1] Brigo, D. & Mercurio, F. (2001). *Interest Rate Models, Theory and Practice*. Springer Finance.
- [2] Duffie, D. & Kan, R. (1996). A yield-factor model of interest rates, *Mathematical Finance* **6**(4), 379–406.
- [3] Hull, J. & White, A. (1990). Pricing interest-rate derivative securities, *The Review of Financial Studies* **3**(4), 573–592.
- [4] Hull, J. & White, A. (1994). Numerical procedures for implementing term structure models II: two-factor models, *Journal of Derivatives* **2**, 37–47.
- [5] Jamshidian, F. (1989). An exact bond option formula, *Journal of Finance* **44**, 205–209.
- [6] Vasicek, O. (1977). An equilibrium characterization of the term structure, *Journal of Financial Economics* **5**(2), 177–188.

Related Articles

Bermudan Swaptions and Callable Libor Exotics; Forward and Swap Measures; Heath–Jarrow–Morton Approach; LIBOR Rate; LIBOR Market Model; Markovian Term Structure Models; Term Structure Models.

ALEXANDRE V. ANTONOV

Quadratic Gaussian Models^a

Quadratic Gaussian (QG) models are factor models for the pricing of interest-rate derivatives, where interest rates are quadratic functions of underlying Gaussian factors. The QG model of interest rates was first introduced by Beaglehole and Tenney [3] and by El Karoui *et al.* [11]. Similar models had been introduced in epidemiology [16]. Jamshidian [14], under restrictive hypothesis on the dynamic of the factors, obtained closed formulas for the prices of vanilla options in the QG model. Durand and El Karoui [10] have detailed some properties and statistical analysis of the QG model. One may refer, for example, to [1, 6] for general studies of quadratic term-structure models.

Quadratic Gaussian Model

Uncertainty is represented by an n -dimensional Brownian motion $\hat{W}_t = (\hat{W}_t^1, \hat{W}_t^2, \dots, \hat{W}_t^n)^*$ on the filtered probability space (Ω, F_t, P) , where F_t is the natural augmented filtration of Brownian motion $\hat{W}$ and P the historical probability. We then add two assumptions:

Hypothesis 1 *The asset price processes in all the different markets are regular deterministic functions of n state variables.*

Hypothesis 2 *The state variables have a Gaussian–Markovian distribution with respect to all the risk-neutral probabilities (which includes the forward-neutral probabilities).*

The first hypothesis is common and is used in numerical methods such as finite difference methods to price and hedge options; the second is also natural if we want to have results as explicit as possible. We specify, more precisely, the diffusion of the state variable Z under the risk-neutral probability (*see Ornstein–Uhlenbeck Processes*):

$$dZ_t = (A_t Z_t + \mu_t) dt + \Sigma_t dW_t \quad (1)$$

where (W_t) is a Brownian motion under the domestic risk-neutral probability $\mathcal{Q}$.

Theorem 1 [11]. *Under these assumptions, there exist a family of symmetrical matrices $\hat{\Gamma}(t, T)$, a family of vectors $\hat{b}(t, T)$, and a family of scalars $\hat{a}(t, T)$ such that the zero-coupon price at t of the bond $P(t, T)$ is given by*

$$P(t, T) = \exp [Z_t^* \hat{\Gamma}(t, T) Z_t + \hat{b}^*(t, T) Z_t + \hat{a}(t, T)] \quad (2)$$

In particular, the functional dependence of the short rate r_t on the factors is quadratic affine:

$$\begin{aligned} r_t &= -\partial_T \ln P(t, T)_{t=T} \\ &= [Z_t^* \Gamma(t, T) Z_t + b^*(t, T) Z_t + a(t, T)] \end{aligned} \quad (3)$$

where

$$\begin{cases} \Gamma(t, T) = -\partial_T \hat{\Gamma}(t, T) \\ b(t, T) = -\partial_T \hat{b}(t, T) \\ a(t, T) = -\partial_T \hat{a}(t, T) \end{cases} \quad (4)$$

To describe completely the dynamics of the interest rates, we need to make explicit the computation of the matrices $\hat{\Gamma}(t, T)$, $\hat{b}(t, T)$, and $\hat{a}(t, T)$. Once again, the no-arbitrage assumption will give us the solution.

Theorem 2 [11]. *Given the short rate parameters, $\Gamma(t, t)$, $b(t, t)$, and $a(t, t)$, the matrices $\hat{\Gamma}(t, T)$, $\hat{b}(t, T)$, and $\hat{a}(t, T)$ are the solutions of the backward differential system with respect to the current date t*

$$\begin{cases} \partial_t U_t = -(A_t^* U_t + U_t A_t) + 2U_t^* \Sigma_t \Sigma_t^* U_t - \Gamma(t, t) \\ \partial_t u_t = -A_t^* u_t + 2U_t \Sigma_t^* \Sigma_t u_t - 2U_t \mu_t - b(t, t) \\ \partial_t \alpha_t = -tr[\Sigma_t \Sigma_t^* U_t] - u_t^* \mu_t + \frac{1}{2} u_t^* \Sigma_t \Sigma_t^* u_t - a(t, t) \end{cases} \quad (5)$$

with the initial conditions

$$\begin{cases} U_T = 0 \\ u_T = 0 \\ \alpha_T = 0 \end{cases} \quad (6)$$

These two results, which are direct consequences of the hypothesis, completely describe the diffusion of the interest rates under the risk-neutral probability. In the stationary case, let us denote $\hat{\Gamma}_\infty$ the solution of the algebraic Riccati equation on U :

$$A^* U + U A - 2U^* \Sigma \Sigma^* U + \Gamma = 0 \quad (7)$$

If $A - 2\hat{\Gamma}_\infty \Sigma \Sigma^*$ has only negative eigenvalues, then the equations of Theorem 2 will only have nonexploding solutions. In those cases, one can even obtain

a simple expression for the limit of the zero-coupon interest rate when the maturity increases to infinity; this limit does not depend on the time (see [10]). The QG model admits, for example, the following models as particular cases:

- the Gaussian model of Vasicek [15], its generalization by Hull and White [13], and its multidimensional version by Heath, Jarrow, and Morton (HJM) [12] (see **Gaussian Interest-Rate Models**);
- Cox, Ingersoll, and Ross (CIR) [8], Chen and Scott [7], and Duffie and Kan [9]; also see **Cox–Ingersoll–Ross (CIR) Model**.

Statistical Justification

In addition to the practical and computational justification of the hypothesis that lead to the QG model, this model can also be justified by statistical studies of the interest rates. The most intuitive justification comes from principal component analysis (PCA). For most currencies, using a PCA will lead to keeping two factors to explain the diffusion of the interest yield curves. Then, the residual noise may be explained by a quadratic form of the first two factors of the PCA. For a detailed description of this analysis, we refer to [10].

Option Pricing

In the QG model, the instantaneous volatility of the zero-coupon interest rates $R(t, T)$ is an affine form of the factors Z_t , and when we consider two factors or more, the instantaneous volatility of $R(t, T)$ cannot be written as a deterministic form of $R(t, T)$: in some way, we can therefore consider the QG model as a stochastic volatility model.^c Its main interest will be in interest-rate exotic option pricing and hedging.

Closed Form for Vanilla Option

Quasi-closed form for vanillas options are the consequence of the following result.

Theorem 3 [11]. *Let us denote by $m^T(t, u)$ and $V^T(t, u)$ the conditional mean and the conditional variance of Z_u under the forward-neutral probability of maturity T $\mathbb{Q}_t^T$. The conditional mean $m^T(t, T)$ and*

variance $V^T(t, T)$ of the factors are solutions of the differential system given by equation (8):

$$\begin{cases} \frac{\partial V^T(t, T)}{\partial T} = V^T(t, T)(A_T)^* + A_T V^T(t, T) \\ \quad - 2V^T(t, T)\Gamma(T, T)V^T(t, T) + \Sigma_T \Sigma_T^* \\ \frac{\partial m^T(t, T)}{\partial T} = A_T m^T(t, T) \\ \quad - 2V^T(t, T)\Gamma(T, T)m^T(t, T) \\ \quad + 2V^T(t, T)b(T, T) + \mu_T \\ V^T(t, t) = 0, \quad m^T(t, t) = Z_t \end{cases} \quad (8)$$

and under the risk-neutral probability $\mathbb{Q}_t$, the conditional mean $m(t, T)$ and variance $V(t, T)$ of the factors are solutions of the differential system:

$$\begin{cases} \frac{\partial V(t, T)}{\partial T} = V(t, T)(A_T)^* + A_T V(t, T) + \Sigma_T \Sigma_T^* \\ \frac{\partial m(t, T)}{\partial T} = A_T m(t, T) + \mu_T \\ V(t, t) = 0 \quad m(t, t) = Z_t \end{cases} \quad (9)$$

Assefa [2] has detailed swaption prices in the QG model and has described how approximations may be obtained using Fourier transform. Some analytical approximations of caps, floors, and swaptions are also proposed in [4].

Pricing of Exotic Options

To use the QG model to price and hedge exotic options, one needs to add some constraints to simplify the different equations and to accelerate the different numerical schemes used in the different steps of calibrating the model on plain vanilla option prices and used to price exotic options. Some calibration results of a two-factor QG model on USD Libor caps and swaptions are shown in [2]. Depending on the shape of the smile and on the payoff of the exotic option studied, one uses a one- or two-factor model: empirically, by using a one-factor QG model one will generate increasing or decreasing volatility smiles; with a two-factor QG model, “U-shape” volatility smiles may be generated. The QG model can then be successfully calibrated on vanilla option prices and on interest-rate correlation to price and hedge exotic interest-rate options. We recommend to choose a calibration set for each exotic option and not to try to calibrate simultaneously to the entire volatility cube. We also recommend to specify a correlation structure

on the forward interest rate and not to deduce it implicitly from the market implied volatilities.

Summary

Owing to its capacity to calibrate interest-rate volatility smiles and correlation term structure, the QG is an interesting model to price and hedge exotic interest-rate derivatives, as multicancelable options on interest-rate spreads. Several banks have developed proprietary models based on the QG model, each choosing its own parameterization, mainly as a consequences of choices made on the numerical schemes used.

End Notes

^aThis article reflects the author's point of view and do not necessarily represents the point of view of Banque de France.

^bHere x^* denotes transpose (x).

^cIn fact, as we can deduce the factors Z_t from the interest-rate curve, we can write the instantaneous volatility of $R(t, \theta)$ as a deterministic function of interest rate of different maturities; in a theoretical point of view, the QG model may be not a stochastic volatility model as defined in [5].

References

- [1] Ahn, D.-H., Dittmar, R.F. & Gallant, A.R. (1999). *Quadratic Term Structure Models: Theory and Evidence*, working paper, University of North Carolina.
- [2] Assefa, S. (2007). *Calibrating and Pricing in a Multi-Factor Quadratic Gaussian Model*, research paper, Quantitative Finance Research Centre, University of Technology Sydney.
- [3] Beaglehole, D. & Tenney, M. (1991). General solutions of some interest rate contingent claim pricing equations, *Journal of Fixed Income* **1**, 69–83.
- [4] Boyarchenko, N. & Levendorski, S. (2007). The eigenfunction expansion method in multi-factor quadratic term structure models, *Mathematical Finance* **17**, 503–539.
- [5] Casassus, J., Collin-Dufresne, P. & Goldstein, B. (2005). Unspanned stochastic volatility and fixed income derivatives pricing, *Journal of Banking and Finance* **29**, 2723–2749.
- [6] Chen, L., Filipovic, D. & Poor, H.V. (2004). Quadratic term structure models for risk-free and defaultable rates, *Mathematical Finance* **14**, 515–536.
- [7] Chen, R.-R. & Scott, L. (1992). Pricing interest rate options in a two-factor cox-ingersoll-ross model of the term structure, *Review of Financial Studies* **5**, 613–636.
- [8] Cox, J., Ingersoll, J. & Ross, S. (1985). A theory of the term structure of interest rates, *Econometrica* **53**, 385–407.
- [9] Duffie, D. & Kan, R. (1992). *A Yield-Factor Model of Interest Rates*, working paper, Stanford University.
- [10] Durand, Ph. & El Karoui, N. (1998). *Interest Rates Dynamics and Option Pricing with the Quadratic Gaussian Model in Several Economies*, working paper.
- [11] El Karoui, N., Myneni, R. & Viswanathan, R. (1991). *Arbitrage Pricing and Hedging of Interest Rate Claims with State Variables, Theory and Applications*, working paper, Universite Paris VI.
- [12] Heath, D., Jarrow, R. & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation, *Econometrica* **60**, 77–105.
- [13] Hull, J. & White, A. (1990). Pricing interest rate derivative securities, *Review of Financial Studies* **3**, 573–592.
- [14] Jamshidian, F. (1996). Bond, futures and option evaluation in the quadratic interest rate model, *Applied Mathematical Finance* **3**, 93–115.
- [15] Vasicek, O. (1977). An equilibrium characterisation or the term structure, *Journal of Financial Economics* **5**, 177–188.
- [16] Woodbury, M.A., Manton, K.G. & Stallard, E. (1979). Longitudinal analysis of the dynamics and risk of coronary heart disease in the framingham study, *Biometrics* **35**, 575–585.

PHILIPPE DURAND

Affine Models

Definition

Notation 1 Throughout the article, $\langle \cdot, \cdot \rangle$ denotes the standard scalar product on $\mathbb{R}^N$.

Definition 1 Let r_t be a short-rate model specified as an affine function of an N -dimensional Markov process X_t with state space $D \subseteq \mathbb{R}^N$:

$$r_t = l + \langle \lambda, X_t \rangle \quad (1)$$

for some (nontime-dependent) constants $l \in \mathbb{R}$ and $\lambda \in \mathbb{R}^N$. This is called an affine term structure model (ATSM) if the zero-coupon bond price has exponential-affine form, that is,

$$\begin{aligned} P(t, T) &= \mathbb{E} \left[e^{-\int_t^T r_s ds} \middle| X_t \right] \\ &= e^{G(t, T) + \langle H(t, T), X_t \rangle} \end{aligned} \quad (2)$$

where $\mathbb{E}$ denotes the expectation under a risk neutral probability measure.

Early Examples

Early well-known examples are the Vasiček [14] and the Cox *et al.* [5] (see **Term Structure Models; Cox–Ingersoll–Ross (CIR) Model**) time-homogeneous one-factor short-rate models. In equation (1), both models are characterized by $N = 1$, $l = 0$, and $\lambda = 1$.

Vasiček Model

X_t follows an Ornstein–Uhlenbeck process on $D = \mathbb{R}$,

$$dX_t = (b + \beta X_t) dt + \sigma dW_t, \quad b, \beta \in \mathbb{R}, \quad \sigma \in \mathbb{R}_+ \quad (3)$$

where W_t is a standard Brownian motion. Under these model specifications, bond prices can be explicitly calculated and the corresponding coefficients G and H in equation (2) are given by

$$H(t, T) = \frac{1 - e^{\beta(T-t)}}{\beta}$$

$$\begin{aligned} G(t, T) &= \frac{\sigma^2}{2} \int_t^T H^2(s, T) ds \\ &\quad + b \int_t^T H(s, T) ds \end{aligned} \quad (4)$$

provided that $\beta \neq 0$ (see also **Term Structure Models**).

Cox–Ingersoll–Ross Model

X_t is defined as the solution of the following affine diffusion process on $D = \mathbb{R}_+$, known as *Feller square root process*,

$$\begin{aligned} dX_t &= (b + \beta X_t) dt + \sigma \sqrt{X_t} dW_t \\ b, \sigma &\in \mathbb{R}_+, \quad \beta \in \mathbb{R} \end{aligned} \quad (5)$$

Like in the Vasiček model, there is a closed-form solution for the bond price. If $\sigma \neq 0$, G and H in equation (2) are then of the form:

$$\begin{aligned} G(t, T) &= \frac{2b}{\sigma^2} \ln \left(\frac{2\gamma e^{(\gamma - \beta)(T-t)/2}}{(\gamma - \beta)(e^{\gamma(T-t)} - 1) + 2\gamma} \right) \\ H(t, T) &= \frac{-2(e^{\gamma(T-t)} - 1)}{(\gamma - \beta)(e^{\gamma(T-t)} - 1) + 2\gamma} \end{aligned} \quad (6)$$

where $\gamma := \sqrt{\beta^2 + 2\sigma^2}$ (see also **Cox–Ingersoll–Ross (CIR) Model**).

Since the development of these first one-dimensional term structure models, many multifactor extensions have been considered with the aim to provide more realistic models.

Regular Affine Processes

The generic method to construct ATSMs is to use regular affine processes. A concise mathematical foundation was provided by Duffie *et al.* [8]. Henceforth, we fix the state space $D = \mathbb{R}_+^m \times \mathbb{R}^{N-m}$, for some $0 \leq m \leq N$.

Definition 2 A Markov process X is called *regular affine* if its characteristic function has exponential-affine dependence on the initial state, that is, for

$t \in \mathbb{R}_+$ and $u \in i\mathbb{R}^N$, there exist $\phi(t, u) \in \mathbb{C}$ and $\psi(t, u) \in \mathbb{C}^N$, such that for all $x \in D$

$$\mathbb{E}[e^{\langle u, X_t \rangle} | X_0 = x] = e^{\phi(t, u) + \langle \psi(t, u), x \rangle} \quad (7)$$

Moreover, the functions ϕ and ψ are continuous in t and $\partial_t^+ \phi(t, u)|_{t=0}$ and $\partial_t^+ \psi(t, u)|_{t=0}$ exist and are continuous at $u = 0$.

Regular affine processes have been defined and completely characterized in [8]. The main result is stated below.

Theorem 1 A regular affine process is a Feller semimartingale with infinitesimal generator

$$\begin{aligned} \mathcal{A}f(x) = & \sum_{k,l=1}^N A_{kl}(x) \frac{\partial^2 f(x)}{\partial x_k \partial x_l} \\ & + \langle B(x), \nabla f(x) \rangle - C(x)f(x) \\ & + \int_{D \setminus \{0\}} (f(x + \xi) - f(x) \\ & - \langle \nabla f(x), \chi(\xi) \rangle) M(x, d\xi) \end{aligned} \quad (8)$$

for f in the set of smooth test functions, with

$$A(x) = a + \sum_{i=1}^m x_i \alpha_i, \quad a, \alpha_i \in \mathbb{R}^{N \times N} \quad (9)$$

$$B(x) = b + \sum_{i=1}^N x_i \beta_i, \quad b, \beta_i \in \mathbb{R}^N \quad (10)$$

$$C(x) = c + \sum_{i=1}^m x_i \gamma_i, \quad c, \gamma_i \in \mathbb{R}_+ \quad (11)$$

$$M(x, d\xi) = m(d\xi) + \sum_{i=1}^m x_i \mu_i(d\xi) \quad (12)$$

where m, μ_i are Borel measures on $D \setminus \{0\}$ and $\chi : \mathbb{R}^N \rightarrow \mathbb{R}^N$ some bounded continuous truncation function with $\chi(\xi) = \xi$ in a neighborhood of 0. Furthermore, ϕ and ψ in equation (7) solve the generalized Riccati equations,

$$\partial_t \phi(t, u) = F(\psi(t, u)), \quad \phi(0, u) = 0 \quad (13)$$

$$\partial_t \psi(t, u) = R(\psi(t, u)), \quad \psi(0, u) = u \quad (14)$$

with

$$F(u) = \langle au, u \rangle + \langle b, u \rangle - c$$

$$+ \int_{D \setminus \{0\}} (e^{\langle u, \xi \rangle} - 1 - \langle u, \chi(\xi) \rangle) m(d\xi) \quad (15)$$

$$\begin{aligned} R_i(u) = & \langle \alpha_i u, u \rangle + \langle \beta_i, u \rangle - \gamma_i \\ & + \int_{D \setminus \{0\}} (e^{\langle u, \xi \rangle} - 1 - \langle u, \chi(\xi) \rangle) \mu_i(d\xi) \end{aligned} \quad (16)$$

for $i \in \{1, \dots, m\}$

$$R_i(u) = \langle \beta_i, u \rangle, \quad \text{for } i \in \{m+1, \dots, N\} \quad (17)$$

Conversely, for any choice of admissible parameters $a, \alpha_i, b, \beta_i, c, \gamma_i, m, \mu_i$, there exists a unique regular affine process with generator (8).

Remark 1 It is worth noting that the infinitesimal generator of every Feller process on $\mathbb{R}^N$ has the form of the above integro-differential operator (8) with some functions A, B, C and a kernel M . The specific characteristic of regular affine processes is that these functions are all affine, as described in equations (9–12).

Observe furthermore that by the definition of the infinitesimal generator and the form of F and R , we have

$$\begin{aligned} & \frac{d}{dt} \mathbb{E}[e^{\langle u, X_t \rangle} | X_0 = x] \Big|_{t=0_+} \\ & = (\partial_t^+ \phi(t, u)|_{t=0} + \partial_t^+ \psi(t, u)|_{t=0}) e^{\langle u, x \rangle} \\ & = (F(u) + \langle R(u), x \rangle) e^{\langle u, x \rangle} = \mathcal{A}e^{\langle u, x \rangle} \end{aligned} \quad (18)$$

This gives the link between the form of the operator $\mathcal{A}$ and the functions F and R in the Riccati equations (13) and (14).

Remark 2 The above parameters satisfy certain admissibility conditions guaranteeing the existence of the process in D . These parameter restrictions can be found in Definition 2.6 and equations (2.23)–(2.24) in [8]. We note that admissibility, in particular, means $\alpha_{i,kl} = 0$ for $i, k, l \leq m$ unless $k = l = i$.

Systematic Analysis

Regular Affine Processes and ATSMs

Regular affine processes generically induce ATSMs. This relation is explicitly stated in the subsequent

argument. Under some technical conditions that are specified in [8, Chapter 11], we have for r_t as defined in equation (1),

$$\mathbb{E}\left[e^{-\int_0^t r_s ds} e^{\langle u, X_t \rangle} \middle| X_0 = x\right] = e^{\tilde{\phi}(t, u) + \langle \tilde{\psi}(t, u), x \rangle} \quad (19)$$

where

$$\begin{aligned} \partial_t \tilde{\phi}(t, u) &= \tilde{F}(\tilde{\psi}(t, u)) \\ \partial_t \tilde{\psi}(t, u) &= \tilde{R}(\tilde{\psi}(t, u)) \end{aligned} \quad (20)$$

with $\tilde{F}(u) = F(u) - l$ and $\tilde{R}(u) = R(u) - \lambda$. Setting $u = 0$ in equation (19), one immediately gets equation (2) with $G(t, T) = \tilde{\phi}(T - t, 0)$ and $H(t, T) = \tilde{\psi}(T - t, 0)$.

Diffusion Case

Conversely, for a class of diffusions

$$dX_t = B(X_t) dt + \sigma(X_t) dW_t \quad (21)$$

on D , Duffie and Kan [7] analyzed when equation (2) implies an affine diffusion matrix $A = \frac{\sigma \sigma^T}{2}$ and an affine drift B of form equations (9) and (10), respectively.

One-dimensional Nonnegative Markov Process

For $D = \mathbb{R}_+$, Filipović [9] showed that equation (1) defines an ATSM if and only if X_t is a regular affine process.

Relation to Heath–Jarrow–Morton Framework

Filipović and Teichmann [10] established a relation between the Heath–Jarrow–Morton (HJM) framework (see **Heath–Jarrow–Morton Approach**) and ATSMs: essentially, all generic finite dimensional realizations^a of a HJM term structure model are time-inhomogeneous ATSMs.

Canonical Representation

An ATSM stemming from a regular affine diffusion process X on $\mathbb{R}_+^m \times \mathbb{R}^{N-m}$ can be represented in different ways by applying nonsingular affine transformations to X . Indeed, for every nonsingular $N \times N$ -matrix K and $\kappa \in \mathbb{R}^N$, the transformation

$KX + \kappa$ modifies the particular form of equation (21) and the short-rate process (1), while observable quantities (e.g., the term structure or bond prices) remain unchanged. To group those N -dimensional ATSMs generating identical term structures, Dai and Singleton [6] found $N + 1$ subfamilies $\mathbb{A}_m(N)$, where $0 \leq m \leq N$ is the number of state variables actually appearing in the diffusion matrix (i.e., the dimension of the positive half space). For each class, they specified a canonical representation whose diffusion matrix $\sigma \sigma^T$ is of diagonal form with

$$(\sigma \sigma^T(x))_{kk} = \begin{cases} x_k, & k \leq m \\ 1 + \sum_{i=1}^m \lambda_{k,i} x_i, & k > m \end{cases} \quad (22)$$

where $\lambda_{k,i} \in \mathbb{R}$. For $N \leq 3$ the Dai–Singleton specification comprises all ATSMs generated by regular affine diffusions on $\mathbb{R}_+^m \times \mathbb{R}^{N-m}$. The general situation $N > 3$ was analyzed by Cheridito *et al.* [4].

Empirical Aspects

Pricing

The price of a claim with payoff function $f(X_t)$ is given by the risk neutral expectation formula:

$$\pi(t, x) = \mathbb{E}\left[e^{-\int_0^t r_s ds} f(X_t) \middle| X_0 = x\right] \quad (23)$$

Suppose that f can be expressed by

$$f(x) = \int_{\mathbb{R}^N} e^{\langle C+i\lambda, x \rangle} \tilde{f}(\lambda) d\lambda, \quad \lambda \in \mathbb{R}^N \quad (24)$$

for some integrable function $\tilde{f}$ and some constant $C \in \mathbb{R}^N$. If, moreover,

$$\mathbb{E}\left[e^{-\int_0^t r_s ds} e^{\langle C, X_t \rangle} \middle| X_0 = x\right] < \infty \quad (25)$$

then equation (19) implies

$$\begin{aligned} \pi(t, x) &= \mathbb{E}\left[e^{-\int_0^t r_s ds} \left(\int_{\mathbb{R}^N} e^{\langle C+i\lambda, X_t \rangle} \tilde{f}(\lambda) d\lambda\right) \middle| X_0 = x\right] \\ &= \int_{\mathbb{R}^N} \mathbb{E}\left[e^{-\int_0^t r_s ds} e^{\langle C+i\lambda, X_t \rangle} \middle| X_0 = x\right] \tilde{f}(\lambda) d\lambda \\ &= \int_{\mathbb{R}^N} e^{\tilde{\phi}(t, C+i\lambda) + \langle \tilde{\psi}(t, C+i\lambda), x \rangle} \tilde{f}(\lambda) d\lambda \end{aligned} \quad (26)$$

Hence, the price $\pi(t, x)$ can be computed *via* numerical integration, since the integrands are, in principle, known. For instance, in the case $N = 1$, the payoff function of a European call $(e^x - e^k)^+$, where x corresponds to the log price of the underlying and k to the log strike price, satisfies equation (24). In particular, we have the following integral representation (see [11]):

$$(e^x - e^k)^+ = \frac{1}{2\pi} \int_{\mathbb{R}} e^{(C+i\lambda)x} \frac{e^{k(1-C-i\lambda)}}{(C+i\lambda)(C+i\lambda-1)} d\lambda \quad (27)$$

Therefore, the previous formula to compute the price of the call $\pi(t, x)$ is applicable. An alternative approach leading to the same result can be found in [3].

Estimation

Statistical methods to estimate the parameters of ATSMs have been based on maximum likelihood and generalized method of moments.

Concerning maximum likelihood techniques, the conditional log densities entering into the log likelihood function can, in general, be obtained by inverse Fourier transformation. Since this procedure is computationally costly, several approximations and limited-information estimation strategies have been considered (e.g., [13]). Another possibility is to use closed-form expansions of the log likelihood function, which are available for general diffusions [1] and which have been applied to ATSMs. In the case of Gaussian and Cox–Ingersoll–Ross models, one can forgo such techniques, since the log densities are known in closed form (e.g., [12]).

As conditional moments of the form $\mathbb{E}[X_t^m X_{t-s}^n]$ for $m, n \geq 0$ can be computed from the derivatives of the conditional characteristic function and are, in general, explicitly known up to the solution of the Riccati ordinary differential equations (ODEs) (13) and (14), the generalized method of moments is an alternative to maximum likelihood estimation (e.g., [2]).

Acknowledgments

Authors Christa Cuchiero and Josef Teichmann gratefully acknowledge the support from the FWF-grant Y 328

(START prize from the Austrian Science Fund). Damir Filipovic gratefully acknowledges the support from WWTF (Vienna Science and Technology Fund).

End Notes

^aFor a precise definition, see [10].

References

- [1] Ait-Sahalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusions, *Annals of Statistics* **36**, 906–937.
- [2] Andersen, T.G. & Sørensen, B.E. (1996). GMM estimation of a stochastic volatility model: A Monte Carlo study, *Journal of Business & Economic Statistics* **14**, 328–352.
- [3] Carr, P. & Madan, D. (1998). Option valuation using the fast Fourier transform, *Journal of Computational Finance* **2**, 61–73.
- [4] Cheridito, P., Filipović, D. & Kimmel, R.L. *A note on the Dai-Singleton canonical representation of affine term structure models*. Forthcoming in *Mathematical Finance*.
- [5] Cox, J.C., Ingersoll, J.E. & Ross, S.A. (1985). A theory of the term structure of interest rates, *Econometrica* **53**, 385–407.
- [6] Dai, Q. & Singleton, K.J. (2000). Specification analysis of affine term structure models, *Journal of Finance* **55**, 1943–1978.
- [7] Duffie, D. & Kan, R. (1996). A yield-factor model of interest rates, *Mathematical Finance* **6**, 379–406.
- [8] Duffie, D., Filipović, D. & Schachermayer, W. (2003). Affine processes and applications in finance, *The Annals of Applied Probability* **13**, 984–1053.
- [9] Filipović, D. (2001). A general characterization of one factor affine term structure models, *Finance and Stochastics* **5**, 389–412.
- [10] Filipović, D. & Teichmann, J. (2004). On the geometry of the term structure of interest rates, *Proceedings of The Royal Society of London. Series A. Mathematical, Physical and Engineering Sciences* **460**, 129–167.
- [11] Hubalek, F., Kallsen, J. & Krawczyk, L. (2006). Variance-optimal hedging for processes with stationary independent increments, *Annals of Applied Probability* **16**, 853–885.
- [12] Pearson, N.D. & Sun, T.-S. (1994). Exploiting the conditional density in estimating the term structure: An application to the Cox, Ingersoll, and Ross model, *Journal of Finance* **49**, 1279–1304.
- [13] Singleton, K.J. (2001). Estimation of affine asset pricing models using the empirical characteristic function, *Journal of Econometrics* **102**, 111–141.
- [14] Vasiček, O. (1977). An equilibrium characterization of the term structure, *Journal of Financial Economics* **5**, 177–188.

Related Articles

Approach; Heston Model; Simulation of Square-root Processes; Term Structure Models.

Cox–Ingersoll–Ross (CIR) Model; Gaussian Interest-Rate Models; Heath–Jarrow–Morton

CHRISTA CUCHIERO, JOSEF TEICHMANN &
DAMIR FILIPOVIC

Markovian Term Structure Models

An interest rate model is said to be Markovian (*see Markov Processes*) in N state variables if all discount factors at any future date can be written as a function of an N -dimensional Markov process. HJM [7] and Libor market models are generally not Markov in a limited number of state variables—the full yield curve has to be included as a Markov state variable. The separable forward rate volatility structure in the HJM introduced by Babbs [4], Cheyette [5], Jamshidian [8], and Ritchken and Sankarasubramanian [11] avoids this problem. Specifically, if the dimension of the driving Brownian motion is n , then a model with separable volatility will have a Markov representation in $N = n + n(n + 1)/2$ state variables. We discuss the connection to short rate models in general and Gaussian models in particular, calibration techniques, simulation and finite-difference implementation, and the specification of multifactor separable models.

Non-Markovian Nature of HJM Models

Let $P(t, T)$ be the time t price of a zero-coupon bond with maturity T . The continuously compounded forward rates are given by

$$f(t, T) = -\frac{\partial \ln P(t, T)}{\partial T} \quad (1)$$

and the short rate is given by

$$r(t) = f(t, t) \quad (2)$$

Under the assumption of continuous dynamics and one driving Brownian motion, in [3, 7] it is shown that the absence of arbitrage implies that forward rates evolve according to

$$df(t, T) = \sigma(t, T) \left(\int_t^T \sigma(t, s) ds \right) dt + \sigma(t, T) dW(t) \quad (3)$$

where $\{\sigma(t, T)\}_{t \leq T}$ is a family of volatility processes and W is a Brownian motion under the risk-neutral measure, that is, the martingale measure under

which the bank account $B(t) = \exp\left(\int_0^t r(u) du\right)$ is the numeraire.

This shows that all that is required to construct an arbitrage-free interest rate model that automatically fits the initial yield curve is a specification of the forward rate volatility structure $\{\sigma(t, T)\}_{t \leq T}$.

The problem, however, with arbitrary specification of the volatility structure is that the resulting model will generally not be Markov in a limited number of state variables. In general, the whole continuum $\{f(t, T)\}_{t \leq T}$ has to be used as state variables for the model. This is true regardless of the dimension of the driving Brownian motion and it is also the case for deterministic forward rate volatility structures. It should be stressed that the Libor market model exhibits the same problem. Generally, it will require all the modeled discrete forward rates as Markov state variables.

Separable Volatility Structure

Heath–Jarrow–Morton Approach gives the necessary and sufficient conditions on the forward rate volatility structure for the resulting HJM model to be Markov. An important subset of the general class of Markov HJM models is the separable volatility structure models, independently introduced by Babbs [4], Cheyette [5], Jamshidian [8], and Ritchken and Sankarasubramanian [11]. For the one-factor case, the separable form is to assume that the forward rate volatility structure is given by

$$\sigma(t, T) = g(T)h(t) \quad (4)$$

where g is a deterministic function and h is a process.

Under the assumption (4), equation (3) can be rewritten as

$$f(t, T) = f(0, T) + \frac{g(T)}{g(t)}x(t) + y(t) \int_t^T \frac{g(s)}{g(t)} ds \quad (5)$$

where

$$dx(t) = \left(\frac{g'(t)}{g(t)}x(t) + y(t) \right) dt + g(t)h(t) dW(t), \quad x(0) = 0$$

$$\begin{aligned} dy(t) &= (g(t)^2 h(t)^2 + 2 \frac{g'(t)}{g(t)} y(t)) dt, \\ y(0) &= 0 \end{aligned} \quad (6)$$

By defining $\kappa(t) = -g'(t)/g(t)$, $\eta(t) = g(t)h(t)$ and integrating equation (5) we obtain the more convenient model representation

$$\begin{aligned} P(t, T) &= \frac{P(0, T)}{P(0, t)} e^{-G(t, T)x(t) - \frac{1}{2}G(t, T)^2 y(t)} \\ dx(t) &= (-\kappa(t)x(t) + y(t)) dt + \eta(t) dW(t), \\ x(0) &= 0 \\ dy(t) &= (\eta(t)^2 - 2\kappa(t)y(t)) dt, \quad y(0) = 0 \\ G(t, T) &= \int_t^T e^{-\int_t^s \kappa(u) du} ds \end{aligned} \quad (7)$$

So if we assume $\eta = \eta(t, x(t), y(t))$, then we have a Markov representation of the full yield curve in the state variables x, y . Here, we can interpret x as a stochastic yield curve factor that perturbs the yield curve and y as a locally deterministic convexity term that has to be included to keep the model arbitrage-free.

It should be noted here that the bond prices are exponentially affine in the state variables. The separable model thus belongs to the general class of affine models (*see Affine Models*). In this class it is a special member, as the model's second locally deterministic state variable (y) eliminates the need for $\eta(t, x, y)^2$ to be linear in (x, y) , as in the models studied in [6].

From equation (5) we note that $r(t) = f(0, t) + x(t)$ and consequently that the process for the short rate is

$$\begin{aligned} dr(t) &= \left[\frac{\partial f(0, t)}{\partial t} + \kappa(t)(r(t) - f(0, t)) + y(t) \right] \\ &\quad nd \times dt + \eta(t) dW(t) \end{aligned} \quad (8)$$

If we set $\eta = \lambda(t)r(t)^\beta$, we get a model that, except for the state variable y included in the drift term of the short rate, is very similar to the short rate models by Vasicek [12], and others.^a

For $\beta = 0$, or equivalently η deterministic, y becomes deterministic, and the model is equivalent to the Gaussian model, that is, a general time-dependent parameter version of the Vasicek [12] model. For this

case, equation (7) can be seen as a convenient implementation of a time-dependent Vasicek model. The fact that the separable volatility models have a structure that is very close to the Gaussian models led Babbs [4] and Jamshidian [8] to term these models *quasi-Gaussian* and *pseudo-Gaussian*, respectively.

Another feature shared with the Gaussian model is that for fixed κ the distribution of the rates at time t only depends on the volatility up to time t , $\{\eta(u)\}_{0 \leq u \leq t}$. This means that the model can be bootstrap calibrated to swaption prices, by calibrating the model to one swaption expiry at the time. This is not the case for general short rate models because the bond price $P(t, T)$ generally depends on the short rate volatility over the interval $[t, T]$. So short rate models generally have to be calibrated to swaption prices using global routines.

Model Implementation

The forward par swap rate for swapping over the dates $t_0, t_1, \dots, t_n$ is given by

$$\begin{aligned} S(t) &= \frac{P(t, t_0) - P(t, t_n)}{A(t)}, \\ A(t) &= \sum_{i=1}^n \delta_i P(t, t_i), \quad \delta_i = t_i - t_{i-1} \end{aligned} \quad (9)$$

Following **Forward and Swap Measures**, we have that the swap rate is a martingale under the annuity measure, so we can write

$$dS(t) = \frac{\partial S(t)}{\partial x} \eta(t) dW^A(t) \quad (10)$$

where W^A is a Brownian motion under the martingale measure with the annuity A as numeraire. This can be used for deriving approximations for the value of swaptions using the same techniques as in the Libor market model literature. Specifically, we may, for example, approximate the stochastic differential equation (SDE) (10) by

$$dS(t) = \lambda(\beta S(t) + (1 - \beta)S(0)) dW^A(t) \quad (11)$$

Matching the diffusions in equations (10) and (11) in level and derivative with respect to x along the path $x = y = 0$ yields

$$\begin{aligned}\lambda^2 &= t^{-1} S(0)^{-2} \int_0^t [(S_x(u)^2 \eta(u)^2)]_{x=y=0} du \\ \beta &= \frac{\int_0^t [S_x(u) S_{xx}(u) \eta(u)^2 + S_x(u)^2 \eta(u) \eta_x(u)]_{x=y=0} du}{\lambda^2 S(0) \int_0^t [S_x(u)]_{x=y=0} du}\end{aligned}\quad (12)$$

where we have used subscripts for derivatives, so $S_x = \partial S / \partial x$, $\eta_x = \partial \eta / \partial x$.

From this we have that the swaption prices of the model can be approximated by

$$\begin{aligned}E^A[(S(t) - K)^+] &\approx \frac{1}{\beta} [S(0) \Phi(z_+) - \bar{K} \Phi(z_-)] \\ \bar{K} &= \beta K + (1 - \beta) S(0) \\ z_{\pm} &= \frac{\ln(S(0)/\bar{K})}{\beta \lambda \sqrt{t}} \pm \frac{1}{2} \beta \lambda \sqrt{t}\end{aligned}\quad (13)$$

More refined approximations based on the Markovian projection techniques of Piterbarg [10] can be found in [2].

Let $0 = t_0 < t_1 < \dots$ be a simulation time line. Then using equation (5) it can be shown that

$$\begin{aligned}x(t_{i+1}) &= \frac{g(t_{i+1})}{g(t_i)} x(t_i) + \frac{g(t_{i+1})}{g(t_i)} \left(\int_{t_i}^{t_{i+1}} \frac{g(u)}{g(t_i)} du \right) \\ &\quad \times y(t_i) + \int_{t_i}^{t_{i+1}} \frac{g(t_{i+1})}{g(u)} \eta(u) dW^{t_{i+1}}(u) \\ y(t_{i+1}) &= \left(\frac{g(t_{i+1})}{g(t_i)} \right)^2 y(t_i) + \int_{t_i}^{t_{i+1}} \left(\frac{g(t_{i+1})}{g(u)} \right)^2 \\ &\quad \times \eta(u)^2 du\end{aligned}\quad (14)$$

where W^T is a Brownian motion under the martingale measure with $P(\cdot, T)$ as numeraire, that is, the maturity T forward measure.

If we make the approximation $\eta(t) = \eta(t_i)$ for $t \in [t_i, t_{i+1}]$ then equation (14) provides a simulation scheme that produces bias-free pricing of all bonds, in the sense that if we generate discrete paths of $\{x(t_i), y(t_i)\}_{i=0,1,\dots}$ using equation (14) and use these

for producing bond prices, then for all n

$$\begin{aligned}P(0, t_n) &= \hat{E}[B(t_n)], \\ B(t_n) &= \prod_{i=0}^{n-1} P(t_i, t_{i+1}; x(t_i), y(t_i))\end{aligned}\quad (15)$$

where $\hat{E}[\cdot]$ denotes the simulation mean. This is so because over each time step the scheme (14) is the exact simulation of a Gaussian model.

The pricing partial differential equation (PDE) associated with the model is

$$\begin{aligned}0 &= \frac{\partial V}{\partial t} + D_x V + D_y V \\ D_x &= -\frac{r}{2} + (-\kappa x + y) \frac{\partial}{\partial x} + \frac{1}{2} \eta^2 \frac{\partial^2}{\partial x^2} \\ D_y &= -\frac{r}{2} + (\eta^2 - 2\kappa y) \frac{\partial}{\partial y}\end{aligned}\quad (16)$$

The absence of diffusion term in the second dimension can make the finite-difference solution of the PDE quite challenging, and this suggests the use of upwind and fully implicit schemes to prevent “ringing” in the numerical solution which would reduce the accuracy of the solution. Andreasen [1], however, reports good practical results with the $O(\Delta t^2 + \Delta x^2 + \Delta y^4)$ accurate Mitchell scheme [9]

$$\begin{aligned}&\left[\frac{1}{\Delta t} - \frac{1}{2} D_x \right] U(t) \\ &= \left[\frac{1}{\Delta t} + \frac{1}{2} D_x + D_y \right] V(t + \Delta t) \\ &\left[\frac{1}{\Delta t} - \frac{1}{2} D_y \right] V(t) \\ &= \frac{1}{\Delta t} U(t) - \frac{1}{2} D_y V(t + \Delta t)\end{aligned}\quad (17)$$

used with a standard 3-point discretization of D_x and a 5-point discretization of D_y . The 5-point discretization in the second dimension eliminates the need for the use of upwind schemes at a slightly higher computational cost than would be the case for a standard 3-point scheme.

Multiple Factors

The multifactor counterpart to equation (3) is

$$df(t, T) = \sigma(t, T) \cdot \left(\int_t^T \sigma(t, s) ds \right) dt + \sigma(t, T) \cdot dW(t) \quad (18)$$

where $\{\sigma(t, T)\}_{t \leq T}$ is a family of n -dimensional vector processes, W is an n -dimensional vector Brownian motion, and $\cdot$ denotes vector product. In the n factor separable volatility structure model, the forward rate volatility is given by

$$\sigma(t, T) = h(t)g(T) \quad (19)$$

where g is a deterministic vector function on $\mathbb{R}_+^n$ and h is a matrix process on $\mathbb{R}^{n \times n}$. Defining

$$\kappa_i(t) = -\frac{g_i'(t)}{g_i(t)} \quad (20)$$

and

$$\eta_{ij}(t) = g_i(t)h_{ij}(t) \quad (21)$$

the separable volatility model can be written as follows [5]:

So we have a Markov representation involving $n + n(n + 1)/2$ state variables with $x = (x_i)$ being a vector of stochastic yield curve factors and the symmetric matrix $y = (y_{ij})$ being a locally deterministic convexity term that has to be pulled along in simulation of the model to keep the model arbitrage-free.

The number of state variables grows at a quadratic rate and this prevents the use of finite-difference methods for $n > 1$. There are, however, very significant computational savings associated with using this type of model rather than a general HJM (see **Heath–Jarrow–Morton Approach**) or LIBOR market model (see **LIBOR Market Model**) approach even though Monte Carlo simulations have to be used for the numerical solution. For the case of $n = 4$ driving Brownian motions the number of state variables is 14, which should be compared against the 120 state variables of a 30-year quarterly Libor market model.

If we let the mean-reversion parameters, $\kappa_1, \dots, \kappa_n$ be constant, then

$$\sigma_j(t, t + \tau) = \sum_{i=1}^n e^{-\kappa_i \tau} \eta_{ij}(t) \rightarrow \int e^{-\kappa \tau} \eta_{\kappa j}(t) d\kappa \quad (23)$$

for $n \rightarrow \infty$ and an appropriately chosen sequence $\kappa_1, \kappa_2, \dots$. So model (22) can be seen as a representation of the forward rate volatility structure on a (discrete) basis of exponential functions. The function $\kappa \mapsto \eta_{\kappa j}(t)$ can thus be viewed as the inverse Laplace transform of the j th component of the forward rate volatility structure in the tenor dimension: $\tau \mapsto \sigma_j(t, t + \tau)$.

$$P(t, T) = \frac{P(0, T)}{P(0, t)} e^{-\sum_{i=1}^n G_i(t, T)x_i(t) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n G_i(t, T)y_{ij}(t)G_j(t, T)}$$

$$dx_i(t) = (-\kappa_i(t)x_i(t) + \sum_{j=1}^n y_{ij}(t)) dt + \sum_{j=1}^n \eta_{ij}(t) dW_j(t)$$

$$dy_{ij}(t) = \left(\sum_{k=1}^n \eta_{ik}(t)\eta_{jk}(t) - (\kappa_i(t) + \kappa_j(t))y_{ij}(t) \right) dt$$

$$G_i(t, T) = \int_t^T e^{-\int_t^s \kappa_i(u) du} ds \quad (22)$$

Still, representation (22) is not particularly concrete in relating the dynamics of the state variables to the dynamics of observable rates. To do so we fix a set of tenors $\tau_1, \dots, \tau_n$ and consider the forward rate vector $F(t) = (f(t, t + \tau_1), \dots, f(t, t + \tau_n))'$. We have

$$dF(t) = \Sigma(t) dW(t) + O(dt)$$

$$\Sigma(t) = \begin{bmatrix} \sigma_1(t, t + \tau_1) & \dots & \sigma_n(t, t + \tau_1) \\ \vdots & \ddots & \vdots \\ \sigma_1(t, t + \tau_n) & \dots & \sigma_n(t, t + \tau_n) \end{bmatrix} \quad (24)$$

In model (22) we have

$$dF(t) = \Gamma(t)\eta(t) dW(t) + O(dt)$$

$$\Gamma(t) = \begin{bmatrix} g_1(t, t + \tau_1) & \dots & g_n(t, t + \tau_1) \\ \vdots & \ddots & \vdots \\ g_1(t, t + \tau_n) & \dots & g_n(t, t + \tau_n) \end{bmatrix}$$

$$g_i(t, T) = e^{-\int_t^T \kappa_i(u) du} \quad (25)$$

If we equate the diffusion terms in equations (24) and (25) we get

$$\eta(t) = \Gamma(t)^{-1} \Sigma(t) \quad (26)$$

So for this choice, models (22) and (24) exactly match on the dynamics of the selected forward rates. Andreasen [2] uses this technique to construct a separable volatility structure model that mimics the dynamics of a Libor market model with rate-dependent and stochastic volatility, and it is shown that a four-factor version of such a model is capable of fitting the full cap and swaption market for all expiries and tenors and strikes within quite narrow tolerances.

End Notes

^aThis volatility specification is suggested in [11]. In [1] it is suggested to model volatility to have dependencies of longer tenor rates.

References

- [1] Andreasen, J. (2000). *Turbo-Charging the Cheyette Model*. Working paper, General Re Financial Products.
- [2] Andreasen, J. (2005). Back to the future, *Risk* September, 43–48.
- [3] Babbs, S. (1990). *The Term Structure of Interest Rates: Stochastic Processes and Contingent Claims*. PhD thesis, Imperial College, London.
- [4] Babbs, S. (1993). Generalised Vasicek Models of the term structure, *Applied Stochastic Models and Data Analysis* **1**, 49–62.
- [5] Cheyette, O. (1992). *Markov Representation of the Heath-Jarrow-Morton Model*. Working paper, BARRA.
- [6] Duffie, D. & Kan, R. (1996). A yield-factor model of interest rates, *Mathematical Finance* **6**, 379–406.
- [7] Heath, D., Jarrow, R. & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation, *Econometrica* **60**, 77–106.
- [8] Jamshidian, F. (1991). Bond and option evaluation in the Gaussian interest rate model, *Research in Finance* **9**, 131–170.
- [9] Mitchell, A. & Griffiths, D. (eds) (1980). *The Finite Difference Method in Partial Differential Equations*, John Wiley & Sons, New York.
- [10] Piterbarg, V. Time to smile, (2005). *Risk* May, 52–56.
- [11] Ritchken, P. & Sankarasubramaniam, L. (1993). *On Finite State Markovian Representations of the Term Structure*. Working paper, Department of Finance, University of Southern California.
- [12] Vasicek, O. (1977). An equilibrium characterization of the term structure, *Journal of Financial Economics* **5**, 177–188.

Further Reading

- Andersen, L. & Andreasen, J. (2002). Volatile volatilities, *Risk* December, 163–168.
- Bjork, T. & Landen, C. (2001). A geometric view of interest rate theory, in *Option Pricing, Interest Rates and Risk Management*, E. Jouini, J. Cvitanic & M. Musiela, eds, Cambridge University Press, pp. 241–277.
- Cox, J., Ingersoll, J. & Ross, S. (1985). A theory of the term structure of interest rates, *Econometrica* **53**, 385–408.
- Filipovic, D. (2001). *Consistency Problems for Heath-Jarrow-Morton Interest Rate Models* (Lecture Notes in Mathematics 1760), Springer-Verlag.

Related Articles

Affine Models; Finite Difference Methods for Barrier Options; Gaussian Interest-Rate Models; Heath–Jarrow–Morton Approach; Markov Processes; Partial Differential Equations; Quadratic Gaussian Models.

JESPER ANDREASEN

Swap Market Models

The Black formula [2], (see **Caps and Floors**) is popular among practitioners as a simple tool to price European options on Libor rates, that is, caplets and floorlets (see **Caps and Floors**), and on swap rates (see **LIBOR Rate**), that is, swaptions. More recently, Brace *et al.* [3], Miltersen *et al.* [12], and Jamshidian [11] provided a sound theoretical basis to this practice by introducing a general framework to consistently price interest rate options by no-arbitrage arguments. These works paved the way toward a broader acceptance of the so-called *market models* for interest-rate derivatives by the academic community since they can be recast within the general arbitrage-free framework discussed in [9] (see **Heath–Jarrow–Morton Approach**). These models have the advantage, over those based on the evolution of the spot interest rate, of concentrating on rates that are market observable.

The *Libor market model* [3, 12], (see **LIBOR Market Model**) and the *co-terminal swap market model* [11] are the two major representatives of this class. These models are built by assigning arbitrage-free dynamics on a set of forward Libor rates and of co-terminal forward swap rates, respectively. The advent of new kinds of exotic (over-the-counter) derivatives in fixed income markets has recently inspired the introduction of “hybrid” or “generalized” market models where the underlying variables constitute a mixed set comprising both Libor and swap rates simultaneously. In this context, an extensive study is provided in Galluccio *et al.* [5, 6]. The availability of a general setup to build market models with mixed sets is of interest in applications, for example, to better capture the risk embedded in some complex financial derivatives.

Tenor Structure and Forward Swap Rates

We assume that we are given a prespecified collection of reset/settlement dates $T = \{T_1, \dots, T_M\}$, referred to as the *tenor structure*, with $T_j < T_k, 1 \leq j < k \leq M$, and starting time $T_0 < T_1$. Let us denote the year fraction between any two consecutive dates by $\delta_j = T_j - T_{j-1}$, for $j = 1, \dots, M$. We write $P(t, T_j)$, $j = 1, \dots, M$, to denote the price at time t of a *discount bond* that matures at time $T_j > t$. The

forward swap rate $S(t, T_j, T_k)$, with j and k satisfying $1 \leq j < k \leq M$, is defined through $S(t, T_j, T_k) = (P(t, T_j) - P(t, T_k))/G(t, T_j, T_k)$ for all $t \in [0, T_j]$. Here, $G(t, T_j, T_k)$ is the price of the *annuity* (or level) *numéraire*. Swap market models are based on the continuous time modeling of $S(t, T_j, T_k)$ and, generally, assume that forward swap rates follow a multidimensional diffusion process. In particular, $S(t, T_j, T_k)$ is a P^{T_j, T_k} -martingale so that, under P^{T_j, T_k} ,

$$\frac{dS(t, T_j, T_k)}{S(t, T_j, T_k)} = \lambda(t, T_j, T_k)' dW^{T_j, T_k}(t) \quad \forall t \in [0, T_j] \quad (1)$$

where $\lambda(t, T_j, T_k)$ is a vector valued volatility function. The probability measure P^{T_j, T_k} is equivalent to the historical probability measure P , and is called the *forward swap probability measure* associated with the dates T_j and T_k , or simply the forward swap measure (see **Forward and Swap Measures**). For every $i = 1, \dots, M$, the relative (or “deflated”) bond $P(t, T_i)/G(t, T_i, T_k)$, $\forall t \in [0, \min(T_i, T_{j+1})]$, follows a local martingale process under P^{T_j, T_k} . We denote the corresponding Brownian motion under P^{T_j, T_k} by W^{T_j, T_k} . The *forward Libor rate* $L(t, T_j)$, $j = 1, \dots, M - 1$, defined as

$$L(t, T_j) = \frac{P(t, T_j) - P(t, T_{j+1})}{\delta_{j+1} P(t, T_{j+1})} \quad \forall t \in [0, T_j] \quad (2)$$

is itself a forward swap rate $S(t, T_j, T_k)$ corresponding to $k = j + 1$; its volatility function is denoted by $\lambda(t, T_j)$. Accordingly, we denote by P^{T_j} the corresponding forward probability measure associated to the discount bond price $P(t, T_j)$, and by W^{T_j} a Brownian motion under P^{T_j} . Then, for every $i = 1, \dots, M$, the relative bond price $P(t, T_j)/(\delta_{j+1} P(t, T_{j+1}))$, $\forall t \in [0, \min(T_i, T_{j+1})]$, follows a local martingale under $P^{T_{j+1}}$ (see **Forward and Swap Measures**). We refer to [13] (Chapters 12 and 13) for further material on the theoretical side.

In [5, 6], the authors introduce the so-called *market model approach*, and investigate the weakest condition under which a general specification of a model for observable forward swap rates has a unique specification in all equivalent pricing measures. In this respect, the concept of *admissibility* of a set is introduced, and its theoretical and practical implications are discussed. Interestingly, the properties of these admissible sets can be best understood with

the use of graph theory. This mapping allows to graphically characterize all admissible sets in a simple and intuitive way, so that model selection for a given tenor structure can be performed by visual inspection. Further, it is possible to prove that the class of admissible market models is very large: for a given tenor structure $T = \{T_1, \dots, T_M\}$ comprising M dates, there exist M^{M-2} admissible sets (and models). Admissible models comprise all “standard” market models [3, 11] as special cases. Three major subclasses denominated *co-initial*, *co-sliding*, and *co-terminal* (according to the nature of the family of forward swap rates) can be identified. We hereby briefly discuss their respective features. Remarkably, the Libor market model is the only admissible model of the co-sliding type.

Co-terminal Swap Market Model

The co-terminal swap market model dates back to [11], and is built from an admissible set of forward swap rates with different start dates $\{T_1, \dots, T_{M-1}\}$ and equal maturity date T_M , so that forward swap rates satisfy equation (1). The model is best suited to price *Bermudan swaptions* (and related derivatives; see [5, 6, 15]) where the holder has the right to enter at times $T_1, \dots, T_{M-1}$ into a plain-vanilla swap maturing at T_M . In this case, the only relevant European swaptions from a pricing and hedging perspective are those expiring at $T_1, \dots, T_{M-1}$, and maturing at T_M . Hence, it is natural to introduce a market model where the relevant underlying set coincides with the associated co-terminal forward swap rates. Other derivative securities with similar characteristics include callable cap and reverse floaters, ratchet cap floaters, and Libor knock-in/out swaps; all are good candidates for valuation in a co-terminal framework.

Co-initial Swap Market Model

The co-initial swap market model [5–7] is built from an admissible set of forward swap rates with different end dates $\{T_2, \dots, T_M\}$ and equal start date T_1 , so that forward swap rates satisfy equation (1). The model is best suited to price (complex) European-style derivatives, where the holder owns the right to exercise an option at a single future date T . In this case, the option payoff, no matter how complex, is

measurable with respect to the information available at time T , by definition. Qualitatively speaking, a set of admissible forward swap rates sharing the same initial date T contains all the information needed to evaluate the payoff, the latter being a function of a set of admissible co-initial forward swap rates at that time. Hence, a model market approach based on a set of co-initial forward swap rates provides a powerful tool to price and hedge a large variety of European-style derivatives including forward-start, amortizing, and zero-coupon swaptions.

Co-sliding Swap Market Model

In [5, 6], it is shown that there exists a unique admissible co-sliding swap market model and that it coincides with the Libor market model [3, 12]. The model is built from an admissible set of forward swap rates with start date T_j and end date T_{j+1} ($j = 1, \dots, M-1$), so that forward swap rates satisfy equation (1). They are thus associated to swaps with the same time to maturity. It is easy to see that nonoverlapping forward swap rates of that form are indeed forward Libor rates. The co-sliding model is best suited to price structured constant-maturity swap (CMS)-linked derivatives (with possibly Bermudan features) whose payoff function depends on a set of fixed-maturity instruments (see **Bermudan Swaptions and Callable Libor Exotics**). More precisely, in a CMS (see **Constant Maturity Swap**), the variable coupon that settles at a generic time T_j is linked to the value of a swap rate prevailing at that time (the latter being associated to a swap of a given maturity). In this context, the Libor market model provides an optimal modeling framework since (sliding) CMS rates can be easily described in terms of linear combinations of forward Libor rates.

Numerical Implementation

In the applications, one needs to calibrate a generic swap market model to the available prices of liquid vanilla derivatives to avoid potential arbitrage in the risk-management process. Implied model calibration is a reverse engineering procedure aimed at identifying the relevant model characteristics, such as volatility parameters, from such a set of instruments. In interest-rate derivatives markets, these instruments are plain-vanilla options written on forward swap and

Libor rates, that is, swaptions and caplets, respectively. To achieve a fast and robust model calibration, one should ideally aim at closed or quasi-closed form formulae for plain-vanilla option prices. When these are not available, good analytical approximations are called for. The accuracy of these methods is studied in several papers. They rely on the so-called *freezing approach* [1, 10, 14] or, alternatively, on the *rank-one approximation* method [3]. In turn, the specification of the instantaneous volatility function $\lambda(t, T_j, T_k)$ in equation (1) is sometimes done by introducing flexible functional forms [4, 15] that are meant to reproduce the observed shape of implied swaption and cap/floor volatility term structures through a low-dimensional parameterization. One of the most appreciated features of the instantaneous volatility function among practitioners is *time stationarity*. This is generally imposed to reproduce the analogous temporal evolution of the volatility term structure observed in the market. However, the constraint of perfect model stationarity is generally incompatible with the observed implied volatility market for a generic well-behaved instantaneous volatility function. This market feature forces practitioners to introduce explicit calendar-time dependent functions $\lambda(t, T_j, T_k)$ to mimic a perturbation mode around the time-stationary solution. Efficient simulation algorithms are also available to price exotic interest-rate derivatives by Monte Carlo methods [8] (see **LIBOR Market Models: Simulation**).

Acknowledgments

The author thanks the Swiss NSF for financial support through the NCCR Finrisk.

References

- [1] Andersen, L. & Andreasen, J. (2000). Volatility skews and extensions of the Libor market model, *Applied Mathematical Finance* **7**, 1–32.
- [2] Black, F. (1976). The pricing of commodity contracts, *Journal of Financial Economics* **3**, 167–179.
- [3] Brace, A., Gatarek, D. & Musiela, M. (1997). The market model of interest rate dynamics, *Mathematical Finance* **7**, 127–155.
- [4] Brigo, D. & Mercurio, M. (2001). *Interest Rates Models: Theory and Practice*, Springer-Verlag, Heidelberg.
- [5] Galluccio, S., Huang, Z., Ly, J.-M. & Scaillet, O. (2007). Theory and calibration of swap market models, *Mathematical Finance* **17**, 111–141.
- [6] Galluccio, S. & Scaillet, O. (2007). *Constructive Theory of Market Models for Interest-Rate Derivatives*, BNP Paribas working paper.
- [7] Galluccio, S. & Hunter, C. (2004). The co-initial swap market model, *Economic Notes* **33**, 209–232.
- [8] Glasserman, P. & Zhao, X. (2000). Arbitrage-free discretization of lognormal forward Libor and swap rate models, *Finance and Stochastics* **4**, 35–68.
- [9] Heath, D., Jarrow, R. & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology for contingent claims valuation, *Econometrica* **60**, 77–105.
- [10] Hull, J. & White, A. (2000). Forward rate volatilities, swap rate volatilities and the implementation of the Libor market model, *Journal of Fixed Income* **10**, 46–62.
- [11] Jamshidian, F. (1997). Libor and swap market models and measures, *Finance and Stochastics* **1**, 293–330.
- [12] Miltersen, K., Sandmann, K. & Sondermann, D. (1997). Closed form solutions for term structure derivatives with lognormal interest rates, *Journal of Finance* **70**, 409–430.
- [13] Musiela, M. & Rutkowski, M. (2004). *Martingale Methods in Financial Modelling*, 2nd Edition, Springer-Verlag, Berlin.
- [14] Rebonato, R. (1998). *Interest Rate Option Models*, 2nd Edition, Wiley, Chichester.
- [15] Rebonato, R. (2003). *Modern Pricing of Interest-Rate Derivatives*. Princeton University Press, Princeton.

Related Articles

CMS Spread Products; Forward and Swap Measures; Constant Maturity Swap; LIBOR Market Model.

STEFANO GALLUCCIO & OLIVIER SCAILLET

Markov Functional Models

Market models (see **LIBOR Market Model**) are formulated directly in terms of market observable rates, such as *LIBORs* (see **LIBOR Rate**), their volatilities, and correlations. They were the first models that could calibrate exactly to Black's formula for pricing liquid instruments for all strikes. Though the models provide an excellent framework for selecting and understanding a model, and have become a benchmark in the market place, they do have one significant drawback. An accurate implementation of a market model can only be done by simulation because of the high dimensionality of the model. This is true even when there is only one stochastic driver.

In this article, we describe models that can fit the observed prices of liquid instruments in a similar fashion to the market models, but which also have the advantage that derivative prices can be calculated just as efficiently as in the most tractable *short-rate model* (see **Term Structure Models**). To achieve this, we consider the general class of Markov-functional interest-rate models [4, 8], which, as we shall discuss, can be specified using the numeraire approach restricted to a finite time horizon. The defining characteristic of Markov-functional models is that pure discount bonds prices are at any time a function of some low-dimensional process that is Markovian in some martingale measure. This ensures that the implementation is efficient since it is only necessary to track the driving process, something that is particularly important for Bermudan-style products. Market models do not possess this property for some low-dimensional Markov process and this is the obstacle to their efficient implementation.

Nnumeraire Approach to Specifying Interest-rate Models

Let $\{\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P}\}$ be a filtered probability space. Denote by D_{tT} the value at t of a pure discount bond with maturity T , an asset that pays a unit amount on its maturity date. There are several ways to model the complete term structure of pure discount bonds, $\{D_{tT} : 0 \leq t \leq T \leq \infty\}$ and it is not necessary to specify the model under the "real-world" measure $\mathbb{P}$. The

numeraire approach follows from the definition of a numeraire pair. Let $\mathbb{N}$ be an *equivalent martingale measure* (see **Equivalent Martingale Measures**) corresponding to a numeraire N (see **Change of Numeraire**). Then any numeraire-rebased asset is a martingale under the measure $\mathbb{N}$.

Noting that $D_{TT} = 1$, we see that

$$D_{tT} = \mathbb{N}_t \mathbb{E}_{\mathbb{N}} [N_T^{-1} | \mathcal{F}_t] \quad (1)$$

Thus, once we have specified the numeraire pair $(N, \mathbb{N})$ and the filtration $\{\mathcal{F}_t\}$, we have defined a term structure model under the equivalent martingale measure $\mathbb{N}$.

For practical applications, it is usual to restrict attention to a finite time horizon, $0 \leq t \leq T$. Here, to specify a model using the numeraire approach, in addition to knowledge of the numeraire, we need to know its joint distribution under $\mathbb{N}$ with the pure discount bonds on some boundary curve. In applications that we have encountered, it is sufficient to take these bonds to be D_{tS} , for $S \in [T, T']$ for some fixed $T' \geq T$. Then we can again use the martingale property of numeraire-rebased assets to recover the pure discount bonds at earlier times for maturities up to T' .

Markov-functional Models

We now give a formal definition of a Markov-functional model. We restrict attention to a finite time horizon and so the definition includes the boundary curve mentioned above. This definition is very general and, if we allow the process driving the model to be of high dimension, the definition will encompass nearly all models of practical interest (including market models). The real spirit of Markov-functional modeling is explained in the next section when we discuss how to recover the prices of calibrating liquid instruments *via* a functional sweep.

Definition 1 *An interest-rate model is said to be Markov-functional if there exists some numeraire pair $(N, \mathbb{N})$ and some process x such that*

1. *the process x is a (time-inhomogeneous) Markov process under the measure $\mathbb{N}$;*
2. *the pure discount bonds are of the form*

$$D_{tS} = D_{tS}(x_t), \quad 0 \leq t \leq \min\{S, T\} \quad (2)$$

From our discussion of the numeraire approach, we see that to completely specify a general Markov-functional model, it is sufficient to specify the joint law of the numeraire N and the process x under $\mathbb{N}$, and the functional form of the discount factors on the boundary. This observation forms the basis for setting up a Markov-functional model in practice.

Recap of the Standard Markov-functional Approach

The formal definition above tells us the general properties of the model we wish to develop, but nothing about the practicalities of how to set up a model for a particular pricing problem. Here, we summarize the standard case in which the driving process, denoted by x , is chosen to be of low dimension and Gaussian. This ensures that the model is efficient to implement. We comment on the choice of covariance structure for x below. Assuming x has been chosen, the practical problem we now seek to address is that of setting up a model that is arbitrage-free and calibrates well to a set of vanilla instruments appropriate for the product we wish to price. In this article, we restrict our discussion to a special case—the LIBOR Markov-functional model in the terminal measure. For further details on practical aspects of this model, see [4, 6, 8] and [10].

As in a *LIBOR market model* (see **LIBOR Market Model**), we assume we have a set of contiguous forward LIBORs denoted by L^i for $i = 1, \dots, n$ corresponding to tenor structure $T_1, \dots, T_{n+1}$. We write $S_i := T_{i+1}$ for $i = 1, \dots, n$ and so L^i is the LIBOR corresponding to the period $[T_i, S_i]$.

In this section, we use $\mathbb{N}$ to denote the terminal measure corresponding to taking the bond $D_{\cdot, S_n}$ as numeraire and $\mathbb{E}$ to denote expectations in this measure.

Consider the problem of how to choose the functional forms so that the resulting model calibrates accurately to the prices of the set of caplets (equivalently digital caplets) corresponding to the forward LIBORs $L^1, \dots, L^n$. The model is actually only specified on a grid. That is, we specify the functional forms $D_{T_i T_j}(x_{T_i})$ for $1 \leq i < j \leq n+1$, since this is all that is (typically) needed in practice. Note that here all we need to recover these discount factors are the functional forms of the numeraire for times $T_1, \dots, T_n$. The functional forms are derived numerically from market prices and the martingale properties necessary to make the model arbitrage-free.

The algorithm for finding the functional forms works back iteratively from the terminal time T_n . Suppose we have reached T_i , having already found the functional forms $D_{T_k S_n}(x_{T_k})$, $k = i+1, \dots, n+1$. Trivially, this is true when $i = n$ as there is nothing to know since $D_{T_{n+1} S_n}(x_{T_{n+1}}) = 1$. Then from the martingale property of numeraire-rebased assets we can find

$$\begin{aligned} \hat{D}_{T_i T_{i+1}}(x_{T_i}) &:= \frac{D_{T_i T_{i+1}}(x_{T_i})}{D_{T_i S_n}(x_{T_i})} \\ &= \mathbb{E} \left[\frac{1}{D_{T_{i+1} S_n}}(x_{T_{i+1}}) \mid x_{T_i} \right] \end{aligned} \quad (3)$$

Noting, by definition, that

$$L_{T_i}^i = \frac{1 - D_{T_i T_{i+1}}}{\alpha_i D_{T_i T_{i+1}}} \quad (4)$$

where α_i is the accrual factor for the interval $[T_i, S_i]$, it follows that

$$D_{T_i S_n}(x_{T_i}) = \frac{1}{\hat{D}_{T_i T_{i+1}}(x_{T_i})(1 + \alpha_i L_{T_i}^i(x_{T_i}))} \quad (5)$$

Thus we see that to determine the functional form for the numeraire at time T_i , $D_{T_i S_n}(x_{T_i})$, it is sufficient to find the functional form for $L_{T_i}^i(x_{T_i})$. Equivalently, it is sufficient to find the off-diagonal discount factor $D_{T_i T_{i+1}}(x_{T_i})$.

We begin with the case in which the process x is one dimensional. Here, we view the process x as capturing the overall level of interest rates.

One-dimensional Case. In setting up our model, we make the assumption that the i th forward LIBOR at time T_i , $L_{T_i}^i$, is a monotonic increasing function of the variable x_{T_i} , that is, we assume that $L_{T_i}^i = f^i(x_{T_i})$, for some monotonic increasing function f^i . The functional forms are found using market prices of digital caplets. This is equivalent to calibrating to caplets as we can recover the price of a caplet with strike K , $C^i(K)$, from the prices of digital caplets:

$$C^i(K) = \int_K^\infty V^i(\hat{K}) d\hat{K} \quad (6)$$

where $V^i(K)$ denotes the market value of the digital caplet with strike K (setting at T_i , paying at S_i). In an arbitrage-free model, we must have

$$V^i(K) = D_{0, S_n} \mathbb{E} \left[\frac{D_{T_i S_i}}{D_{T_i S_n}} \mathbf{1}_{\{L_{T_i}^i > K\}} \right] \quad (7)$$

Choose a grid of values x^* and for each x^* calculate

$$\begin{aligned} J_0^i(x^*) &:= D_{0S_n} \mathbb{E} \left[\frac{D_{T_i S_i}}{D_{T_i S_n}}(x_{T_i}) \mathbf{1}\{x_{T_i} > x^*\} \right] \\ &:= D_{0S_n} \mathbb{E} \left[\mathbb{E} \left[\frac{D_{T_{i+1} S_i}}{D_{T_{i+1} S_n}}(x_{T_{i+1}}) | x_{T_i} \right] \mathbf{1}\{x_{T_i} > x^*\} \right] \end{aligned} \quad (8)$$

We can do this (numerically) from what we already know, having calibrated the model at all $T_j, j > i$. Now find the value $K(x^*)$ such that $V^i(K(x^*)) = J_0^i(x^*)$. Equating equations (7) and (9) we find that

$$\begin{aligned} D_{0S_n} \mathbb{E} \left[\frac{D_{T_i S_i}}{D_{T_i S_n}}(x_{T_i}) \mathbf{1}\{L_{T_i}^i(x_{T_i}) > K(x^*)\} \right] \\ &= V^i(K(x^*)) \\ &= J_0^i(x^*) \\ &= D_{0S_n} \mathbb{E} \left[\frac{D_{T_i S_i}}{D_{T_i S_n}}(x_{T_i}) \mathbf{1}\{x_{T_i} > x^*\} \right] \\ &= D_{0S_n} \mathbb{E} \left[\frac{D_{T_i S_i}}{D_{T_i S_n}}(x_{T_i}) \mathbf{1}\{L_{T_i}^i(x_{T_i}) > L_{T_i}^i(x^*)\} \right] \end{aligned} \quad (9)$$

Under the assumption that $L_{T_i}^i(x)$ is increasing in x , we can now conclude that $L_{T_i}^i(x^*) = K(x^*)$, thus, repeating this on a grid of values x^* , we have derived the required functional form.

Multidimensional Case. This is a straightforward generalization of the one-dimensional case. The key to this extension is to ensure in the generalization that we

1. retain the univariate and monotonicity properties that were required to make the functional fitting efficient;
2. capture the desired correlation/covariance structure.

To do this, we introduce the idea of a *prior model*. The prior model expresses each LIBOR as a function of the driving Markov process x , which is now of dimension $k > 1$. This prior model is chosen to capture the basic dynamics of the market, but may admit arbitrage. We discuss the choice of a prior model based on a market model below. Once the prior

model is chosen, the approach now is to regard the Markov-functional sweep as a (small) perturbation of this prior model, which removes the arbitrage.

In particular, we assume that the functional dependence of $L_{T_i}^i$ on the multidimensional x_{T_i} is only via the prior model LIBOR $\tilde{L}_{T_i}^i$. Thus

$$L_{T_i}^i(x_{T_i}) = f_i \left(\tilde{L}_{T_i}^i(x_{T_i}) \right) \quad (11)$$

for some monotonic function f_i . It is this specialization that enables us to achieve the univariate and monotonicity properties in this higher dimensional setting. The last step is the derivation of the functional forms f_i . This is almost identical to the one-factor case. For details, the reader is referred to [6].

Covariance Structure of x -process and Comparison with Market Models. Typically, x is taken to be a k -dimensional Gaussian process with i th component of the form

$$x_t^i = \int_0^t \sigma_s^i dW_s^i \quad (12)$$

where the W^i s are Brownian motions under the measure $\mathbb{N}$, with instantaneous correlations $dW_t^i dW_t^j = \rho_{ij}^i dt$. With x of this form, we have explicit knowledge of all marginal and conditional transition densities and all the required conditional expectations can be computed efficiently.

As in any interest-rate model, care must be taken in the choice of the instantaneous volatilities so that the resulting model has the appropriate qualitative behavior. For example, many authors illustrate the features of a model by using a simple exponential form of the instantaneous volatilities. However, use of this, in practice, would lead to a model having unrealistic hedges as the resulting correlation structure does not change in an appropriate way when the implied volatilities change. One appropriate choice based on a Hull–White short-rate model is given in [1].

Recall that in a k -dimensional Markov-functional model, the prior model is chosen with some desired correlation structure in mind. If the instantaneous volatilities $\sigma^i, i = 1, \dots, n$ are taken to be separable in that for each i , σ_t^i can be written as a vector product of constants depending on i and a common volatility function σ_t , then it is very easy to form a prior model from the corresponding k -factor LIBOR

market model with the same instantaneous volatilities. A first-order approximation to this model could, for example, be obtained by taking the usual SDE and replacing the time-dependent drift with its time-zero value. This would result in a model with something very close to the desired correlation structure, for which all LIBORs are lognormally distributed, but for which there is significant arbitrage. This makes the approximation too poor for use as a model in practice, but it remains adequate as a starting point for an arbitrage-free Markov-functional model. In fact, it is shown in [1] that in the one-dimensional case, Markov-functional and separable LMM models are very similar across a broad range of parameter values.

Generalizations. The discussion above focused on a LIBOR Markov-functional model specified in the terminal measure. For a version of this model that uses the discrete savings account as numeraire and forward induction, see [2]. The Markov-functional method is not restricted to calibrating to the market implied distributions of LIBORs. For a given tenor structure, one can formulate a model that calibrates to any swap rate or LIBOR at each time slice. In addition, the boundary can be extended so that more than one discount factor can be modeled on the final time slice. The details for a swap model can be found in [4, 8]. This is an appropriate choice for pricing *Bermudan swaptions* (see **Bermudan Swaptions and Callable Libor Exotics**) as the resulting model can be calibrated to vanilla swaption prices.

A multicurrency Markov-functional model first presented in [5] is described in [3].

If one is willing to employ Monte Carlo methods, the Markov-functional approach can be extended to formulate a high-dimensional model comparable

with a full-rank market model. See [9] for details. For further developments in the high-dimensional setting, see [7].

References

- [1] Bennett, M. & Kennedy, J. (2005). A comparison of Markov-functional and market models: the one-dimensional case, *The Journal of Derivatives* 1(2), 22–43.
- [2] Fries, C. (2007). *Mathematical Finance: Theory, Modeling, Implementation*, John Wiley & Sons.
- [3] Fries, C. & Rott, M. (2004). *Cross Currency and Hybrid Markov-Functional Models*, SSRN pre-print, at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=532122.
- [4] Hunt, P., Kennedy, J. & Pelsser, A. (2000). Markov-functional interest rate models, in *The New Interest Rate Models*, L. Hughston, ed, Risk Books.
- [5] Hunt, P. (2003). The modelling and risks of prdcs. *Proceedings of the ICBI Global Derivatives Conference*, Barcelona.
- [6] Hunt, P. & Kennedy, J. (2004). *Financial Derivatives in Theory And Practice*, 2nd Edition, John Wiley & Sons.
- [7] Hunt, P. & Kennedy, J. (2005). *Longstaff–Schwarz, Effective Model Dimensionality and Reducible Markov-Functional Models*, SSRN pre-print at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=627921.
- [8] Hunt, P., Kennedy, J. & Pelsser, A. (2000). Markov-functional interest rate models, *Finance and Stochastics* 4(1), 391–408.
- [9] Kaisajuntti, L. & Kennedy, J. (2008). *An n-Dimensional Markov-Functional Interest Rate Model*, SSRN pre-print at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1081337.
- [10] Pelsser, A. (2000). *Efficient Methods For Valuing Interest Rate Derivatives*, Springer Finance.

JOANNE KENNEDY

Hedging of Interest Rate Derivatives

Cash and the Zero Curve

The simplest contract is a unit notional, zero-coupon bond to be paid at time T (the *maturity*). The value of such a bond is denoted by $P(T)$.^a

The function P thus describes the evolution through time of interest rate expectations.

The instantaneous *forward rate* f is defined by $f(T) \equiv -P'(T)/P(T)$. Thus, the *forward curve* f provides a local view of the market forecast for future interest rates. While knowledge of P and f are in principle equivalent, the latter provides a superior framework for practical analysis.

FRAs, Swaps, and Bond Equivalence

A forward rate agreement, *FRA*, is a contract to lend at a previously agreed rate over some time interval—thus it is equivalent to a calendar spread of zero-coupon bonds. A swap is very similar to a succession of FRAs, so its price can nearly be determined from the zero curve. The slight differences between Libor swaps and coupon-paying bonds stem from the differences between the Libor end date and the period payment date, and also (in most currencies) the difference between fixed and floating payment frequencies. For a more detailed description, see **LIBOR Rate**.

Libor Futures

A Libor futures contract (see **Eurodollar Futures and Options**) pays, at its settlement, a proportion of the Libor rate fixed on the futures expiry date. However, since an FRA makes its payment only at its maturity date, its par rate in a risk-neutral world is equal to the expectation under the discount-adjusted measure to that maturity date. The daily updating of posted margins for Libor futures means that profit or loss from rate fluctuations is realized immediately; thus the par futures price reflects the risk-neutral expectation in an undiscounted measure.

Because the resulting “futures convexity adjustment” does not closely track other measures of

volatility, it is traded actively only by a few specialists. For most purposes, we can think of a future as being equivalent to an FRA plus an exogenously specified spread.

Yield Curve Construction

Since the function $P(T)$, or equivalently $f(T)$, practically determines the value of these Libor-based instruments, we price less-liquid instruments by fitting a *yield curve*—any object from which P and f can be computed—to the observed values of the most liquid *build instruments*. Since there will not be above a few dozen such instruments, this fitting problem is severely underconstrained.

One common method is *bootstrapping of zero yields*: we specify that the yield curve will be defined by linear interpolation on the zero-coupon bond yield $y(T) \equiv -\ln P(T)/T$. This restricts the curve's degrees of freedom to one per interpolation point. If we place one interpolation point at the last maturity date (the latest payment or rate end date) of each build instrument, we can solve for each corresponding value of y with a succession of one-dimensional root searches.

Since $f(T) = y(T) + Ty'(T)$, the forward curve thus constructed is gratuitously discontinuous and contains large-scale interpolation artifacts. We do not wish to recommend this construction method or to disparage others, but only wish to note its frequent use and to show a concrete example. For a more complete discussion, see **Yield Curve Construction**.

Hedging on the Yield Curve

Once a yield curve is built, it can be used to price similar trades that are not among the build instruments, such as forward-starting or nonstandard swaps. Such pricing depends on two implicit assumptions: that the yield curve is the underlying of these trades as well as of its build instruments, and that its interpolation methods (or other nonmarket constraints) are sufficiently accurate. In practice, the former is widely accepted for Libor-based products, while the latter is a major arena of competition among market makers.

Any trade priced on the yield curve will have a *forward rate risk*, which we denote by $\delta_f(T)$, so that its change in value for a small curve fluctuation Δ_f is

equal to the change in $\delta_f \Delta f \equiv \int_0^\infty \delta_f(u) \Delta f(u) du$. Formally, we write

$$\delta_f(t) \equiv \lim_{h \rightarrow 0, \epsilon \rightarrow 0} \frac{U(f + h\eta_{t,\epsilon}) - U(f)}{h\epsilon} \quad (1)$$

where

- $U(f)$ is the trade's value for a given yield curve described by the forward rates f ;
- $\eta(z)$ is a C^∞ test function with support in $[0, 1]$ and $\int_0^1 \eta = 1$; and
- $\eta_{t,\epsilon}(z) \equiv \eta((z - t)/\epsilon)$.

For the *linear* trades, that we have so far discussed, δ_f will change very little as f changes. A portfolio of trades with no net δ_f has, at least for that moment, no interest rate risk.

Figure 1 shows the forward rate risk for a swap. The large-scale behavior is unsurprising; the forward rate risk steadily decreases as coupons are paid. The small-scale spikes are caused by overlapping, or in one case underlapping, of the start and end dates for the Libor rates on the floating side. The vertical scale is, of course, proportional to the swap notional amount, and is not shown here.

In practice, especially when trades cannot be exactly represented by equivalent cash flows, we will not know δ_f exactly but we will have rather a numerically computed (e.g., piecewise constant) approximation thereto; but since we can control the

buckets, that is, the intervals over which δ_f is kept constant, this is not a major difficulty.

Response Functions

However, we cannot execute a hedge of the forward rate risk directly; instead, we must choose a set of *hedge instruments* that will allow us to offset it. Often, these hedge instruments are exactly the build instruments. Each hedge instrument will also, of course, have a forward rate sensitivity. In practice, we generally consider the sensitivity, not of the instrument value, but of the *implied par rate* (or just *implied rate*): The implied FRA rate for futures, par coupon for swaps, or yield for bonds implied by the yield curve.

In this case, we can compute a hedge by slightly bumping each instrument's implied rate r_i , rebuilding the curve, repricing the trade being hedged, and measuring its price change. This method has the advantage of enabling very precise p/l explanation, at the cost of requiring repeated yield curve builds.

The resulting *instrument sensitivity* is closely related to the forward rate risk. To be precise, let the *response function* $\beta_i(T) \equiv dF(T)/dr_i$. Then, the instrument sensitivity is exactly $\beta_i \delta_f$. Thus, response functions provide an ideal tool for examining curve build methods.

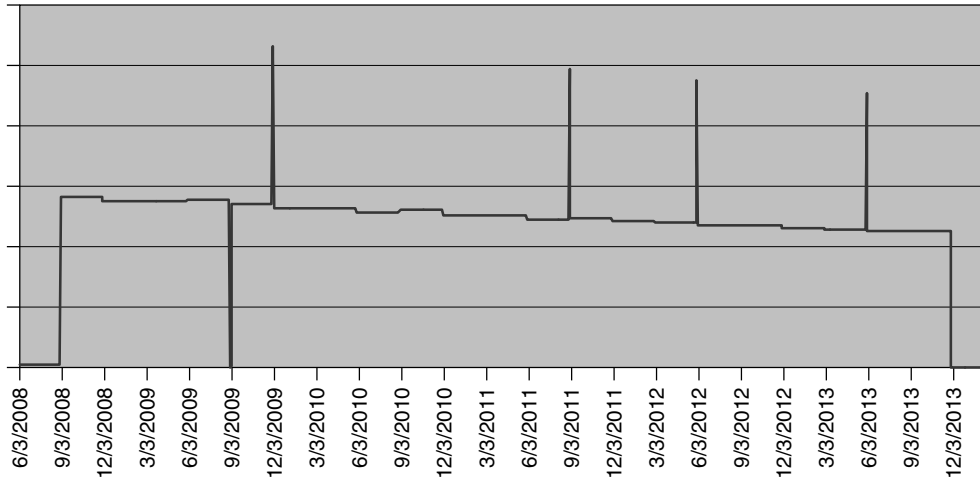


Figure 1 Forward rate risk for a (Payer) swap

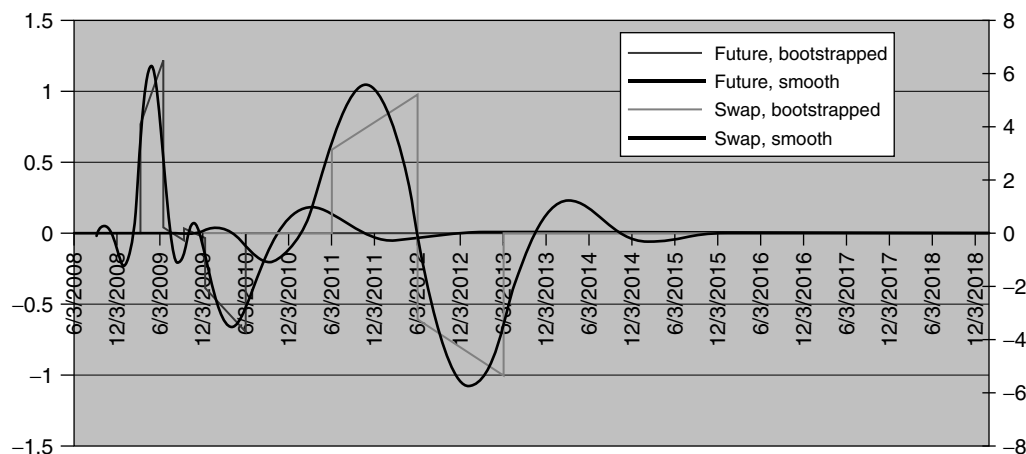


Figure 2 Response of F to fourth future and to 4-year swap

The response functions for two typical build instruments are displayed in Figure 2, for two different curve build methods. The response to a futures rate, shown against the left-hand scale, changes the forwards within the futures period and decreases them in the interval from the last future to the first swap (so that all other build instruments will have unchanged rates); naturally, within the futures period, $dF(T)/dt_i \simeq 1$. The response to a swap rate, which is substantially larger, is shown against the right-hand scale. In both cases, the bootstrapped curve shows the “sawteeth” characteristic of linear interpolation on y , while the smooth curve shows the inevitable loss of locality. This tension between smoothness and locality arises because a smooth curve, by its very nature, must alter values far from the source of a change in order to preserve smoothness; for details see **Yield Curve Construction**.

Bucket Delta Methods

Another common hedge method is to set the bucket end dates to the maturities of the curve build instruments, and then compute the *bucket deltas*: sensitivities $\bar{\delta}_k$ to the forward rate in the k th bucket, computed by applying a parallel shift to those forward rates. We can also define a Jacobian matrix J such that J_{ik} is the sensitivity of the i th instrument’s implied rate to the forward rate in the k th bucket; then the instrument sensitivities are given by $J^{-1}\bar{\delta}$. This is known as the *inverse method*.

A hedge can also be constructed from $\bar{\delta}$ by minimizing the p/l variance of the hedge trade plus a portfolio of hedging instruments; for this we need an estimate of the covariance $\Sigma(T_k, T_m)$ between the forward rates $f(T_k)$ and $f(T_m)$. The variance is then a quadratic form in the hedge instrument notionals, which can easily be minimized. Any other quadratic form, such as a penalty function based on the hedge notionals, can be included without difficulty.

Nonlinear Products

For any product, the sensitivity δ_f is defined in each yield curve state; however, it need not be independent of that state. This nonlinearity is most pronounced for options, especially when they are short-dated and nearly at-the-money. In this case, to lock in an option value by hedging we must dynamically rebalance the hedging instruments, subject to the well-known limitations of payoff replication strategies.

One issue of particular importance is that the local hedge, based on δ_t in the current state of the yield curve, can differ greatly from the variance-minimizing hedge if the rebalancing frequency is finite or if jumps are present. This occurs when the distribution of possible curve shifts is strongly asymmetric, or more frequently when the second derivative of the payoff is highly state-dependent.

These issues are not unique to interest rates, but they can become more pronounced for some payoffs owing to the tendency of short rates to move in

discrete increments (e.g., in response to a central bank action). Nonlinearity is also important to the pricing and hedging of very long-dated swaps and bonds.

Projected Gamma

The second-order “gamma” risk is formally described by an extension of the first-order forward rate risk that we have just discussed; in fact, there is a power series

$$U(f) = U(f_0) + \delta_f \Delta_f + \frac{1}{2} \int \int \gamma_f(u, v) \Delta(u) \Delta(v) du dv + \dots \quad (2)$$

where $\Delta_f(t) \equiv f(t) - f_0(t)$ and δ_f , γ_f , and so on, are computed at $f = f_0$. However, extracting this γ_f in all its detail is prohibitively time-consuming. Instead, we compute the f -dependence of δ_f , which is the same to this order (and is, in fact, the more relevant measure), by computing the delta hedge in various interest rate scenarios.

These scenarios are generally constructed by parallel shifts to the market yield curve. There are several reasons for this:

- Only parallel-shift gamma can realistically be hedged (by trading short-dated options).
- Large unexpected moves in rates, necessitating delta rehedging, tend to be roughly parallel.
- The parallel shift is uniquely easy to define and to explain.

The industry standard is to create a progression of scenarios using parallel shifts, which are multiples of 10 basis points, for example, from -50 to $+50$ in steps of 10. This aids visualization of higher-order contributions to the hedge, which can sometimes be traced to their source; for example, a sudden change in delta indicates an upcoming option expiry or barrier test in that interval.

Vega Hedging

The volatility sensitivity, or “vega”, of an interest rate derivative is characterized by both the time during which its optionality is active, and the time

covered by the underlying rate. In other words, nonlinear interest rate derivatives are sensitive to the Heath–Jarrow–Morton (HJM)^b *forward volatility* $\sigma(t, T)$, the volatility at time t of $f(T)$. This does not mean that we must use an HJM model: any interest rate model defines an expected forward volatility.

A given trade, then, has a “forward volatility footprint” $v(t, T)$ so that its change in value, to first order in volatility, equals the change in $\int \int v(t, T) \sigma(t, T) dt dT$. In more sophisticated models, this change is computed with other parameters (e.g., correlations, elasticity of rates, or volatility of volatility) held fixed. Thus, $v(t, T)$ is precisely analogous to the forward rate sensitivity $\delta_f(T)$.

In addition, any mechanistic calibration technique has a response function

$$B_j(t, T) \equiv \frac{\partial \sigma(t, T)}{\partial \sigma_j} \quad (3)$$

where σ_j is the quoted volatility of the j th calibration instrument and partial derivatives are taken with other calibration instruments fixed. The inner product $\int \int v(t, T) B_j(t, T) dt dT$ is thus the *calibration instrument vega*, which will be computed by bumping, recalibrating, and repricing (“bump-and-grind”). It is worth noting that examination of response functions is by far the best test of the quality of a calibration technique.

Models with few state variables harshly restrict the form of $\sigma(t, T)$; for example, in one-state-variable models we have $\sigma(t, T) = g(t)H(T)$.^c This impacts vega hedging in two main ways. First, the forward vol footprint simply cannot be measured using such a model, since it is the sensitivity to a perturbation which the model cannot reproduce. Second, the response functions must reflect the constraints on $\sigma(t, T)$; thus, they are inevitably nonlocal and often highly unnatural. Such models can be used for vega hedging in restricted environments, such as when the maturity of the hedging instruments is already known; but, in general, they are incapable of “finding” the vega hedge, owing to their intrinsic inability to localize perturbations.

Cost of Funding Complications

So far, we have assumed that the market Libor rate reflects our own cost of funds; this assumption is necessary for a single zero curve to exist. In

practice, many market participants consistently fund their operations above or below Libor; thus they must maintain a zero curve Z_c for their own cashflows, and a different curve Z_r for rate forecasts. To a very good approximation, this effect can be encapsulated by specifying the *funding adjustment* $Z_c(T)/Z_r(T)$.

This is most important in markets where the demand for currencies is highly asymmetric, particularly in Japanese Yen (*JPY*). The steady demand of *JPY*-based issuers for dollars means that foreign dealers can fund their *JPY* debts more cheaply, effectively increasing their own Z_c —this effect drives the currency basis swaps market.

The presence of funding adjustments causes different dealers, observing the same market swap rates, to deduce different forward curves. The effect is always to create an incentive for the party with the higher cost of funds to take the side with initial positive cash flows, thus borrowing a fraction of the swap notional. To minimize this effect, most market quotes are for mutually collateralized swaps.

Cost of Tenor Complications

The frequency of Libor fixings also influences the par rate for a swap; the curve Z_r , which forecasts 3-month Libor will not forecast 1-month or 6-month Libor. In practice, the forward rates are higher for longer-tenor Libor rates; this is generally attributed to credit issues, since a longer tenor entails a greater risk of a downgrade during the loan period.

This can be addressed by constructing separate curves for each tenor, or by adjusting Libor rate forecasts by some *ad hoc* tenor-dependent correction. The floating-for-floating swaps at mismatched tenors, in which these corrections can be traded, are also called *basis swaps*.

Collateral and Repo Complications

For some markets, notably US Treasury bonds, the value of a bond is not fully captured by the value of its cashflows. The complex repo market for these bonds makes some *special* bonds valuable as collateral, giving them a positive convenience yield

and raising their price above that predicted from the zero curve. Such markets are usually treated by building a *general collateral* curve based on bonds with no such added value, and then relating the premium in some bond's price to its expected repo rates. A derivative whose underlying is a special bond is thus exposed to both general collateral rates and the bond's forecast repo rates. For short-expiry trades, this combination of exposures is accurately expressed as an exposure to the special bond's price, but long-dated trades require separate consideration of the two curves.

End Notes

^aEven this simple contract is fraught with credit and collateral issues, which our notation conceals. Most of these are beyond the scope of this discussion.

^bFor details of the HJM approach, see Bjork **Heath–Jarrow–Morton Approach**.

^cHere $\sigma(t, T)$ need not be the normal HJM volatility; in Black–Karasinski models, for example, it is the volatility of the log forward rate in the risk-neutral measure which is separable.

Further Reading

- Romanelli, P. (1997). The yield curve: from the ground up, Bankers Trust Topics in Derivatives Analytics 2.
- Hagan, P. & West, G. (2006). Interpolation methods for yield curve construction, *Applied Mathematical Finance* 3(2), 89–129.
- Heath, D., Jarrow, R. & Morton, A. (1989). *Bond pricing and the term structure of interest rates: a new methodology*, Cornell University, working paper.
- Pennacchi, G., Ritchken, P. & Sankarasubramanian, L. (1996). On pricing kernels and finite state-variable Heath, Jarrow, Morton models, *Review of Derivatives Research* 1, 87–99.

Related Articles

Delta Hedging; Gamma Hedging; Hedging; Mean–Variance Hedging; Yield Curve Construction.

TOM HYER

Inflation Derivatives

The market for financial inflation products started with public sector bonds linked to some measure for inflation of prices of (mainly) goods and services. This dates back to as early as the first half of the eighteenth century when the state of Massachusetts issued bonds linked to the price of silver on the London Exchange [4]. Over time, and particularly, in the last 20 years or so, the dominant index used for inflation-linked bonds has become the consumer price index (CPI). A notable exception is the UK inflation-indexed gilt market, which is linked to the retail price index (RPI).^a The actual cash-flow structure of inflation-indexed bonds varies from issue to issue, including capital-indexed bonds (CIBs), *interest-indexed bonds*, *current pay bonds*, *indexed annuity bonds*, *indexed zero-coupon bonds*, and others. By far the most common cash-flow structure is the CIB, on which we shall focus in the remainder of this article.

Bonds, Asset Swaps, and the Breakeven Curve

Inflation-indexed bonds (of CIB type) are defined by

- N —a notional;
- I —the inflation index;
- L —a lag (often three months);
- $T_i : \{1 \leq i \leq n\}$ —coupon dates;
- $c_i : \{1 \leq i \leq n\}$ —coupon at date T_i (usually, all c_i are equal); and
- $I(T_0 - L)$ —the bond's base index value.

The bond pays the regular coupon payments

$$N c_i \frac{I(T_i - L)}{I(T_0 - L)} \quad (1)$$

plus the inflation-adjusted final redemption, which often contains a capital guarantee according to

$$N \max \left(\frac{I(T_n - L)}{I(T_0 - L)}, 1 \right) \quad (2)$$

Asset swap packages swapping the inflation bond for a floating leg are liquid in some markets such as for bonds linked to the CPTFEMU (also known as the *HICP*) index. Since the present value of an inflation-linked bond can be decomposed into the value of each coupon

$$P_{T_i}(T_0) N c_i \frac{1}{I(T_0 - L)} \mathbb{E}_0^{\mathcal{M}(P_{T_i})} [I(T_i - L)] \quad (3)$$

wherein $\mathbb{E}_\tau^{\mathcal{M}(X)}[\cdot]$ denotes expectation in filtration $\mathcal{F}_\tau$ under the measure induced by choosing X as numéraire, and the value of the final redemption

$$P_{T_n}(T_0) N \frac{1}{I(T_0 - L)} \times \mathbb{E}_0^{\mathcal{M}(P_{T_n})} [\max(I(T_n - L), I(T_0 - L))] \quad (4)$$

these products give us a mechanism to calibrate the *forward curve* $F(t, T)$, where

$$\begin{aligned} F(t, T - L) &:= \text{The index forward for (payment)} \\ &\quad \text{time } T \text{ seen at time } t \\ &:= \mathbb{E}_t^{\mathcal{M}(P_T)} [I(T - L)] \end{aligned} \quad (5)$$

The forward curve is often also referred to as the *breakeven curve*. The realized inflation index fixing level is thus naturally $I(T) = F(T, T)$. Note that while equation (4), strictly speaking, requires a stochastic model for consistent evaluation due to the convexity of the $\max(\cdot, 1)$ function, in practice, the $\max(\cdot, 1)$ part is usually ignored, since its influence on valuation is below the level of price resolution.^b

If there were a multitude of inflation-linked bonds, or associated asset swaps, with well-dispersed coupon dates liquidly available for any given inflation index, then the above argument would be all that is needed for the construction of a forward index curve, that is, breakeven inflation curve. In reality, though, for many inflation markets, there is only a small number of reasonably liquid bonds or asset swaps available. This makes it necessary to use interpolation techniques for forward inflation levels or rates between the attainable index-linked bond's maturity dates. In some cases, this may mean that for the construction of a 10-year (or longer) inflation curve, only three bonds are available, and extreme care must be taken for the choice of interpolation. However, even when a sufficiently large number of bonds is traded, to have a forward inflation rate for each year determined by the bond market, sophisticated interpolation methods are still needed. This is because of inflation's *seasonal* nature. For instance, consumer prices tend to go up significantly more than the annual average just before Christmas and tend to drop (or rise less than the annual average) just after.

The most common approach to incorporate seasonality into the breakeven curve is to analyze the statistical deviation of the month-on-month inflation

rate from the annual average with the aid of historical data, and to overlay a seasonality adjustment on top of an annual inflation average curve in a manner such that, by construction, the annual inflation index growth is preserved. In addition, some authors used to suggest that one may want to add a long-term attenuation function (such as $e^{-\lambda t}$) for the magnitude of seasonality. This was supposed to represent the view that, since we have very little knowledge about long-term inflation seasonality, one may not wish to forecast any seasonality structure. This idea has gone out of fashion though, probably partly based on the realization that, historically, the seasonality of inflation became *more* pronounced over time, not exponentially *less*.

Inflation Derivatives

Daily Inflation Reference

Ultimately, all inflation fixings are based on the publication of a respective index level by a government or cross-government funded organization such as Eurostat for the HICP index series in Europe, or the Bureau of Labor Statistics for the CPI-U in the United States. This publication tends to be monthly, and usually on the same day of the month with a small amount of variability in the publication date. In most inflation bond markets, index-linked bonds are written on the publication of these published index levels in a straightforward manner such as $I(T_i)/I(T_0)$ times a fixed number as discussed in the previous section, with T_i indicating that a certain month's publication level is to be used. For some inflation bonds, however, the inflation reference level is not a single month's published fixing, but instead, an average over the two nearest fixings. In this manner, the fact that a bond's coupon is possibly paid between two index publication dates, and thus should really benefit from a value between the two levels, can be catered for. French OAT i and OAT ϵi bonds, for instance, use this concept of the daily inflation reference (DIR) defined as follows:

$$\begin{aligned} \text{DIR}(T) = & I(T_{m(T)-3}) + \frac{n_{\text{day}}(T) - 1}{n_{\text{days}}(m(T))} \\ & \times [I(T_{m(T)-2}) - I(T_{m(T)-3})] \end{aligned} \quad (6)$$

with $m(T)$ indicating the month in which the reference date T lies, T_i the publication date of month

i , $n_{\text{day}}(T)$ the number of the day of date T in its month, and $n_{\text{days}}(m(T))$ the number of days in the month in which T lies. For example, the DIR applicable to June 21st is $\frac{10}{30}$ times the HICP for March plus $\frac{20}{30}$ times the HICP for April. While the DIR is in itself not an inflation derivative, it is a common building block for derivatives in any market that uses the DIR in any bond coupon definitions. The DIR clause complicates the use of any model that renders inflation index levels as lognormal or similar, since any payoff depending on the DIR thus depends on the weighted sum of two index fixings.

Futures

Futures on the Eurozone HICP and the US CPI have been traded on the Chicago Mercantile Exchange since February 2004. Eurex launched Euro inflation futures based on the Eurozone HICP in January 2008. Both exchanges show, to date, very little actual trading activity in these contracts. An inflation futures contract settles at maturity at

$$M \left(1 - \frac{1}{\Omega} \left(\frac{I(T-L)}{I(T-L-\Omega)} - 1 \right) \right) \quad (7)$$

with M being a contract size multiplier and Ω an additional time offset. The lag L is usually one month. The offset Ω is three months for CPI-U (also known as CPURNSA) on the CME, that is, $\Omega = 1/4$ above. For the HICP (also known as CPT-FEMU) on both Eurex and CME, $\Omega = 1$, that is, one year. Exactly why the inflation trading community has paid little attention to these futures is not entirely clear, though, one explanation may be the difference in inflation linkage between bonds and futures. Both HICP-linked bonds and US Treasury-Inflation Protected Securities (TIPS) pay coupons on an inflation-adjusted notional, that is, they are CIBs. In contrast, both CPI and HICP futures pay period-on-period inflation rates. As a consequence, a futures-based inflation hedge of a single CIB coupon would require a whole sequence of futures positions and would leave the position still exposed to realized period-on-period covariance.

Zero-coupon Swaps

This simple inflation derivative is as straightforward as the simplest derivative in other asset classes: two

counterparties promise to exchange cash at an agreed date in the future. One counterparty pays a previously fixed lump sum, and the other counterparty pays an amount that is given by the published fixing of an agreed official inflation index (times an agreed fixed number). In the world of equity derivatives, this would be called a *forward contract* on the index. Since inflation trading is, by nature and origin, leaning on fixed income markets, this contract is referred to as a *swap*, which is the fixed income market equivalent of a forward contract. Conventionally, swaps are defined by repeated exchange of fixed for floating coupons. Since the inflation index forward contract has no cashflows during the life of the contract, and only at maturity money is exchanged, in analogy to the concept of a *zero-coupon bond*, the inflation index forward contract is commonly known as the *inflation index-linked zero-coupon swap*, or just *zero-coupon swap*^c for short. More precisely speaking, zero-coupon swaps have two legs. The two legs of the swap pay

$$\begin{aligned} \text{inflation leg: } & N \frac{I(T-L)}{I(T_0-L)} \\ \text{fixed leg: } & N(1+K)^{T-T_0} \end{aligned} \quad (8)$$

where N is the notional, T_0 is the start date of the swap, T the maturity, and K the quote. These swaps appear in the market comparatively liquidly as hedges for the inflation bond exposure to the final redemption payment—they act as a mirror. However, this should not mask the fact that the true source of the liquidity is the underlying asset swap/inflation bond.

The Vanilla Option Market

In the inflation option market, the most liquid instruments are as follows.

- **Zero-coupon caps and floors:**

At maturity T , they pay

$$\left[\phi \left(\frac{I(T-L)}{I(T-L-\Omega)} - (1+K)^\tau \right) \right]_+ \quad (9)$$

where Ω is the index offset, L is the index lag, K the annualized strike, τ the year fraction between $(T-L-\Omega)$ and $(T-L)$, and ϕ is +1 for a cap and -1 for a floor. For most of those options, the index in the denominator is actually known (i.e., $\Omega = T$), and

the option premium depends only on the volatility of the index $I(T-L)$.

- **Year-on-year caps and floors:**

The option is a string of year-on-year caplets or floorlets individually paying according to equation (9) with $\Omega = \tau = 1$ at increasing dates spaced by one year. Apart from the front period, these caplets and/or floorlets thus depend on one index forward in their payoff's numerator, and on a second one in their payoff's denominator, whence some authors refer to these products being subject to *convexity*, though this is not to be confused with the usual concept of convexity induced by correlation with interest rates (there is more on this in the section Two Types of Convexity). An alternative view of a year-on-year caplet/floorlet's volatility dependence is to consider *volatility of the year-on-year ratio* as the fundamental driver of uncertainty. In this framework, no convexity considerations are required. The payoff of an inflation caplet/floorlet resembles the payoff of a vanilla option on the return of a money market account if one replaces inflation rates with interest rates.

The Swap Market

A close cousin of inflation caps and floors is the inflation swap. The swap consists of a series of short-tenored *forward starting zero-coupon swaps* each of which pays

$$\frac{I(T-L)}{I(T-L-\Omega)} - (1+K)^\tau \quad (10)$$

at the end of its respective period. Just like an interest rate swap can be seen as a string of *forward interest rate agreements*, an inflation swap can be seen as a string of *forward inflation rate agreements*. Unlike vanilla interest rate swaps, though, the period an inflation swap's individual forward inflation rate agreement is linked to does not have to be equal to the period the associated coupon is nominally associated with. An example for this is an asset swap on an Australian (inflation-linked) government bond whose coupons are typically paid quarterly and are “indexed to the average percentage change in the CPI over the two quarters ending in the quarter that is two quarters prior to that in which the next interest payment falls” [4]. In other words, Ω is six months, $\tau = 1/2$, and L is in the range of one to three months for quarterly coupons.

Total Return Inflation Swaps

These structures pay out a fixed sum linked to an inflation measure over time. With growing inflation concerns in mid-2008, these, together with inflation caps, became increasingly popular with private as well as institutional investors.

The Limited Price Index

A limited price index (LPI) is an instrument that is used in the market to provide a hedge to inflation, but with limited upside and/or downside exposure. When period-on-period inflation is within an agreed range, the LPI grows at the same rate as its underlying official publication index. When period-on-period inflation is outside this range, the LPI grows at the, respectively, capped or floored rate. Given an underlying inflation index I , the LPI $\tilde{I}$ is constructed using I , a base date T_b , a fixing time tenor (i.e., frequency period) τ , an inflation capping level $l_{\max} \in (0, \infty]$, and an inflation flooring level $l_{\min} \in [0, \infty)$. Using T_b and τ , we create a publication sequence for $\tilde{I}$:

$$\tilde{I}_{\text{publication sequence}} = \{T_b, T_b + \tau, T_b + 2\tau, T_b + 3\tau, \dots\} \quad (11)$$

The LPI $\tilde{I}$ can be defined recursively, starting with $\tilde{I}(T_b) = 1$, and continuing with

$$\tilde{I}(T_b + (i+1)\tau) = \tilde{I}(T_b + i\tau) \begin{cases} 1 + l_{\min} & \text{if } \frac{I(T_b + (i+1)\tau)}{I(T_b + i\tau)} \leq 1 + l_{\min} \\ \frac{I(T_b + (i+1)\tau)}{I(T_b + i\tau)} & \text{if } 1 + l_{\min} \leq \frac{I(T_b + (i+1)\tau)}{I(T_b + i\tau)} \leq 1 + l_{\max} \\ 1 + l_{\max} & \text{if } 1 + l_{\max} \leq \frac{I(T_b + (i+1)\tau)}{I(T_b + i\tau)} \end{cases} \quad (12)$$

Given the definition of the LPI, derivatives such as zero-coupon swaps and options on the LPI can be built. LPI-linked products are most common in the UK market. A common comparison made by trading and structuring practitioners is to view an LPI, especially if it is only one sided, as very similar to an inflation cap or floor. It is worth noting that the two structures do have different sensitivities to inflation curve autocorrelation assumptions

whence an exact static replication argument based on inflation caps and floors is not attainable. As a consequence, LPIs require a fully fledged (stochastic) model for relative value comparison with inflation bonds, zero-coupon swaps, and inflation caps and floors.

The Inflation Swaption

This product is in its optionality very similar to a conventional interest rate swaption. However, the underlying swap can have a variety of special features. The start date of the swap T_{start} is on or after the expiry of the option. The swap has one inflation-rate-linked leg, and one (nominal) interest-rate-linked leg. For instance, the underlying swap could be agreed to be given by the following.

- The inflation leg is a sum of n annual forward inflation rate agreements, paid at dates $T_1, T_2, \dots, T_{n-1}, T_n$ with T_n representing the final maturity of the underlying swap. Each inflation leg coupon pays $k^{\text{DIR}(T_i)} / I_{\text{Base}}$, with $\text{DIR}(T_i)$ being the daily inflation reference for date T_i . The leg also pays a final notional redemption floored at 0, which can of course be seen as an enlarged coupon plus a zero-coupon floor struck at 0.
- The Libor leg is a sequence of quarterly forward (nominal) interest rate agreements. A variation

of this is for the interest rate leg to pay Libor plus a fixed margin, or Libor minus a fixed margin floored at zero, which makes this leg alone already effectively a nominal interest rate cap.

There are many variations of inflation-linked swaptions, such as *swaptions on the real rate* (the real rate is explained further on), but most of them are only traded over-the-counter and in moderate size.

Inflation Modeling

For vanilla products such as inflation zero-coupon caps and floors, and year-on-year caplets and floorlets, practitioners tend to use terminologies and conventions relating to the Black and Bachelier models. For instance, to translate the price of an option that pays $(I_T - K)_+$, that is, simply at some point in the future the level of the index minus a previously agreed fixed number, or zero, whichever is greater, practitioners tend to use the concept of implied volatility for the index as if the index evolved in a Black–Scholes modeling framework. For options that pay a year-on-year rate, either Black implied log-normal volatilities or, alternatively, Bachelier, that is, absolute normal volatilities may be used. The latter are also sometimes referred to as *basis points volatilities* to indicate the fact that these are expressed in absolute, rather than relative terms (as Black volatilities would).

Jarrow–Yildirim

Most articles on inflation modeling start out with what is referred to as *the Fisher equation*. This is the definition that *real returns* are given by the relative increase in *buying power* on an investment, not by nominal returns [5]. In other words, the Fisher equation is the *definition*

$$r_{\text{real}} = r - y \quad (13)$$

or

$$r = r_{\text{real}} + y \quad (14)$$

with r being the continuously compounded nominal rate, y the continuously compounded inflation rate, and r_{real} represents a thus defined *real rate*. While this definition is, in practice, of no further consequence to any derivatives modeling, the terminology and concepts are useful to know since they pervade the inflation modeling literature.

One of the earliest publications suggesting a comprehensive set of dynamics for the evolution of inflation (rates) in relation to nominal interest rates is the Jarrow–Yildirim model [8]. The original article discusses the generic setting of a *real economy*, a *nominal economy*, and a translation index between the two, employing the mathematical no-arbitrage HJM apparatus [6]. This results in a framework that is completely analogous to a foreign exchange rate

model with foreign (the *real economy's*) interest rate dynamics, domestic (the *nominal economy's*) interest rate dynamics, and an exchange rate that represents the inflation index. A confusing aspect for people not used to inflation financials is the nomenclature to refer to observable real-world interest rates as *nominal* rates, and to nonobservable (derived) inflation-adjusted return rates (i.e., to nominal interest rates minus inflation rates) as *real* rates. In practice, this model tends to be implemented with both individual economies' interest rate dynamics being given by an extended Vasicek [12] ([7], but also see **Term Structure Models** and **Gaussian Interest-Rate Models**) model, with locally geometric Brownian motion for the real/nominal FX rate, that is, the inflation index. In short, in the nominal money market measure, the dynamics of nominal zero-coupon bonds $P_T(t)$, real zero-coupon bonds $P_{\text{real},T}(t)$, and the inflation index $X(t)$ are governed by the stochastic differential equations:

$$\frac{dP_T(t)}{P_T(t)} = r(t) dt + \sigma_P(t, T) dW_P(t) \quad (15)$$

$$\begin{aligned} \frac{dP_{\text{real},T}(t)}{P_{\text{real},T}(t)} = & (r_{\text{real}}(t) - \rho_{P_{\text{real}},X} \sigma_X(t) \sigma_{P_{\text{real}}}(t, T)) dt \\ & + \sigma_{P_{\text{real}}}(t, T) dW_{P_{\text{real}}}(t) \end{aligned} \quad (16)$$

$$\frac{dX(t)}{X(t)} = (r(t) - r_{\text{real}}(t)) dt + \sigma_X(t) dW_X(t) \quad (17)$$

with

$$\sigma_P(t, T) = \int_t^T \zeta(s) e^{-\int_t^s \kappa(u) du} ds \quad (18)$$

$$\sigma_{P_{\text{real}}}(t, T) = \int_t^T \zeta_{\text{real}}(s) e^{-\int_t^s \kappa_{\text{real}}(u) du} ds \quad (19)$$

The Jarrow–Yildirim model, initially, gained significant popularity with quantitative analysts, though, arguably, primarily because this model could be deployed for inflation derivatives comparatively rapidly since it was already available as a cross-currency Hull–White model in the respective practitioners' group's analytics library. Its main drawbacks are that calibration and operation of the model requires specification of a set of correlation numbers between inflation index and (nominal) interest rates,

inflation index and real rates, and (nominal) interest rates and real rates, of which only the first one is directly observable. Also, this model requires volatility specifications for nonobservable (and nontradable) real rates.

Inflation Forward “Market” Modeling

In an alternative model suggested by Belgrade *et al.* [2], inflation index forward levels $F(t, T_i)$ for a discrete set of maturities $\{T_i\}$ are permitted to evolve according to

$$\frac{dF(t, T_i)}{F(t, T_i)} = \mu(t, T_i) dt + \sigma(t, T_i) dW_i(t) \quad (20)$$

The drift terms $\mu(t, T_i)$ are determined as usual by no-arbitrage conditions, but the volatility functions $\sigma(t, T_i)$ can be chosen freely, as can the correlations between all the Wiener processes W_i . The authors discuss a variety of different possible choices. When interest rates are deterministic, or assumed to be uncorrelated with forward inflation index levels, all drift terms vanish and the model reduces to a set of forward levels that evolve as a multivariate geometric Brownian motion. The main drawback of this model is that it permits highly undesirable forward inflation rate dynamics, specifically with respect to the breakeven curve’s autocorrelation structure, unless complicated modifications of the instantaneous correlation and volatility functions are added. The model also introduces many free parameters that have to be calibrated, which is not ideal in a market that is as illiquid as the inflation option market.

The Exponential Mean-reverting Period-on-period Model

When a model’s purpose is predominantly for the pricing and hedging of contracts that only depend on year-on-year returns of the inflation index, a useful model is to consider the year-on-year ratio process:

$$Y(t) = F_Y(t) e^{-\frac{1}{2} V_0[x(t)] + x(t)} \quad (21)$$

$$dx(t) = -\kappa x(t) dt + \alpha(t) dW(t) \quad (22)$$

with $V_0[x(t)]$ representing the variance of the inflation driver process x from time 0 to time t , and the deterministic function $F_Y(t)$ is implicitly defined by

the choice of measure, that is, it is calibrated such that the model reproduces the breakeven curve (and thus all inflation-linked bonds) correctly. This model is known as the *exponential mean-reverting year-on-year* model. Note that even though the year-on-year ratio process is formulated as a continuous process, it is really only monitored at annual intervals when relevant fixings become available, that is,

$$Y(T_i) = \frac{I(T_i)}{I(T_i - \Omega)} \quad (23)$$

The advantage of this model is its comparative simplicity and its reasonable inflation curve autocorrelation characteristics. This model can also be combined with simple interest rate dynamics (such as given by the Hull–White model). A further benefit is that it promotes the view of a future inflation index level fixing being the result of year-on-year return factors:

$$I(T_n) = I(T_0) \prod_{i=1}^n Y(T_i) \quad (24)$$

The sequence of $Y(T_i)$ fixings are a set of strongly positively correlated lognormal variables (even when interest rates are stochastic in a Hull–White setting) in analogy to the correlation structure of a set of forward starting zero coupon bonds in a standard Hull–White model. As such, they lend themselves very well to an approximation of independence conditional on one or two common factors [1, 3]. Conditional independence permits easy evaluation of the LPI [11]. For other inflation products, the model can also be written on shorter period-on-period returns and can be equipped with more than one inflation driver: $x(t) \rightarrow x_1(t) + x_2(t)$, and so on.

The Multifactor Instantaneous Inflation Model with Stochastic Interest Rates

For more complex products depending on more than just the period-on-period autocorrelation structure, the exponential period-on-period model with Hull–White interest rate dynamics can be taken to the limit of infinitesimal return periods, which makes it structurally similar to the Jarrow–Yildirim model, with the main difference being that we model inflation rates directly, rather than real rates. Dropping all lag-induced convexity considerations for clarity (i.e.,

assuming $L = 0$ and no settlement delay), this gives us, for m inflation and one interest rate factor, in the T -forward measure,

data into unobservable “real rate” volatilities and autocorrelations.

$$I(t) = F(0, t) e^{-B_a(t, T) \text{Cov}_0 \left[\sum_{i=1}^m X_i(t), z(t) \right] - \frac{1}{2} \text{V}_0 \left[\sum_{i=1}^m X_i(t) \right] + \sum_{i=1}^m X_i(t)} \quad (25)$$

$$P_T(t) = \frac{P_T(0)}{P_i(0)} e^{-B_a(t, T) \left(z(t) - \frac{1}{2} B_a(t, T) \text{V}_0[z(t)] \right)} \quad (26)$$

$$z(t) = \int_0^t e^{-a(t-u)} \sigma(u) dW^z(u) \quad (27)$$

$$x_i(t) = \int_0^t e^{-\kappa_i(t-u)} \alpha_i(u) dW^i(u) \quad (28)$$

$$X_i(t) = \int_0^t x_i(s) ds = \int_0^t B_{\kappa_i}(u, T) \alpha_i(u) dW^i(u) \quad (29)$$

$$B_\mu(t_1, t_2) = \int_0^{t_2-t_1} e^{-\mu s} ds \quad (30)$$

wherein all x_i and z are standard Ornstein–Uhlenbeck processes under the T -forward measure, of numéraire P_T , the zero-coupon bond paying in T , and we assume flat mean reversion a and κ_i . It is naturally straightforward to translate these dynamics to other measures (see, e.g., **Forward and Swap Measures**).

A suitable choice of the different drivers’ correlation, volatility, and mean reversion strength permits highly flexible calibration to the inflation curve’s desired autocorrelation structure, which is another major advantage over the Jarrow–Yildirim model. For the specific case when there are two inflation factors, and the mean reversion strength of one of the factors, say x_1 , is taken to the limit $\kappa_1 \rightarrow \infty$ while at the same time keeping α_1^2/κ_1 constant, the cumulative process X_1 becomes a Brownian motion.^d In this case, inflation has one geometric Brownian motion factor, and one mean-reverting inflation rate factor, in complete analogy to the Jarrow–Yildirim model. In this sense, the multifactor instantaneous inflation model with stochastic interest rates encompasses the Jarrow–Yildirim model’s autocorrelation structure as a special case, but, in general, allows for more flexible calibration, while, at the same time, it avoids the need to translate market observable

Common Inflation Modeling Considerations

Two Types of Convexity. In the inflation market, when practitioners talk about *convexity* effects, confusingly, and differently from the derivatives market of most other asset classes, they might be referring to one of the two possible effects whose origins are rather distinct.

The first of the two convexity effects arises in the context of period-on-period inflation rate related products such as year-on-year caps and floors (and obviously also inflation futures). In this context, *convexity* is generated only if the fundamental observable of the underlying inflation market is considered to be the *inflation index* since period-on-period payouts are always governed by the return factor

$$Y(T - L) = \frac{I(T - L)}{I(T - K - \Omega)} \quad (31)$$

which is *hyperbolic*, and thus convex, in the first of the two index fixings entering the return ratio $Y(T - L)$. For inflation derivatives practitioners whose background is firmly in the area of exotic interest rate derivatives, and who prefer to view period-on-period rates, that is, $(Y(T - L) - 1)/\Omega$ or similar, as the fundamental underlying, this convexity, quite naturally, does not exist, and is merely an artifact of

choice of fundamental variables. Given the fact that ultimately all inflation fixing data result in the publication of an index figure, that is, $I(T)$, and the impact this setting has on the mind-set of inflation investors, this *hyperbolicity* effect is worth bearing in mind.

The second of the two convexity effects is the same as can be observed in other asset classes due to correlation with interest rates. It arises from a *timing discrepancy* between the fixing (observation) time of a financial observable and the time at which a payoff based on it is actually paid out. For instance, if an inflation zero-coupon swap payment is to be made with a certain delay, any nonzero correlation with interest rates gives rise to a risk-neutral valuation difference. Intuitively, one can understand this without the use of any model by considering that, in any scenario of high inflation, assuming positive correlation with interest rates, a delay of the payment is likely to incur a higher-than-average discount factor attenuation since, in that scenario, interest rates are also likely to be high. As a consequence, *ab initio*, one would tend to value a delayed inflation payment lower than its nondelayed counterpart (if we assume positive correlation between inflation and interest rates).

Both the hyperbolicity effect's and the interest rate correlation induced convexity effect's valuation formulae depend of course on the specific model used but are, typically, in any tractable model, straightforward to derive.

Fixing Lag Effects. An inconvenient feature of the inflation market is that the prices of tradables, even in the simplest of all possible cases, do not, in general, converge to the level at which they finally settle when the last piece of relevant market information becomes available. Take, for instance, a forward contract on the one month return of a money market account with daily interest rate accrual at the overnight rate. Clearly, the amount of uncertainty in the settlement level of this forward contract decreases at least linearly as we approach the last day of the one month period. On the day before the final overnight rate becomes known, at worst, 1/30 (say) of the uncertainty is left as to where this contract will fix. In contrast, for a similar inflation rate contract, such as a month-on-month caplet or floorlet, the amount of uncertainty in the beginning of the one month period and at its end is very

similar. This is because only limited extra information as to where the inflation index will fix becomes available during the month. As a consequence, it is common practice to consider for all volatility and variance related calculations, the reference date, that is, the point in time assumed to be $t = 0$, to be the *last index publication date*. Arguably, this may or may not be seen as an inconsistency between the framework assumed within any used model and its application, but, in practice, this appears to be a usable compromise.

Another point worth noting for the implementation of models, which is rarely spoken about in publications on the subject, is that all fixing timing tends to be subject to lags, offsets, and so on. This seemingly innocuous feature can, in practice, turn out to require intricate attention to detail. Unfortunately, since inflation structures can have very long final maturities of up to one hundred years, these small differences, if systematically erroneous, can build up to major valuation differences and can thus not be ignored. A specific mistake that is exceptionally easy to make is to assume that a breakeven curve, traded and quoted as zero-coupon swap rates, relates to future inflation index fixing levels in their own T -forward measure, with T being the publication date of the respective index level. This is insidiously wrong since actual zero-coupon swaps pay at a date that is a given tenor after the day of inception, and this could be any day in the month. For instance, if we enter into a five-year zero-coupon swap on the first of a month, ignoring weekends, this will typically settle in five years on the third of the month, and pay the index level published (say) at the beginning of the month three months before the settlement month. If we enter into the same swap on the 26th of the month, it will settle on the 28th five years later, and pay the same index fixing level as the previous example. For an arbitrary zero-coupon swap with lag L , depending on the day of the month at which it is entered into, this means it may settle with an effective lag of L months, or almost $L + 1$ months, or anything in between. Any associated interest rate convexity considerations thus depend not only on time to expiry and nominal lag but also on the day of the month. This may not sound much for a short-maturity zero-coupon swap such as one or two years, but for a 50-year zero-coupon swap, the interest rate convexity incurred makes a prohibitively expensive difference!

Backward Induction for Interest Rate/Inflation Products. It is not uncommon for inflation products to contain elements of a more traditional interest rate derivatives nature as, for example, was mentioned for the comparatively benign inflation swaptions discussed in the section The Inflation Swaption. When more exotic products allow for easy valuation in a strictly forward looking manner, one can of course employ Monte Carlo simulation techniques. However, if a product such as an inflation swaption is equipped with a Bermudan callable feature, as a quantitative analyst, one faces the problem that interest rates are by their nature forward looking, whereas inflation rates are by their nature backward looking. What we mean is that, typically, a natural floating interest rate coupon's absolute value is known *at the beginning* of the period for which it is paid. In contrast, floating inflation coupons are always paid after their associated period has elapsed. In a conventional backwards induction implementation on any type of lattice (e.g., explicit finite differencing solver or tree), this poses the problem that, as one has rolled the valuation back to the beginning of an interest rate period, one has to compare the value of the so induced filtration specific spot Libor rate, in some kind of pay-off formula, with the contemporaneous inflation rate, which is due to the inflation driver's evolution *in the past*, and thus *unknown* on the backwards induction lattice. The problem is ultimately very similar to the valuation of, say, a forward starting equity option. The paradox is that one needs to have simultaneous access to the effect of the interest rate driver's evolution over a future interval *and* to the inflation rate driver's evolution in the past, on any valuation node in the lattice. A practical solution to this dilemma is in fact very similar to the way in which the valuation of a hindsight, or, even simpler, Asian, option in an FX or equity setting can be implemented on a lattice [13]. By enlarging the state space by one dimension, which represents a suitably chosen extra state variable, one can indeed roll back future interest rates in tandem with past inflation rates. In practice, this is in effect nothing other than a parallel roll-back of *delayed equations*. The fact that this may become necessary for what appears to be otherwise relatively vanilla products is a unique feature of the inflation market.

The Inflation Smile. Similar to other asset classes, volatilities implied from inflation options, when

visible in the market, tend to display what is known as "smile" and "skew", and several authors have suggested the pricing of inflation options with models that incorporate a smile, for example, see [9, 10]. Unlike many other asset classes, though, the liquidity in options for different strikes, at the time of this study, is extremely thin. It is therefore arguable whether a sophisticated model that reflects a full smile is really warranted for the management of exotic inflation derivatives structures, or whether a simpler, yet more robust, model, that only permits control over the skew, or possibly even only over the level of volatility, if managed prudently (i.e., conservatively), is perhaps preferable. As and when the inflation options market becomes more liquid, the value differences between management of a trading book with a smiling model and a mere skew model may be attainable by hedging strategies, and then the use of fully fledged smile model for inflation is definitely justified. Since the liquidity in options has not significantly increased in the last few years, it is not clear whether this day of sufficient option liquidity to warrant, for instance, stochastic volatility models for inflation, will come.

End Notes

^aThe main difference between CPI and RPI is that the latter includes a number of extra items mainly related to housing, such as council tax and a range of owner-occupier housing costs, for example, mortgage interest payments, house depreciation, buildings insurance, and estate agents' and conveyancing fees.

^bNote that for the $\max(\cdot, 1)$ function to become active, *the average inflation over the life of the bond must be negative*. While inflation does, in practice, become negative for moderate periods of time, the market tends to assign no measurable value to the risk of inflation being negative on average for decades (the typical maturity of inflation bonds). Also note that inflation bond prices tend to be quoted with a four or five digit round-off rule, and that inflation indices themselves are published rounded (typically) to one digit after the decimal point, for example, the UK RPI for June 2008 was published as 216.8 (based on 1987 at 100).

^cThis can be confusing to newcomers to the inflation market who come from a fixed income background since a zero-coupon swap, by definition as used in the interest rate market, is a contract to never exchange any money whatsoever.

^dThis is a consequence of the fact that an Ornstein–Uhlenbeck process, in the limit $\kappa \rightarrow \infty$ with α^2/κ kept constant converges to the *white noise* process, which, in

turn, is in law equal to the temporal derivative of standard Brownian motion.

References

- [1] Andersen, L. Sidenius, J. & Basu, S. (2003). All your hedges in one basket, *Risk* November, 67–72.
- [2] Belgrade, N. Benhamou, E. & Koehler, E. (2004). *A Market Model For Inflation*. Technical report, CDC Ixis Capital Markets and CNCE, January. ssrn.com/abstract=576081.
- [3] Curran, M. (1994). Valuing Asian and portfolio options by conditioning on the geometric mean price, *Management Science* **40**, 1705–1711.
- [4] Deacon, M. Derry, A. & Mirfendereski, D. (2004). *Inflation-indexed Securities*. John Wiley & Sons. ISBN 0470868120.
- [5] Fisher, I. (1930). *The Theory of Interest*. The Macmillan Company.
- [6] Heath, D. Jarrow, R. & Morton, A. (1992). Bond pricing and the term structure of interest rates, *Econometrica* **61**(1), 77–105.
- [7] Hull, J. & White, A. (1990). Pricing interest rate derivative securities, *Review of Financial Studies* **3**(4), 573–592.
- [8] Jarrow, R. & Yildirim, Y. (2003). Pricing treasury inflation protected securities and related derivatives using an HJM model, *Journal of Financial and Quantitative Analysis* **38**(2), 409–430. forum.johnson.cornell.edu/faculty/jarrow/papers.html
- [9] Kruse, S. (2007). *Pricing of Inflation-Indexed Options Under the Assumption of a Lognormal Inflation Index as Well as Under Stochastic Volatility*. Technical report, S-University – Hochschule der Sparkassen-Finanzgruppe; Fraunhofer Gesellschaft – Institute of Industrial Mathematics (ITWM), April. ssrn.com/abstract=948399
- [10] Mercurio, F. & Moreni, N. (2006). Inflation with a smile, *Risk* **19**(3), 70–75.
- [11] Ryten, M. (2007). *Practical Modelling For Limited Price Index and Related Inflation Products*. In *ICBI Global Derivatives Conference*, Paris.
- [12] Vasicek, O.A. (1977). An equilibrium characterisation of the term structure, *Journal of Financial Economics* **5**, 177–188.
- [13] Wilmott, P. (1998). *Derivatives*. John Wiley & Sons.

Further Reading

- Crosby, J. (2007). Valuing inflation futures contracts, *Risk* **20**(3), 88–90.

PETER JÄCKEL & JEROME BONNETON

Swaps

A *swap* is an over-the-counter (OTC) derivative where two parties exchange regular interest rate payments over the life of the contract based on a principal. The most liquid market for swaps is for maturities between 2 and 10 years. However, in some markets, 30- or 40-year swaps are traded as well, and there exist even longer term deals. Swap legs may be denominated in single currency, as in *interest rate swaps* (IRS), or in different currencies, as in *currency swaps*. We cover the most common contracts of both types that use LIBOR (see **LIBOR Rate**) to fix the interest payments. An extensive discussion of how swaps are priced and risk-managed can be found in [2].

Plain Vanilla Swaps

The simplest IRS product is the *fixed/float swap* (sometimes also called *plain vanilla swap*). It is a contract where the payer pays a fixed rate and the receiver pays a floating rate fixed periodically against some rate index, for example, LIBOR; see Figure 1. The plain vanilla swap contract does not involve the principal exchange; only the interest is paid by the parties.

The purpose of such a transaction is to hedge against variations in the interest rates. In such capacity, vanilla swaps are used as basic hedges for all interest-rate-linked products (see **Hedging of Interest Rate Derivatives**). Another common use of vanilla swaps is to translate the fixed rate liabilities into floating rate ones on the back of bond issuance (see **Bond**). Since both forward rate agreement (FRA) and vanilla swap markets are fairly liquid, both of those instruments are used for yield curve construction (see **Yield Curve Construction**).

Most of the swap contracts are agreed on the terms of International Swaps and Derivatives Association



Figure 1 Fixed/float swap. The payer pays a fixed rate and receives a floating rate

(ISDA). For information about ISDA, conventions that specify swap agreements, see [1]. A swap can be viewed as a set of contiguous FRAs. Unlike in the FRAs, however, the frequencies of payments on the floating and the fixed legs of the swap may not be the same. Another difference is that each leg of a swap is normally making payments at the end of each accrual period. The latter fact allows to treat vanilla swaps as “linear” instruments. There exist other types of swaps for which the linear property does not hold (see **Constant Maturity Swap**).

Each payment amount in a plain vanilla swap equals the corresponding interest rate multiplied by the time fraction of the accrual period and by the notional amount. Such simple payout allows us to write down the present value (PV) formula of the plain vanilla swap for one unit of notional as

$$PV(0) = \sum_{k=1}^N L(0, t_{k-1}, t_k) \cdot \tau_k \cdot P(0, t_k) - R_0 \sum_{j=1}^M \tau_j \cdot P(0, t_j) \quad (1)$$

where the first sum runs over the floating leg payments, the second sum runs over the fixed leg payments, τ_k is the time fraction of accrual period (t_{k-1}, t_k) , R_0 is the swap’s fixed rate, $P(T, S)$ is the value of discount bond at T , and $L(t, T, S)$ is the forward reference rate at t .

The valuation formula in equation (1) demonstrates that a plain vanilla swap, effectively being a linear combination of FRAs, can be valued off a forward curve and a discount function. At the inception of the swap transaction, the fixed rate is chosen such that PV is equal to zero. Such a rate is called *break-even swap rate*.

In addition to the plain vanilla swaps there exist other types of swaps based on LIBOR rates, for example, basis swaps and cross-currency basis swaps.

Basis Swaps

In a *basis swap* the parties exchange floating payments fixed against different rate indices. Very often both legs are LIBOR-based. For example, one leg pays 3M LIBOR plus quoted spread, and the other leg pays 6M LIBOR. The economic rationale behind

the existence of the basis spread between different LIBOR rates in single currency can be explained by the credit and the liquidity considerations. A bank would rather provide an unsecured credit for 3 months, then assess the market situation, and provide credit for another 3 months rather than lend money for 6 months from the very beginning.

Cross-currency Basis Swaps

These contracts are designed to provide funding in one currency through borrowing of funds in another currency. In a *cross-currency basis swap*, the two legs of the swap are each denominated in a different currency. Usually, one floating leg is USD floating rate. In this type of a swap contract, the parties exchange both the principals and the interest payments in the two currencies, based on the foreign exchange rate at the time of the trade. The interest payments are normally fixed against the corresponding currencies' 3M LIBOR rates. The principal exchange, typically occurring at both the start and the end of the contract, is an important aspect of this transaction. Without the principal exchange, the associated swap would not

achieve the goal of transforming liability from one currency into another.

References

- [1] Available at: <http://www.isda.org> (2008).
- [2] Miron, P. & Swannell, P. (1992). *Pricing and Hedging Swaps*, Euromoney Institutional Investor PLC.

Further Reading

- Available at: <http://www.bba.org.uk> (2008).
- Henrard, M. (2007). The irony in the derivatives discounting, *Wilmott Magazine* July, 92–98.
- Traven, S. (2008). *Pricing Linear Derivatives with a Single Discount Curve*. working paper, unpublished.

Related Articles

Constant Maturity Swap; Forward and Swap Measures; LIBOR Rate; Swap Market Models; Trigger Swaps; Yield Curve Construction.

SERGEI TRAVEN

Finite Difference Methods for Barrier Options

Barrier options, options that cease to exist (knock-out barrier options) or that only come into existence (knock-in barrier options) when some observable market parameter crosses a designated level (the barrier), have become ubiquitous in financial contracts for virtually all asset classes, including equities, foreign exchange, fixed income, and commodities.

Pricing of continuously monitored barrier options using partial differential equations (PDEs) enables the use of coordinate transformations to obtain smooth, rapid convergence, highly desirable properties that are difficult to obtain using traditional lattice methods. The finite difference method is perhaps the most straight forward and intuitive approach to the numerical solution of PDEs. Yet even in this seemingly simple approach, there exist subtleties that can be exploited for tremendous gains in accuracy and/or computational efficiency [7]. We explore several ways in which finite difference pricing models for barrier options may be designed for increased accuracy and performance.

Pricing models in which the monitoring of the barrier knockout or knock-in condition is approximated as being continuous have been popular principally because they may yield analytic solutions, at least for simple underlying processes (e.g., Black–Scholes.) Even for underlying process for which numerical PDE solution methods are required (e.g., stochastic volatility models with correlation of the underlying level and its instantaneous variance), the continuously monitored barrier condition is often easier to treat with simple numerical methods.

However, in the vast majority of barrier option contracts, the barrier conditions are monitored at discrete times or dates. For the volatilities observed in most markets, even barrier monitoring as frequently as daily yields barrier option prices surprisingly far from those yielded by a continuous monitoring approximation. Broadie *et al.* [2] derive a formula for shifting a discrete barrier so that pricing with continuously monitored barrier yields the discretely monitored barrier price to lowest order in the monitoring interval for a lognormal process. In many cases of interest, however, direct numerical solution of the discretely monitored barrier pricing

problem is required. Therefore, in this article, the continuously monitored barrier case is discussed fairly briefly, while the majority of space is devoted to the discretely monitored barrier case.

For simplicity of notation, we focus on a simple Black–Scholes [1] pricing PDE but none of the methods discussed require that PDE coefficients be constant, so the methods are directly applicable to local volatility models. When methods of conforming finite difference grids to barrier and/or strike positions are discussed, they apply only to the coordinate representing the financial factor subject to a barrier. Other coordinate grids, for example, the instantaneous volatility in a stochastic volatility model [5], or a second or third asset in a basket model are unaffected. Therefore, the methods presented are applicable to these problems as well.

The case of jump-diffusion models requires further analysis and is discussed in [3] (*see Partial Integro-differential Equations (PIDEs)*).

Continuous Monitoring

The simplest pricing partial differential equations (PDEs) typically encountered is the Black–Scholes equation [1], written in the form

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r - q)S \frac{\partial V}{\partial S} - rV = 0 \quad (1)$$

where V is the option value, S is the underlying asset price, r and q are the forward rate and continuous dividend yield, respectively, and σ is the volatility.

A simple up-out call option is used for illustration. In this case, the payoff condition at option expiration $t = T$ is

$$V(S, T) = \max(S - K, 0) : \quad S < B \quad (2)$$

$$V(S, T) = R : \quad S \geq B \quad (3)$$

where B is the up-out barrier, K is the option strike, and R is a rebate paid immediately upon knockout. The upper grid boundary can be placed at $S = B$ with boundary condition $V(B, t) = R$, while that for the lower boundary $S_{\min}$ is not well defined unless $S_{\min} = 0$. Terminating the grid at $S_{\min} = 0$ may be a quite inefficient use of finite difference grid points, since the option value may be very small over most

of the grid. Rather, one may choose a simple rule of thumb such as

$$S_{\text{Min}} = S_0 e^{-N\sigma\sqrt{T}} \quad (4)$$

which places the lower boundary N standard deviations below the initial asset price S_0 , so that for $N \geq 4$ there is only a minuscule chance of the asset price reaching the strike price before expiration to yield a positive payoff. Then $V(S_{\text{Min}}, t) = 0$ becomes an accurate boundary condition.

The commonly used coordinate transformation $x = \log(S)$ is avoided here for three reasons: (i) since the volatility may have a dependence on S (local volatility), it does not necessarily yield a PDE with constant coefficients. (ii) In a fully numerical PDE solution, constant coefficients are of marginal benefit. (iii) It is far more useful to employ other, more general coordinate transformations.

In particular, the transformation used here is a slight warping of an otherwise uniform grid. The grid is warped so that the strike K is “pinned” midway between two grid points independent of S_{Min} or the number of grid points, while keeping the grid spacing as uniform as possible. The benefit of doing so is that smooth monotonic convergence—superior to that typical of lattice methods—is obtained as demonstrated by Tavella and Randall [7]. In the transformed coordinates, the PDE becomes

$$\begin{aligned} \frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 \frac{S^2(x)}{J(x)} \frac{\partial}{\partial x} \left(\frac{1}{J(x)} \frac{\partial V}{\partial x} \right) \\ + (r - q) \frac{S(x)}{J(x)} \frac{\partial V}{\partial x} - rV = 0 \end{aligned} \quad (5)$$

where $J(x) = \partial S(x)/\partial x$ is the Jacobian of the transformation. The new coordinate x is by convention equally spaced on a grid, $x_0 \leq x_i \leq x_I$, where $0 \leq i \leq I$ and subject to the boundary conditions, $S(x_0) = S_{\text{Min}}$, $S(x_I) = B$, and the “pinning condition” $K = (S_p + S_{p+1})/2$, where p is the index of the grid point just below K .

A method for numerically computing a smoothly varying coordinate transformation $S(x)$ so as to place particular “pinning” points (e.g., K) either exactly on grid points or exactly midway between grid points through the use of spline interpolation, is also discussed in detail in [7].

Table 1 shows the convergence of the numerical results *versus* the number of grid points I for an

Table 1 Finite difference results for a continuously monitored up-out barrier call. Parameters: $T = 1$, $K = 100$, $B = 110$, $R = 0.5$, $r = 0.05$, $q = 0.03$, $\sigma = 0.1$, and $\Delta_0 = 100$

I	$V(S_0)$	Error	Ratio
50	0.8493291	−0.0006945	—
100	0.8498746	−0.0001491	4.65
200	0.8499891	−0.0000345	4.32
400	0.8500147	−0.0000086	3.85
800	0.8500214	−0.0000022	4.07
1600	0.8500231	−0.0000005	3.96
3200	0.8500235	−0.0000001	3.95

at-the-money up-out call option. The option and market parameters are given in the caption. The column labeled $V(S_0)$ is the present value, that labeled *Error* is the finite difference result minus the converged result, and that column labeled *Ratio* is the ratio of *Errors* for grid I and grid $I/2$. The numerical solution converges smoothly and monotonically because the option strike (where the payoff has a discontinuity of slope) always has a fixed relationship to the grid, namely it is always midway between two grid points. The convergence is almost exactly quadratic, that is, *Ratio* is very close to 4 when I is doubled. In order to make time discretization error negligible so that the spatial grid discretization effects of interest could be examined in isolation, 50 000 time steps have been used with a Rannacher [4] time discretization scheme. Time discretization effects with practical numbers of time steps are discussed later.

(One of the benefits of full or partially implicit PDE methods like Crank–Nicolson or Rannacher is that the number of spatial grid points and time steps are independent, whereas in lattice methods they are closely coupled.)

Discrete Monitoring

In modeling continuously monitored barrier options, as in the previous section, it is sufficient to place the barrier(s) on the grid boundaries and enforce a boundary condition $V(B, t) = R$. It is of no consequence that gradient of option value $V(S)$ (the option Δ) is discontinuous at the barriers because the pricing PDE (which includes second derivative terms that become singular) is not solved at the barriers. Boundary conditions are enforced instead.

When modeling discretely sampled barriers, however, the barriers appear inside the solution region,

allowing the option value $V(S)$ to “diffuse” across the barriers between monitoring times. The knockout conditions are enforced only at the monitoring times. Consequently, the pricing PDE is solved at the barrier position, and the grid must resolve the very strong gradients that are periodically created there by the discrete monitoring. To do this efficiently, it is convenient to use a coordinate transformation that results in a concentration of grid points near a specified set of points, for example, the strike K and barrier B .

A transformation useful for this purpose can be obtained as the solution of the ordinary differential equation (ODE)

$$\frac{dS}{dx} = \frac{A}{\left[\sum_i \frac{\alpha^2 + (S - P_i)^2}{\alpha^2 \beta^2 + (S - P_i)^2} \right]^{1/2}} \quad (6)$$

where A is a constant to be determined through the boundary conditions $S(x_{\text{Min}}) = S_{\text{Min}}$ and $S(x_{\text{Max}}) = S_{\text{Max}}$, and the summation is over the set of specified points P_i , $0 \leq i \leq n_p$. The properties of the transformation are easiest to see in the special case of a single specified point, where the ODE simplifies to

$$\frac{dS}{dx} = A \left[\frac{\alpha^2 \beta^2 + (S - P)^2}{\alpha^2 + (S - P)^2} \right]^{1/2} \quad (7)$$

Thus, far from the specified point, $|S - P| > \alpha$, $dS/dx = A$ is a constant. If x is uniformly spaced, then far from the specified point P , S is uniformly spaced as well. In the neighborhood of the specified point $S = P$, $dS/dx = \beta A$ is minimized ($\beta < 1$), and the S grid is finer by the ratio β .

To compute that transformation in equation (6), a uniform grid is first created in the underlying coordinate x and an initial guess for the constant A is made. After enforcing the left boundary condition $S(x_{\text{Min}}) = S_{\text{Min}}$, a simple ODE solver such as Runge–Kutta can be used to step through the x grid, computing $S(x)$. If, at the right end of the grid, the boundary condition $S(x_{\text{Max}}) = S_{\text{Max}}$ is not satisfied to a given precision, then the constant A is adjusted *via* a Newton’s method and the ODE solution repeated until the right boundary condition is satisfied. Convergence is typically rapid. Finally, a secondary transformation is applied, slightly warping the computed S grid so as to place the set of specified points either exactly at grid points or at grid midpoints as described in [7].

Table 2 Finite difference results for a discretely monitored up-out barrier call. Parameters: $T = 1$, $K = 100$, $B = 110$, $R = 0.5$, $r = 0.05$, $q = 0.03$, $\sigma = 0.1$, $S_0 = 100$, 250 monitoring dates, 50 000 time steps, $\alpha = 20$, $\beta = 0.1$

I	$V(S_0)$	Error	Ratio
50	0.9244337	0.0052293	—
100	0.9203730	0.0011686	4.47
200	0.9194832	0.0002788	4.19
400	0.9192736	0.0000692	4.02
800	0.9192216	0.0000172	4.03
1600	0.9192087	0.0000043	4.04
3200	0.9192055	0.0000011	4.01

Table 2 shows the convergence of option value with the number of grid points I for an at-the-money up-out call option, discretely monitored 250 times per year. The option and market parameters are given in the caption. The column labeled $V(S_0)$ is the present value, that labeled *Error* is the finite difference result minus the analytic result, and that column labeled *Ratio* is the ratio of *Errors* for grid I and grid $I/2$. Even in this case of frequent discrete monitoring of the barrier, in which discontinuities of option value V are created periodically throughout the integration of the PDE, the numerical solution converges smoothly and monotonically. This is possible because the option strike and the barrier both always have a fixed relationship to the grid, namely, they are always midway between two grid points. The convergence is again almost exactly quadratic, that is, *Ratio* is very close to 4 when the number of grid points is doubled. Again, a very large number of time steps (50 000) have been used in order to isolate spatial grid discretization effects.

Smooth convergence according to a known scaling law also allows fairly robust extrapolation to the continuum result from several computationally inexpensive sparse grid computations.

$$V(I \rightarrow \infty) \approx \frac{4V(2I) - V(I)}{3} \quad (8)$$

For example, using $I = 100$ in equation (8), the estimated continuum result is 0.919187, which is very close to the presumed converged result 0.919204. Thus, one extrapolates to very near the converged value from two very cheap and fast sparse grid computations. Robust extrapolation from lattice computations of barrier option value is typically not possible.

The converged values of the continuously monitored and daily monitored options differ by over 8%

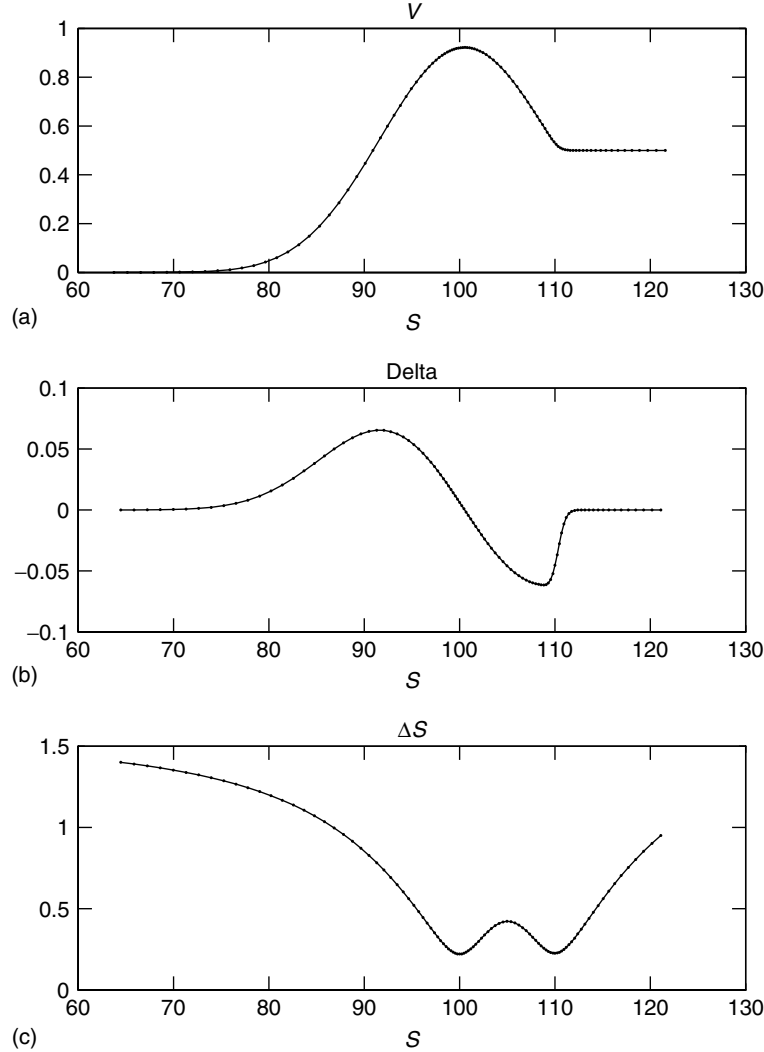


Figure 1 Option value (a), option Δ (b), and grid spacing (c) for a discretely monitored up-out barrier call. Parameters: $T = 1$, $K = 100$, $B = 110$, $R = 0.5$, $r = 0.05$, $q = 0.03$, $\sigma = 0.1$, $S_0 = 100$, 250 monitoring dates, 50 000 time steps, $\alpha = 20$, $\beta = 0.1$, $I = 100$ grid points

even for this relatively low volatility case. The continuity correction formula in [2] yields an option value of 0.9217721, which is much closer to the converged value of the daily monitored option, but still too large by approximately 0.3%. Since the formula is basically an expansion in $\sigma\sqrt{\Delta t_B}$, where Δt_B is the monitoring interval, the accuracy of the correction will degrade for larger volatilities and/or less frequent monitoring. The accuracy of the formula also depends on the proximity of S_0 and B .

The grids used in the computations of Table 2 were concentrated about the option strike and barrier, with coordinate transformation parameters $\alpha = 20$, $\beta = 0.1$. Figure 1 displays the present ($t = 0$) option value $V(S)$, option delta $= \partial V(S)/\partial S$, and the grid spacing $\Delta S(S) = J(S)\Delta x$ when $I = 100$. With the next barrier monitoring date one day hence, the option has significantly larger value at the barrier $S = B = 110$ than the discounted rebate value of approximately $R = 0.5$. The option Δ changes rapidly in

the neighborhood of the barrier, but remains continuous. The grid spacing ΔS is minimized near the two designated points $K = 100$ and $B = 110$ yielding a ratio of largest to smallest spacing on the grid of about a factor of 7. Even with just $I = 100$ grid points, the small grid spacing near the barrier resolves the rapid change of Δ through the region. Similarly, the small spacing near the strike helps resolve the rapid variation of Δ in the region close to option expiration.

Table 3 compares the accuracy of the solution for an up-out call option for three finite difference grids. Grid A is equally spaced ($\beta = 1$) with the barrier exactly on a grid point. (The strike remains at a grid midpoint.) Grid B is also equally spaced, but with the barrier at a midpoint. Finally, Grid C is the nonuniform grid of Table 2, whose data is simply reproduced for comparison.

In the case of Grid A, the fixed relationship of the strike and barrier to the grid yields smooth monotonic

convergence, but it is only linear (*Ratio* is close to 2 when I is doubled) and the observed error is 1–3 orders of magnitude larger than that of Grid C.

In the case of Grid B, the barrier is at a grid midpoint and quadratic convergence (*Ratio* ~ 4) is restored, but because the grid is still uniform, the spacing near the barrier is roughly three times larger than it is for Grid C. As a result, the observed error, while far superior to that of Grid A, is about an order of magnitude larger than that of the nonuniform Grid C. Clearly, use of a nonuniform grid, while taking care to place a discretely sampled barrier at a grid midpoint yields superior convergence. The computational effort involved in computing such a grid is generally negligible compared to the PDE solution itself and obviously worth the effort.

Continuously monitored barriers are optimally priced when the barrier(s) coincide with grid points, while the foregoing numerical results seem to establish that a discretely sampled barrier option is optimally priced when the barrier is midway between two grid points. Of course, as the monitoring frequency is increased, a discretely monitored barrier becomes a continuously monitored one. Therefore, a dimensionless parameter is needed, one that, given a monitoring frequency, determines the optimal grid style to be used. An obvious choice is the ratio of characteristic grid diffusion time near the barrier to the monitoring interval:

$$R_{\Delta t} = \frac{\Delta t_{\text{Grid}}}{\Delta t_B}$$

$$\Delta t_{\text{Grid}} = \frac{(\Delta S/B)^2}{\sigma^2/2}$$

$$\Delta t_B = \frac{T}{n_B} \quad (9)$$

where T is option expiration and n_B is the number of monitoring dates. One expects that for $R_{\Delta t} > 1$, the discretely monitored barrier is effectively continuously monitored to the resolution of the grid, and the barrier should be placed at a grid point. Conversely, when $R_{\Delta t} < 1$ then the barrier should be midway between grid points for optimal accuracy.

This criterion, which is easy to verify numerically, can be used to choose where to place a discretely sampled barrier for optimal accuracy. However, it is easy to see that for almost any discretely sampled barrier option with typical parameters, the barrier(s)

Table 3 Finite difference results for a discretely monitored up-out barrier call using three finite difference grids. Parameters: $T = 1$, $K = 100$, $B = 110$, $R = 0.5$, $r = 0.05$, $q = 0.03$, $\sigma = 0.1$, $S_0 = 100$, 250 monitoring dates, 50 000 time steps. The three grids are described in the text

Grid I	$V(S_0)$	Error	Ratio
Grid A			
50	0.8638922	−0.0553122	—
100	0.878615	−0.0405894	1.36
200	0.8958347	−0.0233697	1.73
400	0.9066875	−0.0125169	1.87
800	0.9127664	−0.0064380	1.94
1600	0.9159481	−0.0032563	1.98
3200	0.9175651	−0.0016393	1.99
Grid B			
50	0.969642	0.0504376	—
100	0.9321862	0.0129818	3.88
200	0.9220514	0.0028470	4.55
400	0.9199134	0.0007090	4.01
800	0.9193805	0.0001761	4.02
1600	0.9192481	0.0000438	4.02
3200	0.9192153	0.0000109	4.00
Grid C			
50	0.9244337	0.0052293	—
100	0.9203730	0.0011686	4.47
200	0.9194832	0.0002788	4.19
400	0.9192736	0.0000692	4.02
800	0.9192216	0.0000172	4.03
1600	0.9192087	0.0000043	4.04
3200	0.9192055	0.0000011	4.01

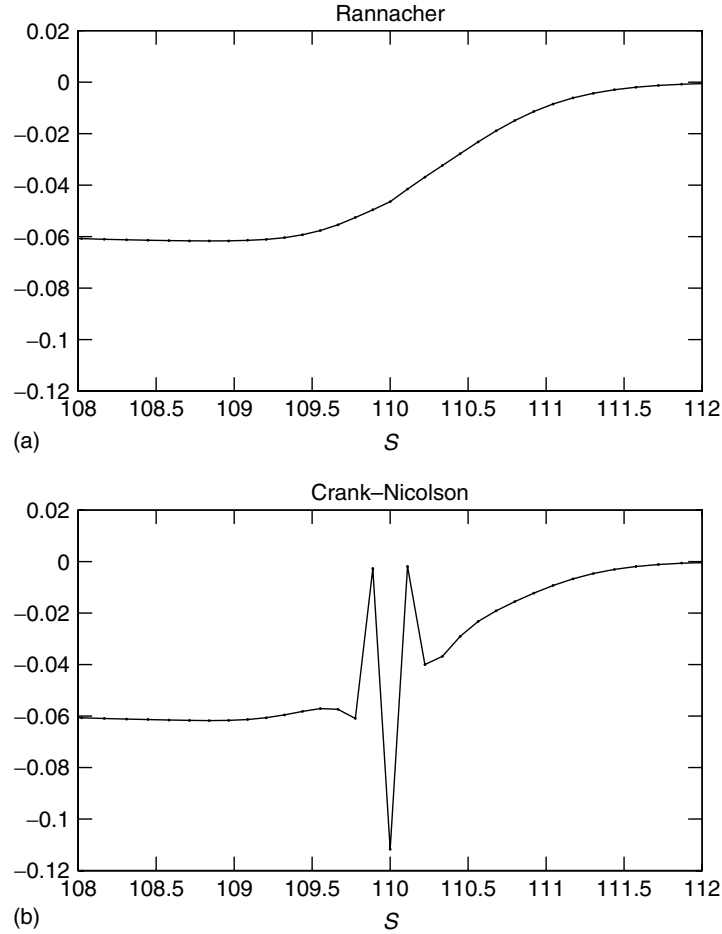


Figure 2 Option Δ in the neighborhood of the barrier for two time discretizations for a discretely monitored up-out barrier call: Rannacher (a), and Crank–Nicolson (b). Parameters: $T = 1$, $K = 100$, $B = 110$, $R = 0.5$, $r = 0.05$, $q = 0.03$, $\sigma = 0.1$, $S_0 = 100$, 250 monitoring dates, 1000 time steps, $\alpha = 20$, $\beta = 0.1$, $I = 100$ grid points

are optimally at grid midpoints. For example, choosing the nonuniform grid with $I = 100$ in Figure 1, one computes $\Delta t_{\text{Grid}} \approx 0.00066$. Since the approximate daily sampling interval is $\Delta t_B \approx 0.004$, the ratio is $R_{\Delta t} \approx 0.16$ and best results are achieved with the barrier midway between grid points. For finer grids, the ratio $R_{\Delta t}$ is even smaller. Thus, for fairly typical parameters, the monitoring interval for which placing barriers on grid points becomes optimal is so small that simply modeling the option as continuously monitored to begin with is probably sufficiently accurate. It is surely much more efficient, because the time steps can be much larger than Δt_{Grid} when semi-implicit time-stepping

methods like Crank–Nicolson or Rannacher [4] are used.

Time Discretization

In the examples thus far, very large numbers of time steps have been used (far in excess of what is practical or desirable), to eliminate time discretization error and isolate spatial discretization error. Considering for the moment the simple heat equation (written in reverse time as it might be in the context of finance),

$$\frac{\partial V}{\partial t} + \kappa \frac{\partial^2 V}{\partial x^2} = 0 \quad (10)$$

the following partially implicit time discretization can be proposed:

$$\frac{V^{n+1} - V^n}{\Delta t} + \theta \kappa \frac{\partial^2 V^{n+1}}{\partial x^2} + (1 - \theta) \kappa \frac{\partial^2 V^n}{\partial x^2} = 0 \quad (11)$$

to advance the solution from t^{n+1} to t^n , where θ is the “implicitness” parameter.

The Crank–Nicolson method, which is often recommended because its truncation error is $O(\Delta t^2)$, corresponds to $\theta = 0.5$. As is well known [5, 6], a Fourier analysis of equation (11) shows that the Crank–Nicolson method is also unconditionally stable. No individual Fourier mode grows exponentially as $n \rightarrow \infty$ no matter the size of Δt relative to the characteristic diffusion time of the grid. However, the Nyquist modes, which change sign at alternate grid points, while stable for large time steps, do not diffuse away, but simply alternate sign on succeeding time steps. On the other hand, a fully implicit method $\theta = 1$ has truncation error $O(\Delta t)$, but the Nyquist modes decay rapidly when Δt is large relative to the diffusion time of the grid.

In pricing discretely sampled barrier options, one enforces the knockout or knock-in conditions by simply changing the option value $V(S)$ appropriately on monitoring dates. Doing so creates discontinuities that add energy to the Nyquist portion of the Fourier spectrum. Thus, while the Nyquist modes formally remain stable under Crank–Nicolson differencing, their amplitude can still grow with each periodic monitoring. The result can be oscillatory solutions near the barrier. And because the Nyquist modes are highly oscillatory, they have a much larger polluting effect on the Greeks Δ and γ than on the value itself [6].

In the Rannacher [4] method, several fully implicit time steps are taken after each barrier monitoring date (or more generally after any event that can result in value or Δ becoming discontinuous), followed by Crank–Nicolson steps. If the number of implicit steps remains constant as the total number of time steps is increased, then the method is $O(\Delta t^2)$. However, it has superior performance when applied to solutions with discontinuities such as digital options or discretely monitored barrier options, since the fully implicit steps in Rannacher drastically reduce the Nyquist modes. An alternative three-level time discretization that likewise has truncation error

of $O(\Delta t^2)$ while eliminating Nyquist modes can also be formulated for the parabolic PDEs common in computation finance. See [7] for a detailed discussion of the three-level scheme for pricing discretely monitored barrier options. Here, the Rannacher scheme is used because of its simplicity.

Figure 2 displays the present ($t = 0$) option delta $= \partial V(S)/\partial S$ in the immediate neighborhood of the barrier for Grid C of Table 3 when $I = 200$, for both Rannacher discretization and Crank–Nicolson. However, the number of time steps is a more practical 1000. The accumulation of Nyquist mode energy near the barrier is evident for Crank–Nicolson differencing, but absent for Rannacher in which two fully implicit time steps were used after each monitoring date. With 250 monitoring dates and 1000 time steps, fully one half of the time steps were implicit. Hence, time discretization error is fairly large despite the stability. However, as the number of time steps is increased, the fraction of fully implicit steps decreases and convergence is quadratic in Δt as shown in Table 4.

Comparing the results of Table 2 ($N = 50\,000$ time steps) to those of Table 4, (N varies with I) it is apparent that most of the error in Table 4 can be ascribed to time discretization, even with a scheme that is clearly quadratically convergent. This is expected for a barrier option with high frequency (approximately daily) monitoring. The monitoring periodically creates discontinuous option values $V(S)$ in the neighborhood of the barrier, and sufficient time steps are required to resolve the evolution of the strong gradients (large Δ s and γ s) created.

References

Table 4 Finite difference results for a discretely monitored up-out barrier call. I is the number of grid points, N the number of time steps. Parameters: $T = 1$, $K = 100$, $B = 110$, $R = 0.5$, $r = 0.05$, $q = 0.03$, $\sigma = 0.1$, $S_0 = 100$, 250 monitoring dates, $\alpha = 20$, $\beta = 0.1$

I	N	$V(S_0)$	Error	Ratio
100	500	0.9284932	0.0092888	—
200	1000	0.9212974	0.0020930	4.44
400	2000	0.9197262	0.0005218	4.01
800	4000	0.9193361	0.0001317	3.96
1600	8000	0.9192377	0.0000333	3.96
3200	16000	0.9192129	0.0000085	3.91

- [1] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- [2] Broadie, M., Glasserman, P. & Kou, S.G. (1997). A continuity correction for discrete barrier options, *Mathematical Finance* **7**, 325–349.
- [3] Cont, R. & Voltchkova, E. (2005). A finite difference scheme for option pricing in jump diffusion and exponential Levy models, *SIAM Journal on Numerical Analysis* **43**(4), 1596–1626.
- [4] Giles, M. & Carter, R. (2006). Convergence of Crank–Nicolson and Rannacher time stepping, *Journal of Computational Finance* **9**, 89–112.
- [5] Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**, 327–343.
- [6] Shaw, W. (1998). *Modeling Financial Derivatives with Mathematica*, Cambridge University Press, Cambridge.
- [7] Tavella, D. & Randall, C. (2000). *Pricing Financial Instruments: The Finite Difference Method*, John Wiley & Sons, New York.

Related Articles

Barrier Options; Corridor Options; Crank–Nicolson Scheme; Finite Difference Methods for Early Exercise Options; Partial Integro-differential Equations (PIDEs); Tree Methods.

CURT RANDALL

Finite Difference Methods for Early Exercise Options

Analytical formulas for the price are not available for options with early exercise possibility. Numerical methods are needed for pricing them. A common way is to derive a partial differential equation (PDE) for the price of the corresponding European-type option (without early exercise possibility) and then modify the equation to allow early exercise. With this approach, most often the underlying partial differential operator is discretized using a finite difference method that we consider in the following. Finite difference methods can be seen as a special (but simpler) case of the finite element method (*see Finite Element Methods*).

An American-type option can be exercised any time during its life (*see American Options*) while a Bermudan-type option can be exercised at specified discrete times during its life. Most typical examples are American put and call options, but American and Bermudan exercise features can be added virtually to any option. In the following, we use an American put option under the Black–Scholes model as an example. Many of the considered methods can be used in a straightforward manner for other types of models and options, as discussed below.

PDE Formulation

We denote the price of an option by V , which is a function of the value of the underlying asset S and time t . As it is more common to consider problems forward in time instead of backward, we use the inverted time variable $\tau = T - t$ instead of t in the following. The payoff function g gives the value of the option at the expiry date T , and also if it is exercised early. For example, for a put option, it is $g(S, \tau) = \max\{K - S, 0\}$, where K is the strike price.

The price V of a European option is given by a parabolic PDE

$$V_\tau + LV = 0 \quad (1)$$

with the initial condition $V = g$ at $\tau = 0$ together with boundary conditions, where L is linear partial

differential operator. Under the Black–Scholes model [6], the operator L is defined by

$$LV = -\frac{1}{2}\sigma^2 S^2 V_{SS} - rSV_S + rV \quad (2)$$

for $S > 0$, where σ is the volatility and r is the interest rate (*see Black–Scholes Formula*).

At the moment when the owner of an American or Bermudan-type option exercises it, she/he will receive a payment defined by the payoff function g . Hence the value V of such an option cannot be less than g , otherwise there would be an arbitrage opportunity. This leads to an early exercise constraint

$$V \geq g \quad (3)$$

which holds whenever the option can be exercised. As this inequality constrains the price V , it does not satisfy the PDE (1) everywhere; instead, it satisfies an inequality

$$V_\tau + LV \geq 0 \quad (4)$$

whenever equation (3) holds. Furthermore, either equation (3) or equation (4) has to hold with the equality at each point. Combining these conditions for an American-type option leads to a linear complementarity problem (LCP)

$$\begin{aligned} V_\tau + LV &\geq 0, \quad V \geq g, \\ (V_\tau + LV)(V - g) &= 0 \end{aligned} \quad (5)$$

For a Bermudan option, the LCP (5) holds when the option can be exercised, and at other times, the PDE (1) holds. Another possibility is to formulate a variational inequality for the price V ; see [1, 18, 27], for example.

At time τ , the space can be divided into two parts: an early exercise (stopping) region $E(\tau)$ and a hold (continuation) region $H(\tau)$. In the early exercise region, it is optimal to exercise the option, whereas in the hold region, it is optimal not to exercise it. For example, under the Black–Scholes model, these regions can be defined as

$$\begin{aligned} E(\tau) &= \{S > 0 : V(S, \tau) = g(S)\}, \\ H(\tau) &= \{S > 0 : V(S, \tau) > g(S)\} \end{aligned} \quad (6)$$

The boundary between these regions is called the *early exercise boundary* $S_f(\tau)$. This is a time-dependent free boundary whose location is not known before solving the LCP (5) or some equivalent problem.

Alternatively, the price V can be computed as the solution of a free boundary problem in which the function $S_f(\tau)$ is also unknown. When the smooth pasting principle holds, the first derivative of V is continuous across the boundary $S_f(\tau)$, and this gives an additional boundary condition on $S_f(\tau)$ that can be used to locate the free boundary. We remark that the smooth pasting principle does not hold, for example, when the model has $\sigma = 0$ like the variance gamma model (see **Variance-gamma Model**). Under the Black–Scholes model, we can formulate the problem for a put option:

$$\begin{aligned} V_\tau + LV &= 0 \quad \text{for } S > S_f(\tau), \\ V(S_f(\tau), \tau) &= g(S_f(\tau)) = S_f(\tau) - K, \\ V_S(S_f(\tau), \tau) &= g_S(S_f(\tau)) = -1 \end{aligned} \quad (7)$$

with the initial condition $V(S, 0) = g(S)$. One advantage of this formulation is that it can give a good approximation for the free boundary $S_f(\tau)$. The domain $(S_f(\tau), \infty)$ in which the PDE needs to be satisfied is time varying. This makes the use of a finite difference method more complicated. An approach used in [32, 43] is to use a time-dependent change of variable in such a way that the computational domain is independent of time. The free boundary problem (7) is nonlinear, and devising a simpler and efficient solution procedure can be also a challenging task. We do not consider this formulation in the following.

A semilinear formulation for American put and call options under the Black–Scholes model was described in [4, 28]. For a put option, it leads to a semilinear PDE

$$V_\tau - LV = q \quad (8)$$

for $S > 0$, where

$$q(S, \tau) = \begin{cases} 0 & \text{if } V(S, \tau) > g(S) \\ rK & \text{if } V(S, \tau) \leq g(S) \end{cases} \quad (9)$$

This PDE has a simple form, and it is posed in the fixed domain $(0, \infty)$. The discontinuous q can make the numerical solution of this problem difficult. The

PDE (8) with a regularized q was solved using an explicit finite difference method in [5].

Finite Difference Scheme

Here, we consider the finite difference discretization of the LCP formulation (5) for an American option. Furthermore, we use the Black–Scholes operator L in equation (2) as an example. For discussion on finite difference discretizations, see [1, 15, 36, 37].

First, the domain $(0, \infty)$ is truncated into a sufficiently large interval $(0, S_{\max})$ and an artificial boundary condition is introduced at $S_{\max}$. For a put option, one possibility is $V(S_{\max}, \tau) = 0$. Next, we define a grid S_i , $i = 0, \dots, p$, such that $0 = S_0 < S_1 < \dots < S_p = S_{\max}$. We allow the grid to be nonuniform, that is, the grid steps $\Delta S_{i+1} = S_{i+1} - S_i$ can vary. An alternative approach would be to use a coordinate transformation together with a uniform grid (see **Finite Difference Methods for Barrier Options**). In the finite difference discretization, we seek the value of V at the grid points S_i . For better accuracy, it is desirable to have finer grid where V changes more rapidly such as near the strike price K for a put and call option, and near where the option price is desired. For example, we can construct a finer grid near the strike price K using a formula

$$S_i = \left(1 + \frac{\sinh(\mu(i/p - \xi))}{\sinh(\mu\xi)}\right) K \quad (10)$$

where the constant μ is solved numerically from the equation $S_p = S_{\max}$. By choosing ξ , we control the amount of grid refinement.

The adopted space finite difference discretization leads to an approximation

$$\begin{aligned} (LV)(S_i) &\approx -\alpha_i V_{i-1} \\ &\quad + (\alpha_i + \beta_i + r)V_i - \beta_i V_{i+1} \end{aligned} \quad (11)$$

at the internal grid points S_i , $i = 1, \dots, p-1$, where the grid point values of V are denoted by $V_i = V(S_i)$. The coefficients α_i and β_i are defined by

$$\begin{aligned} \alpha_i &= \frac{\sigma^2 S_i^2}{\Delta S_i (\Delta S_{i+1} + \Delta S_i)} \\ &\quad - \frac{r S_i}{\Delta S_{i+1} + \Delta S_i}, \end{aligned}$$

$$\beta_i = \frac{\sigma^2 S_i^2}{\Delta S_{i+1}(\Delta S_{i+1} + \Delta S_i)} + \frac{r S_i}{\Delta S_{i+1} + \Delta S_i} \quad (12)$$

if $\Delta S_i \leq \sigma^2 S_i / r$, and otherwise by

$$\alpha_i = \frac{\sigma^2 S_i^2}{\Delta S_i(\Delta S_{i+1} + \Delta S_i)},$$

$$\beta_i = \frac{\sigma^2 S_i^2}{\Delta S_{i+1}(\Delta S_{i+1} + \Delta S_i)} + \frac{r S_i}{\Delta S_{i+1}} \quad (13)$$

The reason to switch over to the latter formulas based on a one-sided difference for V_S when the above condition does not hold is to always have positive coefficients α_i (and β_i). The latter formulas are less accurate. Usually, it is necessary to use them only for a few grid points near $S = 0$, and this has minor (or no) influence on accuracy. We form a vector

$$\mathbf{V} = \begin{pmatrix} V_0 \\ V_1 \\ \vdots \\ V_p \end{pmatrix} \in \mathbb{R}^{p+1} \quad (14)$$

and a $p+1 \times p+1$ tridiagonal matrix $\mathbf{A}$ with $A_{i+1,i} = -\alpha_i$, $A_{i+1,i+1} = \alpha_{i+1} + \beta_{i+1} + r$, and $A_{i+1,i+2} = -\beta_{i+1}$ for $i = 1, \dots, p-1$. The first and last row of $\mathbf{A}$ depends on the boundary conditions. For example, for a put option, we choose them to be zero rows. The matrix vector multiplication $\mathbf{A}\mathbf{V}$ results in a vector that contains the approximations of LV at the grid points.

The space finite difference discretization leads to a semidiscrete LCP for the vector function $\mathbf{V}(\tau)$ given by

$$\mathbf{V}_\tau - \mathbf{A}\mathbf{V} \geq \mathbf{0}, \quad \mathbf{V} \geq \mathbf{g},$$

$$(\mathbf{V}_\tau - \mathbf{A}\mathbf{V})^T (\mathbf{V} - \mathbf{g}) = 0 \quad (15)$$

for $\tau \in (0, T]$. The initial value $\mathbf{V}(0)$ and the vector $\mathbf{g}$ contain the grid point values of the payoff function g . In the above and following, the inequalities hold componentwise.

The finite difference discretization of the LCP gives the value of V only at the grid points. Thus, any simple approximation of the early exercise region $E(\tau)$ and the free boundary $S_f(\tau)$ can have only the same accuracy as the grid step size.

The time discretization approximates the vector function $\mathbf{V}$ at times τ^n such that $0 = \tau^0 < \tau^1 < \dots < \tau^m = T$. In the following, $\mathbf{V}$ at the approximation time τ^n is denoted by $\mathbf{V}^n = \mathbf{V}(\tau^n)$ and the time step between τ^n and τ^{n+1} is denoted by $\Delta\tau^{n+1} = \tau^{n+1} - \tau^n$. The popular θ time stepping scheme leads to a sequence of discrete LCPs

$$\mathbf{B}\mathbf{V}^{n+1} \geq \mathbf{b}^{n+1} \quad \mathbf{V}^{n+1} \geq \mathbf{g},$$

$$(\mathbf{B}\mathbf{V}^{n+1} - \mathbf{b}^{n+1})^T (\mathbf{V}^{n+1} - \mathbf{g}) = 0 \quad (16)$$

for $n = 0, 1, \dots, m-1$ with the initial vector $\mathbf{V}^0 = \mathbf{g}$, where we have used the notations

$$\mathbf{B} = \mathbf{I} + \theta^{n+1} \Delta\tau^{n+1} \mathbf{A},$$

$$\mathbf{b}^{n+1} = [\mathbf{I} + (1 - \theta^{n+1}) \Delta\tau^{n+1} \mathbf{A}] \mathbf{V}^n \quad (17)$$

Different choices of θ^{n+1} lead to the following commonly used method: the explicit Euler method $\theta^{n+1} = 0$, the Crank–Nicolson method $\theta^{n+1} = 1/2$, and the implicit Euler method $\theta^{n+1} = 1$. The Rannacher scheme is obtained by taking a few (say, four) first time steps with the implicit Euler method ($\theta_n = 1$, $n = 1, 2, 3, 4$) and then using the Crank–Nicolson method ($\theta_n = 1/2$, $n = 5, \dots, m$) (see **Crank–Nicolson Scheme**).

Usually, the exercise boundary $S_f(\tau)$ moves rapidly near the expiry and more slowly away from it. For example, for a put option under the Black–Scholes model the boundary behaves like

$$S_f(\tau) \approx K \left(1 - \sigma \sqrt{-\tau \log \tau} \right) \quad (18)$$

near the expiry $\tau = 0$ [29]. Because of this, with uniform time steps, time discretization errors are likely to be much larger during the first time steps than during the later steps. This suggests that it is beneficial to use variable time steps. By gradually increasing the length of time steps, the errors can be made more equidistributed. This way, better accuracy can be obtained with a given number of time steps. Thus, we can reduce the computational effort to reach a desired accuracy.

One possible way to choose the time steps is so that the exercise boundary moves approximately the same amount at each time step. For the Rannacher scheme, by neglecting the logarithm term in the exercise boundary estimate given by equation (18), this approach leads to the approximation

times

$$\begin{aligned}\tau^n &= \left(\frac{n}{2m-4}\right)^2 T, \quad n = 1, 2, 3, 4, \\ \tau^n &= \left(\frac{n-2}{m-2}\right)^2 T, \quad n = 5, \dots, m\end{aligned}\quad (19)$$

where the lengths of the four implicit Euler steps are further reduced by the factor $\frac{1}{2}$.

Another approach is to use adaptive time step selector that uses already computed time steps to predict a good length for the next time step. In [17], the time step $\Delta\tau^{n+1}$ was suggested to be selected according to

$$\Delta\tau^{n+1} = C \left(\min_i \left[\frac{\max\{|V_i^{n-1}|, |V_i^n|, D\}}{|V_i^n - V_i^{n-1}|} \right] \right) \Delta\tau^n \quad (20)$$

where C is a target relative change over the time step and D is a scale of value of the option (for example, we could have $D = 1$ if the value of the option is of the order of one monetary unit).

In the following sections, we describe common ways to solve the discrete LCPs (16), or approximate them and then solve resulting problems.

Solution Methods for LCPs

In the following, we consider the solution of the model LCP

$$\begin{aligned}\mathbf{BV} &\geq \mathbf{b}, \quad \mathbf{V} \geq \mathbf{g}, \\ (\mathbf{BV} - \mathbf{b})^T (\mathbf{V} - \mathbf{g}) &= 0\end{aligned}\quad (21)$$

arising at each time step.

A commonly used iterative method for LCPs is the projected successive over relaxation (PSOR) method [11, 12, 22]. It reduces to the projected Gauss–Seidel method when the relaxation parameter denoted by ω is 1. The basic idea of the projected Gauss–Seidel method is to solve components V_i successively, using the i th row of the system $\mathbf{BV} = \mathbf{b}$ and then project V_i to be g_i if it is below it. The PSOR method over corrects each component by the factor ω before the projection. The following pseudocode performs one iteration with the PSOR method for the LCP (21). The vector $\mathbf{V}$ contains the initial guess for the solution,

which is usually the value from the previous time step or the vector $\mathbf{V}$ returned by the previous iteration.

Algorithm PSOR($\mathbf{B}$, $\mathbf{V}$, $\mathbf{b}$, $\mathbf{g}$)

For $i = 1, p + 1$

$$r_i = b_i - \sum_j B_{i,j} V_j$$

$$V_i = \max\{V_i + \omega r_i / B_{i,i}, g_i\}$$

End For

The PSOR method is guaranteed to converge when the matrix $\mathbf{B}$ is strictly diagonal dominant with positive diagonal entries ($\sum_{j \neq i} |B_{i,j}| < B_{i,i}$ for all i) and the relaxation parameter ω has value in $(0, 1]$. For more precise and general convergence results, we refer to [11, 22]. The convergence rate reduces as the number of grid points grows. On the other hand, smaller time steps make the matrix $\mathbf{B}$ more diagonally dominant and convergence improves. Overall, the convergence slows down somewhat when both space and time steps are reduced at the same rate. The relaxation parameter ω has a big influence on the convergence. Usually, on coarse grids, the optimal value of ω is closer to 1 and on finer grids, it approaches 2. There is no formula for the optimal value; however, it is possible to form a reasonable estimate for it. Even then, quite often, ω is chosen by hand tuning it for a given grid.

A grid-independent convergence rate can be obtained using multigrid methods (see **Multigrid Methods**). For LCPs, suitable projected multigrid methods have been considered in [7, 33, 34], for example. The basic idea is to use a sequence of coarser grids to get better corrections with a small computational effort. These methods are more involved and thus it takes more effort to implement them. Nevertheless, for higher dimensional models, it may be necessary to use these methods to keep computation times feasible.

The LCP (21) can be equivalently formulated as a linear programming (LP) problem [14] when $\mathbf{B}$ is a Z-matrix (offdiagonal entries are nonpositive, that is, $B_{i,j} \leq 0$ for all $i \neq j$) and it is strictly diagonal dominant with positive diagonal entries. The solution $\mathbf{V}$ of equation (21) is given by the LP

$$\min_{\mathbf{V} \in F} \mathbf{c}^T \mathbf{V}, \quad F = \{\mathbf{V} \in \mathbb{R}^{p+1} : \mathbf{V} \geq \mathbf{g}, \mathbf{BV} \geq \mathbf{b}\} \quad (22)$$

for any fixed $\mathbf{c} > 0$ in $\mathbb{R}^{p+1}$. The LP problems can be solved using the (direct) simplex method or an

(iterative) interior point method. For discussion on this approach see [14].

When the matrix $\mathbf{B}$ is symmetric ($\mathbf{B}^T = \mathbf{B}$) and strictly diagonal dominant with positive diagonal entries, the LCP (21) can be equivalently formulated as a quadratic programming (QP) problem, which reads

$$\min_{\mathbf{V} \geq \mathbf{g}} \frac{1}{2} \mathbf{V}^T \mathbf{B} \mathbf{V} - \mathbf{b}^T \mathbf{V} \quad (23)$$

There exists a host of methods to solve such QP problems. Our example of discretization above does not lead to a symmetric matrix $\mathbf{B}$. However, it is often possible to symmetrize the problem by performing a coordinate transformation for the underlying operator or by applying a diagonal similarity transform to LCP. Particularly, these two approaches are applicable for the Black–Scholes operator. For example, we could employ a common transformation to the heat equation to symmetrize $\mathbf{B}$.

The Brennan–Schwartz algorithm [8] is a direct method to solve the LCP (21) with a tridiagonal matrix $\mathbf{B}$. In [27], it has been shown that the algorithm gives the solution of the LCP when $\mathbf{B}$ is a strictly diagonal dominant matrix with positive diagonal entries and there exists k such that $V_i = g_i$ for all $i \leq k$ and $V_i > g_i$ for all $i > k$. The latter condition means that the early exercise region is $(0, S_k)$. This is the case with a put option. A similar direct method has been considered in [16].

Effectively, the algorithm forms an $\mathbf{UL}$ decomposition for $\mathbf{B}$ such that $\mathbf{L}$ is a bidiagonal lower-triangular matrix and $\mathbf{U}$ is a bidiagonal upper-triangular matrix with ones on the diagonal. Then it performs the steps of solving the system $\mathbf{ULV} = \mathbf{b}$ with the modification that in the backsubstitution step employing $\mathbf{L}$, the components of $\mathbf{V}$ are projected to be feasible ($V_i = \max\{V_i, g_i\}$) as soon as they are computed. The following pseudocode solves the LCP (21) using the Brennan–Schwartz algorithm.

Algorithm BS($\mathbf{B}, \mathbf{V}, \mathbf{b}, \mathbf{g}$)

$w_{p+1} = B_{p+1,p+1}$

For $i = p, 1, -1$

$w_i = B_{i,i} - B_{i,i+1}B_{i+1,i}/w_{i+1}$

$V_i = V_i - B_{i,i+1}V_{i+1}/w_{i+1}$

End For

$V_1 = \max\{V_1/w_1, g_1\}$

For $i = 2, p + 1$

$V_i = \max\{(V_i - B_{i,i-1}V_{i-1})/w_i, g_i\}$

End For

When the continuous early exercise region is $(S_k, S_{\max})$ for some k , there are two possible ways to use a Brennan–Schwartz-type algorithm. This is the case with a call option. The first one is to use the above code with the reverse index numbering for the vectors and matrices. Alternatively, the algorithm can be changed to use an $\mathbf{LU}$ decomposition instead of the $\mathbf{UL}$ decomposition.

The Brennan–Schwartz algorithm is usually the fastest way to solve LCPs when it is applicable. Direct methods for LCP with tridiagonal matrices can also be developed for more general exercise regions [13]. These methods are more complicated to implement and they are computationally more expensive.

The LCP (21) can be equivalently formulated using a Lagrange multiplier λ as

$$\begin{aligned} \mathbf{B}\mathbf{V} - \lambda &= \mathbf{b}, \\ \lambda &\geq \mathbf{0}, \quad \mathbf{V} \geq \mathbf{g} \\ \lambda^T (\mathbf{V} - \mathbf{g}) &= 0 \end{aligned} \quad (24)$$

or, alternatively, [21] as

$$\begin{aligned} \mathbf{B}\mathbf{V} - \lambda &= \mathbf{b}, \\ \lambda - \max\{\lambda + \eta(\mathbf{g} - \mathbf{V}), \mathbf{0}\} &= \mathbf{0} \end{aligned} \quad (25)$$

for any $\eta > 0$. For these formulations, several active set solution strategies have been developed; see [1, 21, 42] and references therein.

Penalty Methods

The penalty methods enforce the early exercise constraint by penalizing its violations. For the model LCP (21) a typical power penalty approximation is given by

$$\mathbf{B}\mathbf{V} = \mathbf{b} + \frac{1}{\epsilon} [\max\{\mathbf{g} - \mathbf{V}, \mathbf{0}\}]^k \quad (26)$$

where the maximum and the power function are applied componentwise, and $\epsilon > 0$ is a small penalty parameter. The linear penalty $k = 1$ and the quadratic penalty $k = 2$ are the most common ones. For example, the linear penalty was considered for the Black–Scholes model in [17], and both linear and quadratic penalties were considered for the Heston stochastic volatility model in [45]. The penalty

parameter ϵ controls the quality of the approximation. A small value for ϵ enforces the constraint more strictly. The following pseudocode performs one (semismooth) Newton iteration for the system of nonlinear equations (26). The vector $\mathbf{V}$ contains the initial guess for the solution or the vector $\mathbf{V}$ returned by the previous iteration.

Algorithm Penalty($\mathbf{B}$, $\mathbf{V}$, $\mathbf{b}$, $\mathbf{g}$)

$\mathbf{J} = \mathbf{B}$

For $i = 1, p + 1$

$r_i = b_i - \sum_j B_{i,j} V_j$

If $V_i < g_i$ Then

$r_i = r_i + \frac{1}{\epsilon}(g_i - V_i)^k$

$J_{i,i} = J_{i,i} + \frac{1}{\epsilon}k(g_i - V_i)^{k-1}$

End If

End For

Solve $\mathbf{Jd} = \mathbf{r}$

$\mathbf{V} = \mathbf{V} + \mathbf{d}$

For our example of discretization, the Jacobian matrix $\mathbf{J}$ is tridiagonal and the system $\mathbf{Jd} = \mathbf{r}$ can be solved efficiently using an **LU** decomposition (or the **UL** decomposition used by Algorithm BS), for example. With higher dimensional models, it is probably necessary to solve these problems iteratively to obtain reasonable computational times. Under the Heston model, the BiCGSTAB method with an incomplete **LU** preconditioner was used in [45] and a multigrid method was used in [25].

The penalty approximation (26) leads to $\mathbf{V}$ that violates the early exercise constraint $V \geq g$ by a small amount in the early exercise region. The size of the violation depends on the penalty parameter ϵ and the violation vanishes when ϵ approaches zero. The use of very small values for ϵ can lead to numerical difficulties. A modified penalty approximation in [26],

$$\mathbf{BV} = \mathbf{b} + \max\{\bar{\lambda} + (\mathbf{g} - \mathbf{V})/\epsilon, \mathbf{0}\} \quad (27)$$

with $\bar{\lambda} = \max\{\mathbf{B}\mathbf{g} - \mathbf{b}, \mathbf{0}\}$ can be shown to lead always to $\mathbf{V}$ satisfying the constraint $\mathbf{V} \geq \mathbf{g}$. For example, the above (semismooth) Newton method can be modified to solve the system (27). Another penalty function enforcing the constraint strictly was considered in [32].

Other Approximations for LCP

The simplest way to enforce the early exercise constraint $V \geq g$ is to treat it explicitly. In this so-called explicit payoff method, one first solves the system of linear equations

$$\mathbf{B}\tilde{\mathbf{V}} = \mathbf{b} \quad (28)$$

and then the intermediate solution $\tilde{\mathbf{V}}$ is projected to be feasible by setting

$$\mathbf{V} = \max\{\tilde{\mathbf{V}}, \mathbf{g}\} \quad (29)$$

Typically, this approach leads the order of accuracy to be only $\Delta\tau$. With the explicit Euler method, the LCP (21) reduces to the steps in equations (28) and (29) with $\mathbf{B}$ being the $p + 1 \times p + 1$ identity matrix. Owing to the stability restriction of the explicit Euler method, the time step $\Delta\tau$ has to be of order $(\Delta S)^2$. Thus with the explicit Euler method, the order of accuracy usually is $\Delta\tau \sim (\Delta S)^2$.

Another approach to approximate the LCP with a system of linear equations and a correction enforcing the constraint is to employ an operator splitting method [23, 25]. This method is based on the Lagrange multiplier formulation (24). The basic idea is to use the Lagrange multiplier λ from the previous time step to form a linear system, and then update the solution and Lagrange multiplier to satisfy the constraints pointwise. In the first step, the system of linear equations reads

$$\mathbf{B}\mathbf{V}^{n+1} = \mathbf{b}^n + \lambda^n \quad (30)$$

and the correction in the second step is

$$\begin{aligned} \mathbf{V}^{n+1} - \tilde{\mathbf{V}}^{n+1} &= \lambda^{n+1} - \lambda^n, \\ \lambda^{n+1} &\geq \mathbf{0}, \quad \mathbf{V}^{n+1} \geq \mathbf{g}, \\ (\lambda^{n+1})^T (\mathbf{V}^{n+1} - \mathbf{g}) &= 0 \end{aligned} \quad (31)$$

For higher dimensional models like the Heston model or options on several assets, the LCPs are much more challenging to solve than those with one-dimensional models. One approach is to approximate the resulting problems with a sequence of problems corresponding to a one-dimensional model as in alternating direction implicit (ADI) schemes (see **Alternating Direction Implicit (ADI) Method**). With early exercise possibility, this approach usually leads to solving LCPs with Black–Scholes-type

one-dimensional models. When the solutions have suitable form, these LCPs can be solved efficiently using the Brennan–Schwartz algorithm. Such methods have been considered for American options in [24, 25, 39].

Example

As an example, we consider pricing an American put option with the parameters: $S = K = 100$, $r = 0.1$, $T = 0.25$ years, and $\sigma = 0.2$. The same option was also priced in [17]. For discretization, we truncate the semi-infinite interval at $S_{\max} = 400$, and we construct nonuniform time–space grids that are refined near the expiry date $\tau = 0$ and the strike price K . The space grid is constructed using the formula (10) with $\xi = 0.4$. We choose the time steps explicitly for the Rannacher time stepping according to the formulas (19).

We compute the option price using the Brennan–Schwartz algorithm, the linear penalty method

with $\epsilon = (\Delta\tau^m)^2$, and the PSOR method. For a comparison, we also use the explicit payoff method. Table 1 reports the results. The errors with the explicit and implicit treatments of the constraint are given for a sequence of grids. All implicit methods give essentially the same accuracy. The total iteration counts are reported for the penalty method and the PSOR method. We have optimized the relaxation parameter ω for each grid. CPU times show that it is possible to price hundreds of options with a good precision in one second under the Black–Scholes model.

We have plotted a very coarse grid and an approximation of the early exercise boundary $S_f(\tau)$ in Figure 1. It shows a typical jagged boundary obtained using the LCP formulation. Though the approximation of $S_f(\tau)$ can be only ΔS Accurate, the order of accuracy for the price appears to be $(\Delta S)^2$ and $(\Delta\tau)^2$ with the implicit treatment of the early exercise constraint. This behavior is studied in [17], which concludes that with suitably chosen time steps a quadratic convergence rate is attainable.

Table 1 Numerical results with different methods for an American put option. The parameters are $S = K = 100$, $r = 0.1$, $T = 0.25$ years, and $\sigma = 0.2$. The reference price is 3.0701067. The number of time steps is m and the number of space grid points is $p + 1$. Ratio is the ratio of successive errors. Iter is the total number of iterations. Time is the CPU time in seconds on a 3.2-GHz Pentium 4 PC

Grid		Explicit		Direct	Implicit		BS	Penalty		PSOR	
m	p	error at K	ratio	time	error at K	ratio	time	iter	time	iter	time
18	80	-3.1×10^{-2}		0.0002	-1.5×10^{-2}		0.0003	24	0.0004	204	0.0010
34	160	-1.2×10^{-2}	2.5	0.0008	-3.7×10^{-3}	4.0	0.0009	47	0.0013	511	0.0045
66	320	-5.3×10^{-3}	2.3	0.0032	-9.5×10^{-4}	3.9	0.0034	91	0.0053	1236	0.0212
130	640	-2.5×10^{-3}	2.1	0.0124	-2.4×10^{-4}	3.9	0.0139	179	0.0208	3205	0.1071
258	1280	-1.2×10^{-3}	2.1	0.0506	-6.0×10^{-5}	4.0	0.0561	356	0.0858	8315	0.5645

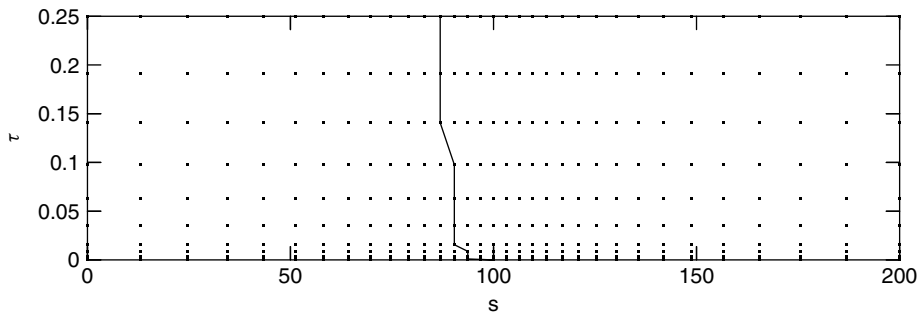


Figure 1 A part of 41×10 space–time grid and an approximation of the early exercise boundary $S_f(\tau)$ given by the Brennan–Schwartz algorithm

Other Models and Options

Often it is desirable to add jumps to the model of the underlying asset (*see Exponential Lévy Models; Variance-gamma Model; Jump-diffusion Models*). Models with jumps lead to a partial integro differential equation (PIDE) for the price of European options (*see Partial Integro-differential Equations (PIDEs)*) and an LCP with same operator can be derived for the price of American options. The discretization of these problems can lead to nonsparse matrices and the efficient solution of the resulting equations is more challenging. Finite-difference-based methods for pricing American options under jump models have been considered in [3, 10, 19, 31, 37, 38, 40, 41].

Stochastic volatility models like the Heston model (*see Heston Model*) lead to LCPs with partial differential operators with two space dimensions. These can be discretized with finite differences in a fairly straightforward manner. Owing to the correlation between the asset value and its volatility the resulting partial differential operator has a second-order cross derivative that can lead to numerical issues. No direct method like the Brennan–Schwartz algorithm is available for these problems. Furthermore, because of two space dimensions, the systems resulting from the discretization are much larger. Finite difference methods for options under stochastic volatility models have been considered in [9, 24, 25, 33, 45]. Similar discretizations and solution methods can be used when interest rates or dividends are modeled as stochastic.

Asian options lead to partial differential operator with an additional dimension for the average of the underlying asset. Similar finite difference methods can be used for American-style Asian options as for the above-mentioned options. The partial differential operator has only the first derivative present in the direction of the new dimension, and this makes discretizing the operator more involved. American-style Asian options have been priced numerically in [20, 44], for example.

The payoff of a multiasset option depends on several underlying assets. Each underlying asset adds one dimension to the model, leading to high dimensional problems. With a few underlying assets, the standard finite difference methods can still be used. For example, American basket options on two stocks

were priced in [39]. With the standard finite differences, the computational cost grows to be high with several underlying assets; see [34]. A special sparse grid technique can be used to reduce the size of the discrete problem (*see Sparse Grids*). This technique relies on the regularity properties of the price function V . The early exercise possibility reduces the regularity of V and thus the straightforward application of the sparse grid technique for American options might lead to reduced accuracy for the price. Nevertheless, in [35] it was observed that the accuracy for American options was still good.

Related Topics

The grids and time steps are chosen usually on the basis of the behavior of discretization error in numerical experiments. Error estimation gives a more systematic way to choose (nearly) optimal discretizations where the number of grid points and time steps are minimized to reach a desired accuracy. In [30] finite difference discretizations were constructed on the basis of an error estimate for European multiasset options. For finite elements, error estimation has been considered in [1] for European options and in [2] for American options.

References

- [1] Achdou, Y. & Pironneau, O. (2005). *Computational Methods for Option Pricing*, *Frontiers in Applied Mathematics*, SIAM, Philadelphia, Vol. 30.
- [2] Allegretto, W., Lin, Y. & Yan, N. (2006). A posteriori error analysis for FEM of American options, *Discrete Continuous Dynamical Systems Series B* **6**, 957–978.
- [3] Almendral, A. & Oosterlee, C.W. (2007). Accurate evaluation of European and American options under the CGMY process, *SIAM Journal of Scientific Computing* **29**, 93–117.
- [4] Benth, F.E., Karlsen, K.H. & Reikvam, K. (2003). A semilinear Black and Scholes partial differential equation for valuing American options, *Finance and Stochastics* **7**, 277–298.
- [5] Benth, F.E., Karlsen, K.H. & Reikvam, K. (2004). A semilinear Black and Scholes partial differential equation for valuing American options: approximate solutions and convergence, *Interfaces and Free Boundaries* **6**, 379–404.
- [6] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.

- [7] Brandt, A. & Cryer, C.W. (1983). Multigrid algorithms for the solution of linear complementarity problems arising from free boundary problems, *SIAM Journal on Scientific and Statistical Computing* **4**, 655–684.
- [8] Brennan, M.J. & Schwartz, E.S. (1977). The valuation of American put options, *Journal of Finance* **32**, 449–462.
- [9] Clarke, N. & Parrott, K. (1999). Multigrid for American option pricing with stochastic volatility, *Applied Mathematical Finance* **6**, 177–195.
- [10] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, Chapman & Hall/CRC, Boca Raton.
- [11] Cottle, R.W., Pang, J.-S. & Stone, R.E. (1992). *The Linear Complementarity Problem*, Academic Press, Boston.
- [12] Cryer, C.W. (1971). The solution of a quadratic programming problem using systematic overrelaxation, *SIAM Journal on Control* **9**, 385–392.
- [13] Cryer, C.W. (1983). The efficient solution of linear complementarity problems for tridiagonal Minkowski matrices, *ACM Transaction on Mathematical Software* **9**, 199–214.
- [14] Dempster, M.A.H. & Hutton, J.P. (1999). Pricing American stock options by linear programming, *Mathematical Finance* **9**, 229–254.
- [15] Duffy, D.J. (2006). *Finite Difference Methods in Financial Engineering*, Wiley Finance Series, John Wiley & Sons, Chichester.
- [16] Elliott, C.M. & Ockendon, J.R. (1982). *Weak and Variational Methods for Moving Boundary Problems*, Research Notes in Mathematics, Pitman, Boston, Vol. 59.
- [17] Forsyth, P.A. & Vetzal, K.R. (2002). Quadratic convergence for valuing American options using a penalty method, *SIAM Journal of Scientific Computing* **23**, 2095–2122.
- [18] Glowinski, R. (1984). *Numerical Methods for Nonlinear Variational Problems*, Springer Series in Computational Physics, Springer-Verlag, New York.
- [19] d'Halluin, Y., Forsyth, P.A. & Labahn, G. (2004). A penalty method for American options with jump diffusion processes, *Numerische Mathematik* **97**, 321–352.
- [20] d'Halluin, Y., Forsyth, P.A. & Labahn, G. (2005). A semi-Lagrangian approach for American Asian options under jump diffusion, *SIAM Journal of Scientific Computing* **27**, 315–345.
- [21] Hintermüller, M., Ito, K. & Kunisch, K. (2003). The primal-dual active set strategy as a semismooth Newton method, *SIAM Journal on Optimization* **13**, 865–888.
- [22] Huang, J. & Pang, J.-S. (1998). Option pricing and linear complementarity, *The Journal of Computational Finance* **2**, 31–60.
- [23] Ikonen, S. & Toivanen, J. (2004). Operator splitting methods for American option pricing, *Applied Mathematics Letters* **17**, 809–814.
- [24] Ikonen, S. & Toivanen, J. (2007). Componentwise splitting methods for pricing American options under stochastic volatility, *International Journal of Theoretical and Applied Finance* **10**, 331–361.
- [25] Ikonen, S. & Toivanen, J. (2007). Efficient numerical methods for pricing American options under stochastic volatility, *Numerical Methods Partial Differential Equations* **24**, 104–126.
- [26] Ito, K. & Kunisch, K. (2006). Parabolic variational inequalities: The Lagrange multiplier approach, *Journal de Mathématiques Pures et Appliquées* **85**, 415–449.
- [27] Jaillet, P., Lamberton, D. & Lapeyre, B. (1990). Variational inequalities and the pricing of American options, *Acta Applicandae Mathematicae* **21**, 263–289.
- [28] Kholodnyi, V.A. (1997). A nonlinear partial differential equation for American options in the entire domain of the state variable, *Nonlinear Analysis* **30**, 5059–5070.
- [29] Kuske, R.A. & Keller, J.B. (1998). Optimal exercise boundary for an American put option, *Applied Mathematical Finance* **5**, 107–116.
- [30] Lötstedt, P., Persson, J., von Sydow, L. & Tysk, J. (2007). Space-time adaptive finite difference method for European multi-asset options, *Computers & Mathematics with Applications* **53**, 1159–1180.
- [31] Matache, A.-M., Nitsche, P.-A. & Schwab, C. (2005). Wavelet Galerkin pricing of American options on Lévy driven assets, *Quantitative Finance* **5**, 403–424.
- [32] Nielsen, B.F., Skavhaug, O. & Tveito, A. (2002). Penalty and front-fixing methods for the numerical solution of American option problems, *The Journal of Computational Finance* **5**, 69–97.
- [33] Oosterlee, C.W. (2003). On multigrid for linear complementarity problems with application to American-style options, *Electronic Transactions on Numerical Analysis* **15**, 165–185.
- [34] Reisinger, C. & Wittum, G. (2004). On multigrid for anisotropic equations and variational inequalities: pricing multi-dimensional European and American options, *Computing and Visualization in Science* **7**, 189–197.
- [35] Reisinger, C. & Wittum, G. (2007). Efficient hierarchical approximation of high-dimensional option pricing problems, *SIAM Journal of Scientific Computing* **29**, 440–458.
- [36] Seydel, R.U. (2006). *Tools for Computational Finance*, 3rd Edition, Universitext, Springer-Verlag, Berlin.
- [37] Tavella, D. & Randall, C. (2000). *Pricing Financial Instruments: The Finite Difference Method*, John Wiley & Sons, Chichester.
- [38] Toivanen, J. (2008). Numerical valuation of European and American options under Kou's jump-diffusion model, *SIAM Journal of Scientific Computing* **30**, 1949–1970.
- [39] Villeneuve, S. & Zannette, A. (2002). Parabolic ADI methods for pricing American options on two stocks, *Mathematics of Operations Research* **27**, 121–149.
- [40] Wang, I.R., Wan, J.W.L. & Forsyth, P.A. (2007). Robust numerical valuation of European and American options under the CGMY process, *The Journal of Computational Finance* **10**, 31–69.
- [41] Zhang, X.L. (1997). Numerical analysis of American option pricing in a jump-diffusion model, *Mathematics of Operations Research* **22**, 668–690.

- [42] Zhang, K., Yang, X.Q. & Teo, K.L. (2006). Augmented Lagrangian method applied to American option pricing, *Automatica Journal of IFAC* **42**, 1407–1416.
- [43] Zhu, Y.-L., Chen, B.-M., Ren, H. & Xu, H. (2003). Application of the singularity-separating method to American exotic option pricing, *Advances in Computational Mathematics* **19**, 147–158.
- [44] Zvan, R., Forsyth, P.A. & Vetzal, K.R. (1998). Robust numerical methods for PDE models of Asian options, *The Journal of Computational Finance* **1**, 39–78.
- [45] Zvan, R., Forsyth, P.A. & Vetzal, K.R. (1998). Penalty methods for American options with stochastic volatility, *Journal of Computational and Applied Mathematics* **91**, 199–218.

Related Articles

Alternating Direction Implicit (ADI) Method; American Options; Asian Options; Crank–Nicolson Scheme; Jump-diffusion Models; Method of Lines; Monotone Schemes; Multigrid Methods; Partial Differential Equations; Partial Integro-differential Equations (PIDEs); Sparse Grids.

JARI TOIVANEN

Partial Differential Equations

In their seminal 1973 article [3], Black and Scholes derived a partial differential equation (PDE) for the call option price by considering a portfolio containing the option and the underlying asset and using absence of arbitrage arguments. Later, in the 1980s, Harrison, Kreps, and Pliska [10, 11] pioneered the use of stochastic calculus in mathematical finance and introduced martingale methods for option pricing in continuous time. This article is an overview of various contexts in finance where PDEs arise, in particular, for option pricing, portfolio optimization, and calibration. The PDE approach of Black–Scholes and the martingale method are related through the Feynman–Kac formula (*see Markov Processes*).

First, we recall the basic derivation of the Black–Scholes PDE and then present the PDEs for various exotic options and, in particular, American options. A paragraph is devoted to Hamilton–Jacobi–Bellman (HJB) equations, which are nonlinear PDEs arising from stochastic control and portfolio optimization. Finally, we discuss some PDEs associated with calibration problems. Computational aspects are treated in companion entries (*see Finite Difference Methods for Barrier Options; Finite Difference Methods for Early Exercise Options; Finite Element Methods*).

The Black–Scholes PDE

We recall the original arguments in the derivation of the Black–Scholes PDE for option pricing [3, 14]. In the Black–Scholes model, we consider a market with a risk-free bond of constant interest rate r , and a stock with a price process S evolving according to a geometric Brownian motion:

$$dS_t = bS_t dt + \sigma S_t dW_t \quad (1)$$

where the drift rate b and the volatility $\sigma > 0$ are assumed to be constant and W is a standard Brownian motion. We now consider a European call option, characterized by its payoff $(S_T - K)_+$ at the maturity T , and with strike K . We denote by V the value of the call option: $V = V(t, S_t)$ is a function of the spot

price of the underlying asset S_t at time t . At the expiry date of the option, we have $V(T, S_T) = (S_T - K)_+$, and for $t < T$, by assuming that the function V is smooth, we get by Itô's formula

$$\begin{aligned} dV &= \frac{\partial V}{\partial t} dt + \frac{\partial V}{\partial S} dS + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} dt \\ &= \left(\frac{\partial V}{\partial t} + bS \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} \right) dt \\ &\quad + \sigma S \frac{\partial V}{\partial S} dW \end{aligned} \quad (2)$$

Now consider a portfolio consisting of the option and a short position Δ in the asset: the portfolio value is then equal to $\Pi = V - \Delta S$, and its self-financed dynamics is given by

$$\begin{aligned} d\Pi &= dV - \Delta dS \\ &= \left(\frac{\partial V}{\partial t} + bS \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} - \Delta bS \right) dt \\ &\quad + \sigma S \left(\frac{\partial V}{\partial S} - \Delta \right) dW \end{aligned} \quad (3)$$

The random component in the evolution of the portfolio Π may be eliminated by choosing

$$\Delta = \frac{\partial V}{\partial S} \quad (4)$$

This results in a portfolio with deterministic increment:

$$d\Pi = \left(\frac{\partial V}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} \right) dt \quad (5)$$

Now, by arbitrage-free arguments, the rate of return of the riskless portfolio Π must be equal to the interest rate r of the bond, that is,

$$\frac{\partial V}{\partial t} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = r\Pi \quad (6)$$

and recalling that $\Pi = V - (\partial V / \partial S)S$, this leads to the *Black–Scholes partial differential equation*:

$$rV - \frac{\partial V}{\partial t} - rS \frac{\partial V}{\partial S} - \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} = 0 \quad (7)$$

This PDE together with the terminal condition $V(T, S) = (S - K)_+$ is a linear parabolic Cauchy

problem, whose solution is analytically known, and this is the celebrated Black–Scholes formula. Moreover, this formula can also be computed as an expectation:

$$V(t, S) = \mathbb{E}^{\mathbb{Q}} \left[e^{-r(T-t)} (S_T - K)_+ \middle| S_t = S \right] \quad (8)$$

where $\mathbb{E}^{\mathbb{Q}}$ denotes the expectation under which the drift rate b in equation (1) is replaced by the interest rate r . $\mathbb{Q}$ is called *risk-neutral probability*.

Linear PDEs for European Options

The derivation presented in the previous paragraph is prototypical. Besides the absence of arbitrage argument, the key points for the derivation of the PDE satisfied by the option price is the Markov property of the stochastic processes describing the market factors. The relation with risk-neutral probability is achieved through the Feynman–Kac formula (see **Markov Processes**). In its basic (multidimensional) version, the Feynman–Kac representation is formulated as follows: let us consider the stochastic differential equation on $\mathbb{R}^n$

$$dX_s = b(X_s) ds + \sigma(X_s) dW_s \quad (9)$$

where b and σ are measurable functions valued, respectively, on $\mathbb{R}^n$ and $\mathbb{R}^{n \times d}$, and W is a n -dimensional Brownian motion. Consider the Cauchy problem

$$rv - \frac{\partial v}{\partial t} - b(x) D_x v - \frac{1}{2} \text{tr}(\sigma \sigma'(t, x) D_x^2 v) = 0 \quad \text{on } [0, T] \times \mathbb{R}^n \quad (10)$$

$$v(T, x) = g(x) \quad \text{on } \mathbb{R}^n \quad (11)$$

Here, $D_x v$ is the gradient, $D_x^2 v$ is the Hessian matrix of v , with respect to the x variable, σ' is the transpose of σ , and tr denotes the trace of a matrix. Then, the solution to this Cauchy problem may be represented as

$$v(t, x) = \mathbb{E} \left[e^{-\int_t^T r(u, X_u^{t,x}) ds} g(X_T^{t,x}) \right] \quad (12)$$

where $X_s^{t,x}$ is the solution to equation (9) starting from x at $s = t$. The Feynman–Kac formula is a

probabilistic interpretation of the integral representation of the solution using the *Green's function*, which is in this case the density of the underlying asset. The Black–Scholes price is a particular case with r constant, $b(t, x) = rx$, $\sigma(t, x) = \sigma x$, and $g(x) = (x - K)_+$. More generally, the interest rate r and the volatility σ may depend on time and spot price.

In the general case, we do not have analytical expression for v , and we have to resort to numerical methods for option pricing. The probabilistic representation (12) is the basis for Monte Carlo methods in option pricing while deterministic numerical methods (finite differences and finite elements) are based on the PDE (11).

Barrier Options

The payoff of these options depends on the fact that the underlying asset crossed or not, some given barriers during the time interval $[0, T]$ (see **Barrier Options**). For example, a down-and-out call option has a payoff $(S_T - K)_+ 1_{\inf_{t \in [0, T]} S_t > L}$. Its price $v(t, x)$ at time t for a spot price $S_t = x$, satisfies the boundary value problem:

$$\begin{aligned} rv - \frac{\partial v}{\partial t} - rx \frac{\partial v}{\partial x} - \frac{1}{2} \sigma^2 x^2 \frac{\partial^2 v}{\partial x^2} &= 0, \\ (t, x) &\in [0, T] \times (L, \infty) \\ v(t, L) &= 0 \\ v(T, x) &= (x - K)_+ \end{aligned} \quad (13)$$

Lookback Options

The payoff of these options involve the maximum or minimum of the underlying asset (see **Lookback Options**). For instance, the floating strike lookback put option pays at maturity $M_T - S_T$ where $M_t = \sup_{0 \leq t \leq T} S_t$. The pair (S_t, M_t) is a Markov process, and the price at time t of this lookback option is equal to $v(t, S_t, M_t)$ where the function v is defined for $t \in [0, T]$, $(S, M) \in \{(S, M) \in \mathbb{R}_+^2 : 0 \leq S \leq M\}$, and satisfies the Neumann problem:

$$\begin{aligned} rv - \frac{\partial v}{\partial t} - rS \frac{\partial v}{\partial S} - \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 v}{\partial S^2} &= 0 \\ \frac{\partial v}{\partial M}(t, M, M) &= 0 \\ v(T, S, M) &= M - S \end{aligned} \quad (14)$$

Asian Options

These options involve the average of the risky asset (see **Asian Options**). For example, the payoff of an Asian call option is $(A_T - K)_+$ where $A_t = \frac{1}{t} \int_0^t S_u du$. The pair (S_t, A_t) is a Markov process, and the price at time t of this Asian option is equal to $v(t, S_t, A_t)$ where the function v is defined for $t \in [0, T]$, $(S, A) \in \mathbb{R}_+^2$, and satisfies the Cauchy problem (see, e.g., [16]):

$$\begin{aligned} rv - \frac{\partial v}{\partial t} - rS \frac{\partial v}{\partial S} - \frac{1}{t}(S - A) \frac{\partial v}{\partial A} \\ - \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 v}{\partial S^2} = 0 \\ v(T, S, A) = (A - K)_+ \end{aligned} \quad (15)$$

American Options and Free Boundary Problems

With respect to European options presented so far, American options give the holder the right to exercise his/her right at any time up to maturity (see **American Options**). For an American put option of payoff $(K - S_t)_+$, $0 \leq t \leq T$, its price at time t and for a spot stock price $S_t = x$ is given by

$$v(t, x) = \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}^{\mathbb{Q}} \left[e^{-r(\tau-t)} (K - S_\tau)_+ \mid S_t = x \right] \quad (16)$$

where $\mathcal{T}_{t,T}$ denotes the set of stopping times valued in $[t, T]$. In terms of PDE, and within the Black–Scholes framework, this leads *via* the dynamic programming principle (see [7]) to a variational inequality:

$$\begin{aligned} \min \left[rv - \frac{\partial v}{\partial t} - rx \frac{\partial v}{\partial x} - \frac{1}{2} \sigma^2 x^2 \frac{\partial^2 v}{\partial x^2}, \right. \\ \left. v(t, x) - (K - x)_+ \right] = 0 \end{aligned} \quad (17)$$

together with the terminal condition: $v(T, x) = (K - x)_+$. This variational inequality maybe written equivalently as

$$rv - \frac{\partial v}{\partial t} - rx \frac{\partial v}{\partial x} - \frac{1}{2} \sigma^2 x^2 \frac{\partial^2 v}{\partial x^2} \geq 0 \quad (18)$$

which corresponds to the supermartingale property of the discounted price process $e^{-rt} v(t, S_t)$

$$v(t, x) \geq (K - x)_+ \quad (19)$$

which results directly from the fact that by exercising his/her right immediately, one receives the option payoff, and

$$\begin{aligned} rv - \frac{\partial v}{\partial t} - rx \frac{\partial v}{\partial x} - \frac{1}{2} \sigma^2 x^2 \frac{\partial^2 v}{\partial x^2} = 0, \\ \text{for } (t, x) \in \mathcal{C} = \{v(t, x) > (K - x)_+\} \end{aligned} \quad (20)$$

which means that as long as we are in the continuation region $\mathcal{C}$, that is, the value of the American option price is strictly greater than its payoff, the holder does not exercise his/her right early. The formulation (18–20) is also called a *free-boundary problem*, and in the case of the American put, there is an increasing function $x^*(t)$, the free-boundary or critical price, which is smaller than K , and such that $\mathcal{C} = \{(t, x) : x > x^*(t)\}$. This free boundary is an unknown part of the PDE and separates the continuation region from the exercise region where the option is exercised, that is, $v(t, x) = (K - x)_+$. The above conditions do not determine the unknown free boundary $x^*(t)$. An additional condition is required, which is the continuous differentiability of the option price across the boundary $x^*(t)$:

$$v(t, x^*(t)) = K - x^*(t), \quad \frac{\partial v}{\partial x}(t, x^*(t)) = -1 \quad (21)$$

This general property is known in optimal stopping theory as the *smooth fit principle*. However, the American option price v is not C^2 , and the nonlinear PDE (17) should be interpreted in a weak sense by means of distributions (see [2] or [12]), or in the viscosity sense (see [5]). Notice that the main difference between PDE for American options and European options is the nonlinearity of the equation

in the former case. This makes the theory and the numerical implementation more difficult than for the European options.

Stochastic Control and Bellman Equations

Stochastic control problems arise in continuous-time portfolio management. This may be formulated in a roughly general framework as follows: we consider a controlled diffusion process in the form

$$dX_s = b(X_s, \alpha_s) ds + \sigma(X_s, \alpha_s) dW_s, \quad \text{in } \mathbb{R}^n \quad (22)$$

where W is a d -dimensional Brownian motion on some filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t), \mathbb{P})$, and $\alpha = (\alpha_t)$ is an adapted process valued in a Borel set $A \subset \mathbb{R}^m$, the so-called control process, which influences the dynamics of the state process X through the drift coefficient b and the diffusion coefficient σ . A stochastic control problem (in a finite horizon) consists of maximizing over the control processes a functional objective in the form

$$\mathbb{E} \left[\int_0^T f(X_t, \alpha_t) dt + g(X_T) \right] \quad (23)$$

where f and g are real-valued measurable functions. The method used to solve this problem, and initiated by Richard Bellman in the 1950s, is to introduce the value function $v(t, x)$, that is the maximum of the objective when starting from state x at time t , and to apply the dynamic programming principle (DPP). The DPP formally states that if a control is optimal from time t until T , then it is also optimal from time $t+h$ until T for any $t+h > t$. Mathematically, the DPP relates the value functions at two different dates t and $t+h$, and by studying the behavior of the value functions when h tends to zero, one obtains a PDE satisfied by v , the so-called Hamilton–Jacobi–Bellman equation:

$$\begin{aligned} -\frac{\partial v}{\partial t} - \sup_{a \in A} \left[b(x, a) D_x v + \frac{1}{2} \text{tr}(\sigma \sigma'(x, a) D_x^2 v) \right. \\ \left. + f(x, a) \right] = 0, \quad \text{on } [0, T) \times \mathbb{R}^n \end{aligned} \quad (24)$$

together with the terminal condition $v(T, x) = g(x)$. Here, and in the sequel, the prime symbol “’” is

for the transpose. The most famous application of Bellman equation in finance is the portfolio selection of Merton [13]. In this problem, an investor can choose to invest at any time between a riskless bond of interest rate r or a stock with Black–Scholes dynamics of rate of return β and volatility γ . By denoting α_t , the proportion of wealth X_t invested in the stock, this corresponds to a controlled wealth process X in equation (22) with $b(x, a) = ax(\beta - r) + rx$ and $\sigma(x, a) = ax\gamma$. The objective of the investor is to maximize his/her expected utility from terminal wealth, which corresponds to a functional objective of the form (23) with $f = 0$, and g an increasing, concave function. A usual choice of utility function is $g(x) = x^p$, with $0 < p < 1$, in which case, there is an explicit solution to the corresponding HJB equation (24). Moreover, the optimal control attaining the argument maximum in equation (24) is a constant equal to $b - r/(1 - p)\gamma^2$. In the general case, there is no explicit solution to the HJB equation. Moreover, the solution is not smooth C^2 and one proves actually that the value function is characterized as the unique weak solution to the HJB equation in the viscosity sense. We refer to the books [9] and [15] for an overview of stochastic control and viscosity solutions in finance.

Related nonlinear PDEs also arise in the uncertain volatility model, where one computes the cost of (super)hedging an option when the volatility is only known to be inside a band (*see Uncertain Volatility Model*).

Diffusion Models and the Dupire PDE

It is well known that the constant coefficient Black–Scholes model is not consistent with empirical observations in the markets. Indeed, given a call option with quoted price C_M on the market, one may associate the so-called *implied volatility*, that is, the volatility σ_{imp} such that the price given by the Black–Scholes formula coincides with C_M . If the Black–Scholes model was sharp, then the implied volatility would not depend on the strike and maturity of the option. However, it is often observed that the implied volatility is far from constant and is actually a convex function of the strike price, a phenomenon known as the volatility *smile*. Several extensions of the Black–Scholes model have been proposed in the literature. We focus here on local volatility models

(see **Local Volatility Model**) where the volatility is a function $\sigma(t, S_t)$ of time and the spot price. In this model, the price $C(t, S_t)$ of a call option of strike K and maturity T satisfies the PDE:

$$rC - \frac{\partial C}{\partial t} - rS \frac{\partial C}{\partial S} - \frac{1}{2} \sigma^2(t, S) S^2 \frac{\partial^2 C}{\partial S^2} = 0, \quad (t, S) \in [0, T) \times (0, \infty)$$

$$C(T, S) = (S - K)_+ \quad (25)$$

The calibration problem in this local volatility model consists of finding a function $\sigma(t, S)$, which reproduces the observed call option prices on the market for all strikes and maturities. In other words, we want to determine $\sigma(t, S)$ in such a way that the prices computed, for example, with the above PDE coincide with the observed prices. The solution to this problem has been provided by Dupire [6]. By fixing the date t and the spot price S , and by denoting $C(t, S, T, K)$ the call option price with strike K and maturity $T \geq t$, Dupire showed that it satisfies the forward (with initial condition) parabolic PDE:

$$\frac{\partial C}{\partial T} + rK \frac{\partial C}{\partial K} - \frac{1}{2} \sigma^2(T, K) K^2 \frac{\partial^2 C}{\partial K^2} = 0, \quad (T, K) \in [t, \infty) \times \mathbb{R}_+ \quad (26)$$

$$C(t, S, t, K) = (S - K)_+ \quad (27)$$

This PDE may be obtained by at least two methods. The first one is based on Itô-Tanaka formula on the risk-neutral expectation representation of the call option price, while the second one is derived by PDE arguments based on the Fokker-Planck equation for the diffusion process (S_t) (see, e.g., [1]). The PDE (26) can be used, in principle, to compute the local volatility function from the call options prices observed at various strikes K and maturities T :

$$\sigma^2(T, K) = 2 \frac{\frac{\partial C}{\partial T} + rK \frac{\partial C}{\partial K}}{K^2 \frac{\partial^2 C}{\partial K^2}} \quad (28)$$

This is known as *Dupire's formula* (see **Dupire Equation**). Notice that equation (28) cannot be used directly since only a finite number of options are quoted on the market. We refer to [1] or [4] for a more advanced discussion of such PDEs.

Semilinear PDEs

We have presented so far three types of PDE in finance: linear PDEs arising from European options and Feynman-Kac formulas, variational inequalities arising from American options and optimal stopping, and nonlinear HJB arising from stochastic control problems. *Semilinear* PDEs the form

$$-\frac{\partial v}{\partial t} - b(x) D_x v - \frac{1}{2} \text{tr}(\sigma \sigma'(x) D_x^2 v) - f(x, v, D_x v' \sigma) = 0, \quad (t, x) \in [0, T) \times \mathbb{R}^n$$

$$v(T, x) = g(x), \quad x \in \mathbb{R}^n \quad (29)$$

Such PDEs arise, for instance, in finance in option pricing with large investor models or in indifference pricing (see **Expected Utility Maximization**). Backward stochastic differential equations [8] provide probabilistic approaches for solving such PDEs arising in finance (see **Backward Stochastic Differential Equations**).

References

- [1] Achdou, Y. & Pironneau, O. (2005). *Computational Methods for Option Pricing*, Frontiers in Applied Mathematics, SIAM.
- [2] Bensoussan, A. & Lions, J.L. (1982). *Applications of Variational Inequalities in Stochastic Control*, North Holland, Amsterdam.
- [3] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy*, **81**, 637–654.
- [4] Cont, R. & BenHamida, S. (2005). Recovering volatility from option prices by evolutionary optimization, *Journal of Computational Finance* **8**(3), 1–34.
- [5] Crandall, M., Ishii, H. & Lions, P.L. (1992). User's guide to viscosity solutions of second order partial differential equations, *Bulletin of the American Mathematical Society* **27**, 1–167.
- [6] Dupire, B. (1994). Pricing with a smile, *Risk* **7**, 18–20.
- [7] El Karoui, N. (1981). Les aspects probabilistes du contrôle stochastique, *Lecture Notes in Mathematics* **876**, 73–238.
- [8] El Karoui, N., Peng, S. & Quenez, M.C. (1997). Backward stochastic differential equations in finance, *Mathematical Finance*, **7**, 1–71.
- [9] Fleming, W. & Soner, M. (1993). *Controlled Markov Processes and Viscosity Solutions*, Springer Verlag.
- [10] Harrison, M. & Kreps, D. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory*, **20**, 381–408.

- [11] Harrison, M. & Pliska, S. (1981). Martingales and Stochastic integrals in the theory of continuous trading, *Stochastic Processes and Applications*, **11**, 215–260.
- [12] Jaillet, P., Lamberton, D. & Lapeyre, B. (1990). Variational inequalities and the pricing of American options, *Acta Applicandae Mathematicae* **21**, 263–289.
- [13] Merton, R. (1973). Optimum consumption and portfolio rules in a continuous-time model, *Journal of Economic Theory* **3**, 373–413.
- [14] Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science*, **4**(1), 141–183.
- [15] Pham, H. (2007). *Optimisation et contrôle stochastique appliqués à la finance*, Springer Verlag.
- [16] Rogers, C. & Shi, Z. (1995). The value of an Asian option, *Journal of Applied Probability*, **32**, 1077–1088.

HUYÊN PHAM

Alternating Direction Implicit (ADI) Method

In the Black–Scholes model, option values are characterized as solutions of certain partial differential equations (PDEs) for European options or partial differential inequalities for American options (*see Partial Differential Equations*). In general, these option values have to be evaluated numerically, for example using the finite difference method (*see Finite Difference Methods for Barrier Options; Finite Difference Methods for Early Exercise Options*). Unfortunately, when considering the pricing of options depending on several assets, the finite difference method suffers from the curse of dimensionality. The alternating direction implicit (ADI) algorithm of Peaceman–Rachford [8], first proposed for the numerical solution of heat equation in two space dimensions, is to reduce the numerical solution of higher dimensional PDEs to a sequence of steps involving only one-dimensional finite difference operators, involving a simple tridiagonal matrix. The resulting algorithms are memory efficient and easy to parallelize. We illustrate this method in a Black–Scholes model with two risky assets.

Black–Scholes Equation for Multiasset Options

Consider a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, IP)$ and let $(W_t)_{t \geq 0}$ be a standard bidimensional $\mathcal{F}_t$ Brownian motion on it. We consider options written on two dividend-paying stocks whose price S_t^1, S_t^2 satisfies the stochastic differential equation:

$$\frac{dS_t^i}{S_t^i} = (r - \delta_i) dt + \sum_{j=1}^2 \sigma_{ij} dW_t^j, \quad i = 1, 2 \quad (1)$$

where r is the interest rate, δ_i is the dividend rate of the stock i , and the matrix $\Sigma = (\sigma_{ij})_{1 \leq i, j \leq n}$ is assumed to be invertible, which ensures that the market is complete. In this setting, the value at time t of an European option with maturity T and payoff function ψ is given by $U(t, S_t)$ where

$$U(t, x) = \mathbb{E} \left(e^{-r(T-t)} \psi(S_T^{t,x}) \right) \quad (2)$$

while the value at time t of an American option with maturity T and payoff function ψ is given by $U_{am}(t, S_t)$ where

$$U_{am}(t, x) = \sup_{\tau \in \mathcal{T}_{0,T-t}} \mathbb{E} \left(e^{-r\tau} \psi(S_\tau^x) \right) \quad (3)$$

where $\mathcal{T}_{0,T-t}$ is the set of all stopping times with values in $[0, T - t]$.

To see this, we shall use the following notation:

- μ is the vector with components $(r - \delta_i - \frac{1}{2} \sum_{j=1}^2 \sigma_{ij}^2)_{i=1,2}$.
- for $x = (x_1, x_2) \in \mathbb{R}^2$; $\exp(x) = (e^{x_1}, e^{x_2})$

Hence, if we denote by $S_t^{\exp(x)}$ the solution of equation (1) with $S_0 = \exp(x)$, we have

$$S_t^{\exp(x)} = \exp(x + \mu t + \Sigma W_t) \quad (4)$$

We make the following change in variables, $\alpha = \Sigma^{-1}\mu$, $\rho = r + (||\alpha||^2/2)$. The Girsanov theorem (**Equivalence of Probability Measures**) ensures that there exists a probability measure $IP^{(\alpha)}$ defined on $(\Omega, \mathcal{F}_T)$ by $\frac{dIP^{(\alpha)}}{dIP} = M_T^{(\alpha)}$, where $M_T^{(\alpha)} = e^{-\alpha \cdot W_T - \frac{|\alpha|^2}{2} T}$ such that $(W_t^{(\alpha)} = W_t + \alpha t)_{0 \leq t \leq T}$ is a standard $IP^{(\alpha)}$ -Brownian motion. Therefore, we have for $x \in \mathbb{R}^2$,

$$\begin{aligned} U(t, \exp(x)) &= \mathbb{E} \left(e^{-r(T-t)} M_{T-t}^{(\alpha)} \right. \\ &\quad \times e^{(\alpha \cdot W_{T-t} + |\alpha|^2/2(T-t))} \psi(\exp(x + \Sigma W_{T-t}^{(\alpha)})) \Big) \\ &= \mathbb{E}^{(\alpha)} \left(e^{-\rho(T-t)} e^{\alpha \cdot W_{T-t}^{(\alpha)}} \psi(\exp(x + \Sigma W_{T-t}^{(\alpha)})) \right) \end{aligned} \quad (5)$$

Define for $y \in \mathbb{R}^2$,

$$V(t, y) = \mathbb{E}^{(\alpha)} \left(e^{-\rho(T-t)} \phi(y + W_{T-t}^{(\alpha)}) \right) \quad (6)$$

with $\phi(y) = e^{\alpha \cdot y} \psi(\exp(\Sigma y))$. We have $U(t, \exp(\Sigma y)) = e^{-\alpha \cdot y} V(t, y)$, the valuation of an European option is now reduced to the computation of V . By means of the same transformation, we have $U_{am}(t, \exp(\Sigma y)) = e^{-\alpha \cdot y} V_{am}(t, y)$, where

$$V_{am}(t, y) = \sup_{\tau \in \mathcal{T}_{0,T-t}} \mathbb{E}^{(\alpha)} \left(e^{-\rho\tau} \phi(y + W_\tau^{(\alpha)}) \right) \quad (7)$$

Introduce the parabolic operator $\mathcal{L}$ defined by

$$\mathcal{L}v = \frac{\partial v}{\partial t} + \frac{1}{2} \Delta v - \rho v \quad (8)$$

where Δ stands for the Laplacian. Then the function V is the solution of the following PDE:

$$\begin{cases} \mathcal{L}v = 0 \\ v(T, \cdot) = \phi \end{cases} \quad (9)$$

while the function V_{am} satisfies the following obstacle problem on $[0, T] \times \mathbb{R}^2$

$$\begin{cases} \max(\mathcal{L}v, \phi - v) = 0 \\ v(T, \cdot) = \phi \end{cases} \quad (10)$$

either in the sense of variational inequalities [7] or in the sense of viscosity solutions [9, Prop 1.2] (*see Monotone Schemes*).

ADI Methods

The idea of Peaceman–Rachford ADI methods can be outlined as follows: consider a two-dimensional PDE arising from the valuation of a European option

$$\frac{\partial u}{\partial t} + \mathcal{A}u = 0 \quad (11)$$

where the operator $\mathcal{A}$ can be decomposed into $\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2$ where each term acts on one variable. The ADI methods consist in splitting each time interval of length Δt in two subintervals and applying the implicit scheme for $\mathcal{A}_1$ and the explicit scheme for $\mathcal{A}_2$ on the time interval $[t_n, t_{n+1/2}]$ and the implicit scheme for $\mathcal{A}_2$ and the explicit scheme for $\mathcal{A}_1$ on the time interval $[t_{n+1/2}, t_{n+1}]$. If we denote by u_n the vector $u(t_n, \cdot)$, the ADI methods compute u_n from u_{n+1} in two steps as follows: first we compute an intermediate value function $u_{n+1/2}$ applying the implicit scheme for the differential operator $\mathcal{A}_1$

$$u_{n+1} - u_{n+1/2} + \frac{\Delta t}{2} (\mathcal{A}_1 u_{n+1/2} + \mathcal{A}_2 u_{n+1}) = 0 \quad (12)$$

next, we compute u_n from $u_{n+1/2}$ applying an implicit scheme to $\mathcal{A}_2$

$$u_{n+1/2} - u_n + \frac{\Delta t}{2} (\mathcal{A}_1 u_{n+1/2} + \mathcal{A}_2 u_n) = 0 \quad (13)$$

Each intermediate equation are afterwards discretized in space using finite difference approximation.

ADI Methods for American Options

We now treat the example of pricing of two-asset American options. The first step is to formulate (10) in a bounded domain, for example on $Q_l = [0, T] \times \Omega_l$ where $\Omega_l =]-l, l[^2$:

$$\begin{cases} \max(\mathcal{L}v_{am}, \phi - v_{am}) = 0 \\ v_{am}(T, \cdot) = \phi \end{cases} \quad (14)$$

with a Dirichlet boundary condition $v_{am} = \phi$ on $]0, T[\times \partial\Omega_l$.

For the numerical resolution of the obstacle problem (14) by finite difference methods, we shall introduce a grid of mesh points (nk, ih, jh) where h, k are mesh parameters that are thought of as tending to zero. Denote by $N = \lceil T/k \rceil$ and M the greatest integer such that $(M + \frac{1}{2})h \leq l$. For each point $x_{ij} = (ih, jh)$, consider a square

$$C_{ij}^{(h)} =](i - 1/2)h, (i + 1/2)h[\times](j - 1/2)h, (j + 1/2)h[\quad (15)$$

and define

$$\begin{aligned} \Omega_h &= \{x_{ij} ; C_{ij}^{(h)} \subset \Omega_l\} \\ &= \{x_{ij} ; -M \leq i, j \leq M\} \end{aligned} \quad (16)$$

In the sequel, V_h is the space generated by $\chi_{ij}^{(h)}$ where $\chi_{ij}^{(h)}$ is the indicator function of the squares $C_{ij}^{(h)}$. If $u_h \in V_h$, we write $u_h(x) = \sum_{i,j=-M}^M u_{ij} \chi_{ij}^{(h)}(x)$. Note that $u_{ij} = u(ih, jh)$. Moreover, we denote by $\phi_{h,k}$ the approximation of the payoff function ϕ in the grid defined by

$$\begin{aligned} \phi_{h,k}(t, x) &= \sum_{n=0}^N \phi_h(x) \mathbf{1}_{[nk, (n+1)k[}(t) \\ &= \sum_{n=0}^N \left(\sum_{i,j=-M}^M \phi_{ij} \chi_{ij}^{(h)}(x) \right) \mathbf{1}_{[nk, (n+1)k[}(t) \end{aligned} \quad (17)$$

where $\phi_{ij} = \phi(x_{ij})$ and $\mathbf{1}_I$ is the indicator function of the interval I . We replace the Laplacian operator by a finite difference approximation and denote throughout

the paper A, B the linear operators defined on V_h by

$$(Au_h)(x) = \sum_{i,j=-M}^M (Au_h)_{ij} \chi_{ij}^{(h)}(x) \quad (18)$$

$$(Bu_h)(x) = \sum_{i,j=-M}^M (Bu_h)_{ij} \chi_{ij}^{(h)}(x) \quad (19)$$

where

$$(Au_h)_{ij} = 1/2(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \quad (20)$$

$$(Bu_h)_{ij} = 1/2(u_{i,j+1} - 2u_{ij} + u_{i,j-1}) \quad (21)$$

with the convention $u_{ij} = 0$ if $|i| \geq M+1$ and $|j| \geq M+1$. Finally, we shall denote by $(\cdot, \cdot)_l$ the inner product on Ω_l and $|\cdot|_l = \sqrt{(\cdot, \cdot)_l}$.

Dynamic Programming and ADI Method

Barles *et al.* [1] discuss a splitting method, which can be viewed as an analytic version of the dynamic programming principle: one builds the approximate solution

$$u_{h,k} = \sum_{n=0}^N u_h^n(x) \mathbf{1}_{[nk, (n+1)k]}(t) \quad (22)$$

in a recursive way, starting from $u_h^N = \phi$ and computing u_h^n for $0 \leq n \leq N$ in two steps:

- *First step:* One solves the following Cauchy problem on $[nk, (n+1)k] \times \Omega_l$ with Dirichlet boundary conditions.

$$\begin{cases} \mathcal{L}v = 0 \\ v(n+1, \cdot) = u_h^{n+1}(\cdot) \end{cases} \quad (23)$$

and denote the solution by $S(k)[u_h^{n+1}]$.

- *Second step:* One computes

$$u(n, \cdot) = \max(\phi_h(\cdot), S(k)[u_h^{n+1}]) \quad (24)$$

The ADI method consists in splitting the initial system into two intermediate *unidimensional* linear systems: u^{n+1} given in V_h with $u_{ij}^{n+1} = \phi_{ij}$ if $|i| = M+1$ or $|j| = M+1$.

One computes an intermediate value function $v_h^{n+1/2} = (v_{ij}^{n+1/2})_{-M \leq i, j \leq M}$ by

- for $|i| \leq M-1$ and $|j| \leq M-1$,

$$\begin{aligned} & \frac{u_{ij}^{n+1} - v_{ij}^{n+1/2}}{\frac{k}{2}} + \frac{v_{i+1,j}^{n+1/2} - 2v_{ij}^{n+1/2} + v_{i-1,j}^{n+1/2}}{2h^2} \\ & + \frac{u_{i,j+1}^{n+1} - 2u_{ij}^{n+1} + u_{i,j-1}^{n+1}}{2h^2} \\ & - \frac{\rho}{2} (u_{ij}^{n+1} + v_{ij}^{n+1/2}) = 0 \end{aligned} \quad (25)$$

- for $i = M$ and $|j| \leq M-1$ (right boundary conditions),

$$\begin{aligned} & \frac{u_{M,j}^{n+1} - v_{M,j}^{n+1/2}}{\frac{k}{2}} + \frac{\phi_{M+1,j} - 2v_{M,j}^{n+1/2} + v_{M-1,j}^{n+1/2}}{2h^2} \\ & + \frac{u_{M,j+1}^{n+1} - 2u_{M,j}^{n+1} + u_{M,j-1}^{n+1}}{2h^2} \\ & - \frac{\rho}{2} (u_{M,j}^{n+1} + v_{M,j}^{n+1/2}) = 0 \end{aligned} \quad (26)$$

- for $i = -M$ and $|j| \leq M-1$ (left boundary conditions),

$$\begin{aligned} & \frac{u_{-M,j}^{n+1} - v_{-M,j}^{n+1/2}}{\frac{k}{2}} + \frac{\phi_{-M-1,j} - 2v_{-M,j}^{n+1/2} + v_{-M+1,j}^{n+1/2}}{2h^2} \\ & + \frac{u_{-M,j+1}^{n+1} - 2u_{-M,j}^{n+1} + u_{-M,j-1}^{n+1}}{2h^2} \\ & - \frac{\rho}{2} (u_{-M,j}^{n+1} + v_{-M,j}^{n+1/2}) = 0 \end{aligned} \quad (27)$$

and the symmetric equations for $|j| = M$. In a more compact form:

$$\begin{aligned} & \frac{u^{n+1} - v^{n+1/2}}{\frac{k}{2}} + \frac{Av^{n+1/2} + a}{h^2} + \frac{Bu^{n+1} + b}{h^2} \\ & - \frac{\rho}{2} (u^{n+1} + v^{n+1/2}) = 0 \end{aligned} \quad (28)$$

with

$$\begin{aligned} a_{Mj} &= \frac{1}{2}\phi_{M+1,j}, & a_{-Mj} &= \frac{1}{2}\phi_{-M-1,j} \\ a_{ij} &= 0 & \text{for } |i| \leq M-1 \end{aligned} \quad (29)$$

$$\begin{aligned} b_{iM} &= \frac{1}{2}\phi_{i,M+1}, \quad b_{i,-M} = \frac{1}{2}\phi_{i,-M-1} \\ b_{ij} &= 0 \quad \text{for } |j| \leq M-1 \end{aligned} \quad (30)$$

In the same manner, one computes $v_h^n = (v_{ij}^n)_{-M \leq i, j \leq M}$ by

$$\begin{aligned} \frac{v^{n+1/2} - v^n}{\frac{k}{2}} + \frac{Av^{n+1/2} + a}{h^2} + \frac{Bv^n + b}{h^2} \\ - \frac{\rho}{2}(v^n + v^{n+1/2}) = 0 \end{aligned} \quad (31)$$

Equations (28) and (31) give

$$\begin{aligned} \left[\left(1 + \frac{\rho k}{4} \right) I - \frac{k}{2h^2} A \right] v_h^{n+1/2} \\ = \left[\left(1 - \frac{\rho k}{4} \right) I + \frac{k}{2h^2} B \right] u^{n+1} + \frac{k}{2h^2} (a + b) \end{aligned} \quad (32)$$

and

$$\begin{aligned} \left[\left(1 + \frac{\rho k}{4} \right) I - \frac{k}{2h^2} B \right] v_h^n \\ = \left[\left(1 - \frac{\rho k}{4} \right) I + \frac{k}{2h^2} A \right] v_h^{n+1/2} + \frac{k}{2h^2} (a + b) \end{aligned} \quad (33)$$

If we set $\eta_{h,k} = \frac{k}{2h^2}(a + b)$,

$$\begin{aligned} A_{h,k}^+ &= \left(1 + \frac{\rho k}{4} \right) I - \frac{k}{2h^2} A \\ A_{h,k}^- &= \left(1 - \frac{\rho k}{4} \right) I + \frac{k}{2h^2} A \end{aligned} \quad (34)$$

and

$$\begin{aligned} B_{h,k}^+ &= \left(1 + \frac{\rho k}{4} \right) I - \frac{k}{2h^2} B \\ B_{h,k}^- &= \left(1 - \frac{\rho k}{4} \right) I + \frac{k}{2h^2} B \end{aligned} \quad (35)$$

one obtains

$$\begin{cases} A_{h,k}^+ v^{n+1/2} = B_{h,k}^- u^{n+1} + \eta_{h,k} \\ B_{h,k}^+ v^n = A_{h,k}^- v^{n+1/2} + \eta_{h,k} \end{cases} \quad (36)$$

Finally, the computation of u_h^n may be summarized by:

$$u_h^n = \max(\phi_h, T_{h,k}[u_h^{n+1}])$$

where the operator $T_{h,k}$ is defined on V_h by:

$$\begin{aligned} T_{h,k}[u_h] &= (A_{h,k}^+)^{-1} (B_{h,k}^+)^{-1} A_{h,k}^- B_{h,k}^- [u_h] \\ &\quad + (A_{h,k}^+)^{-1} (B_{h,k}^+)^{-1} A_{h,k}^- \eta_{h,k} \\ &\quad + (B_{h,k}^+)^{-1} \eta_{h,k} \end{aligned} \quad (37)$$

in which one implicitly used the fact that $(A_{h,k}^+)^{-1}$, $(B_{h,k}^+)^{-1}$, $A_{h,k}^-$ and $B_{h,k}^-$ commute.

Villeneuve and Zanette have proved the stability and the convergence of this scheme (see [9, Proposition 2.4 and Theorem 2.1]). Under a condition on the mesh parameters of the form

$$\lim_{h,k \downarrow 0} \frac{k}{h^2} = 0 \quad (38)$$

From a numerical viewpoint, the systems (28) and (31) involve a tridiagonal matrix. Therefore, each intermediate system can be easily solved by Gaussian elimination, which takes part in the accuracy of the ADI method.

Linear Complementarity Problem and ADI Method

In this section, we describe a second numerical method, which adapts the ADI algorithm to solve the linear complementarity problem (LCP) arising from the discretization of the parabolic variational inequalities related to the pricing of American options.

When the American option value is computed using standard finite differences approximation, one obtains finite dimensional LCP as follows:

$$\begin{aligned} AU &\leq b \\ U &\geq \psi \\ (U - \psi)^T (AU - b) &= 0 \end{aligned} \quad (39)$$

where U is the $(M+1)^2$ vector of American option values on the space grid and A is a block tridiagonal matrix.

There is an extensive literature on the resolution of LCPs and a complete survey can be found in [3]. In particular, the matrix A of the LCP arising from a variational inequality exhibits special properties like sparseness, which make it possible to use efficient

algorithms including projected SOR [4] or direct pivoting methods [5].

Once again, the idea of the ADI method is to exploit the rapid LU decomposition algorithm for tridiagonal matrices by solving recursively a sequence of *one-dimensional* linear complementarity problems involving a tridiagonal matrix. Speed and flexibility of ADI methods come again from this decomposition:

$$\begin{aligned} u_{h,k}(t, x) = & \sum_{n=0}^{N-1} (u_h^n(x) \mathbf{1}_{[nk, (n+1/2)k]}(t) \\ & + u_h^{n+1/2}(x) \mathbf{1}_{[(n+1/2)k, (n+1)k]}(t)) \\ & + u_h^N \mathbf{1}_{[Nk, (N+1/2)k]}(t) \end{aligned} \quad (40)$$

where $u_h^0, u_h^{1/2}, \dots, u_h^N$ are the elements of the vector space V_h satisfying the variational inequalities:

$$\begin{cases} u_h^N = \phi_h \text{ and } \forall n \leq N-1 \\ \forall v_h \geq \phi_h \quad \left(\frac{u_h^{n+1} - u_h^{n+1/2}}{\frac{k}{2}} + \frac{1}{h^2} A u_h^{n+1/2} + \frac{1}{h^2} B u_h^{n+1} - \frac{\rho}{2} (u_h^{n+1/2} + u_h^{n+1}), v_h - u_h^{n+1/2} \right)_l \leq 0 \\ \forall v_h \geq \phi_h \quad \left(\frac{u_h^{n+1/2} - u_h^n}{\frac{k}{2}} + \frac{1}{h^2} A u_h^{n+1/2} + \frac{1}{h^2} B u_h^n - \frac{\rho}{2} (u_h^{n+1/2} + u_h^n), v_h - u_h^n \right)_l \leq 0 \end{cases} \quad (41)$$

with the Dirichlet conditions

$$\begin{aligned} \forall n \quad u_h^n(x_{ij}) = u_h^{n+1/2}(x_{ij}) = \phi_{ij} \\ \text{for } |i| = M \text{ and } |j| = M \end{aligned} \quad (42)$$

As usual, we had rather write the system in the following more compact form:

$$\begin{cases} u_h^N = \phi_h \text{ and } \forall n \leq N-1 \\ \forall v_h \geq \phi_h \quad \left(\left(I + \frac{k}{2h^2} B - \frac{\rho}{2} \right) u^{n+1} - \left(I - \frac{k}{2h^2} A + \frac{\rho}{2} \right) u_h^{n+1/2}, v_h - u_h^{n+1/2} \right)_l \leq 0 \\ \forall v_h \geq \phi_h \quad \left(\left(I + \frac{k}{2h^2} B - \frac{\rho}{2} \right) u^n - \left(I - \frac{k}{2h^2} A + \frac{\rho}{2} \right) u_h^{n+1/2}, v_h - u_h^n \right)_l \leq 0 \end{cases} \quad (43)$$

From a theoretical viewpoint, Villeneuve and Zanette [9, Theorem 3.2] have proved the convergence of this approximation procedure, using the weaker notion of quadratic convergence.

From a numerical viewpoint, there are mainly two families for solving variational inequalities in finite dimension by exploiting their link with linear

complementarity problems (see the next proposition): pivoting methods (algorithms of Cryer [5], Brennan–Schwartz [2]) and iterative methods (e.g., PSOR [4]) (see **Finite Difference Methods for Early Exercise Options**). In the sequel, we specify the computational treatment of equation (41). The inner product in R^{d^2} is denoted by (u, v) and we write $u \geq v$ if $u_i \geq v_i$ for all $i \in \{1, \dots, d^2\}$. The variational inequality (41) becomes a linear complementarity problem in finite dimensions:

Proposition 1 (Linear complementarity problem [3, 7]). *Let A a $d^2 \times d^2$ matrix and $u, \theta, \phi \in R^{d^2}$. The following systems are equivalent:*

$$\begin{aligned} \bullet \quad (S) \quad & Au \geq \theta, \quad u \geq \phi \\ & (Au - \theta, \phi - u) = 0 \end{aligned} \quad (44)$$

$$\begin{aligned} \bullet \quad (S') \quad & u \geq \phi, \quad \forall v \geq \phi \\ & (Au - \theta, v - u) \geq 0 \end{aligned} \quad (45)$$

For any matrix $u = (u_{ij})_{1 \leq i, j \leq d}$, we choose one of the most obvious methods of ordering

$$u = [u_{11}, \dots, u_{d1}, \dots, u_{1d}, \dots, u_{dd}] \quad (46)$$

The ADI scheme (41) consists in approximating $u(nk, ih, jh)_{0 \leq n \leq N; -M \leq i, j \leq M}$ by (u_{ij}^n) ordered in the way defined above by

$$u_{ij}^N = \phi(ih, jh) = \phi_{ij} \quad \text{for } -M \leq i, j \leq M \quad (47)$$

and for $0 \leq n \leq N - 1$ and $-M \leq i, j \leq M$,

$$\left\{ \begin{array}{l} u_{ij}^{n+1/2} \geq \phi_{ij} \text{ and } u_{ij}^n \geq \phi_{ij} \\ u_{ij}^{n+1/2} = \phi_{ij} \text{ and } u_{ij}^n = \phi_{ij} \text{ for } |i| = M + 1 \text{ or } |j| = M + 1 \\ \frac{-1}{2}u_{i+1,j}^{n+1/2} + \alpha u_{ij}^{n+1/2} - \frac{1}{2}u_{i-1,j}^{n+1/2} \geq \frac{-1}{2}u_{i+1,j}^{n+1} + \alpha u_{ij}^{n+1} - \frac{1}{2}u_{i-1,j}^{n+1} \\ \left(\left(\frac{-1}{2}u_{i+1,j}^{n+1/2} + \alpha u_{ij}^{n+1/2} - \frac{1}{2}u_{i-1,j}^{n+1/2} \right) - \frac{1}{2}u_{i+1,j}^{n+1} - \alpha u_{ij}^{n+1} + \frac{1}{2}u_{i-1,j}^{n+1}, u_{ij}^{n+1/2} - \phi_{ij} \right) = 0 \\ \frac{-1}{2}u_{i,j+1}^n + \alpha u_{ij}^n - \frac{1}{2}u_{i,j-1}^n \geq \frac{-1}{2}u_{i,j+1}^{n+1/2} + \alpha u_{ij}^{n+1/2} - \frac{1}{2}u_{i,j-1}^{n+1/2} \\ \left(\left(\frac{-1}{2}u_{i,j+1}^n + \alpha u_{ij}^n - \frac{1}{2}u_{i,j-1}^n \right) - \left(\frac{-1}{2}u_{i,j+1}^{n+1/2} + \alpha u_{ij}^{n+1/2} - \frac{1}{2}u_{i,j-1}^{n+1/2} \right), u_{ij}^n - \phi_{ij} \right) = 0 \end{array} \right. \quad (48)$$

where $\alpha = (1 + \frac{\rho k}{4} - \frac{k}{2h^2})$. The Kronecker product of two matrices $M, N \in \mathcal{M}_d(R)$ is the $d^2 \times d^2$ matrix denoted by $M \otimes N$ with entries $(M \otimes N)_{ij} = M_{ij}N$ for $1 \leq i, j \leq d$. To take into account the boundary conditions, for $u \in \mathcal{M}_{d^2}(R)$ with $d = 2M + 1$, we define $\bar{u} \in \mathcal{M}_{d^2}(R)$ with components $i = 1$

$$\left\{ \begin{array}{l} \bar{u}_{11} = u_{11} - \frac{k}{h^2}(\phi(-(M+1)h, -Mh) + \phi(-Mh, -(M+1)h)) \\ \bar{u}_{1j} = u_{1j} - \frac{k}{h^2}(\phi(-(M+1)h, -(M+1+j)h)) \quad 2 \leq j \leq d-1 \\ \bar{u}_{1d} = u_{1d} - \frac{k}{h^2}(\phi(-(M+1)h, Mh) + \phi(-Mh, (M+1)h)) \end{array} \right. \quad (49)$$

$2 \leq i \leq d-1$

$$\left\{ \begin{array}{l} \bar{u}_{i1} = u_{i1} - \frac{k}{h^2}(\phi((-M+1+i)h, -(M+1)h)) \\ \bar{u}_{ij} = u_{ij} \bar{u}_{id} = u_{id} - \frac{k}{h^2}(\phi((-M+1+i)h, (M+1)h)) \end{array} \right. \quad (50)$$

$i = d$

$$\left\{ \begin{array}{l} \bar{u}_{d1} = u_{d1} - \frac{k}{h^2}(\phi((M+1)h, -Mh) + \phi(Mh, -(M+1)h)) \\ \bar{u}_{dj} = u_{dj} - \frac{k}{h^2}(\phi((M+1)h, -(M+1+j)h)) \quad 2 \leq j \leq d-1 \\ \bar{u}_{dd} = u_{dd} - \frac{k}{h^2}(\phi((M+1)h, Mh) + \phi(Mh, (M+1)h)) \end{array} \right. \quad (51)$$

Then, the linear complementary problem can be written as

$$\left\{ \begin{array}{l} u^{n+1/2} \geq \phi_h \text{ and } u^n \geq \phi_h \\ \left[I \otimes \left(I + \frac{\rho k}{2} I - \frac{k}{2h^2} T \right) \right] u^{n+1/2} \geq \left[I \otimes \left(I - \frac{\rho k}{2} I + \frac{k}{2h^2} T \right) \right] \bar{u}^{n+1} \\ \left(I \otimes \left(I + \frac{\rho k}{2} I - \frac{k}{2h^2} T \right) u^{n+1/2} - I \otimes \left(I - \frac{\rho k}{2} I + \frac{k}{2h^2} T \right) \bar{u}^{n+1}, u^{n+1/2} - \phi \right) = 0 \\ \left[\left(I + \frac{\rho k}{2} I - \frac{k}{2h^2} T \right) \otimes I \right] u^n \geq \left[\left(I - \frac{\rho k}{2} I + \frac{k}{2h^2} T \right) \otimes I \right] \bar{u}^{n+1/2} \\ \left(\left[\left(I + \frac{\rho k}{2} I - \frac{k}{2h^2} T \right) \otimes I \right] u^n - \left[\left(I - \frac{\rho k}{2} I + \frac{k}{2h^2} T \right) \otimes I \right] \bar{u}^{n+1/2}, u^n - \phi \right) = 0 \end{array} \right. \quad (52)$$

with

$$T = \begin{pmatrix} -2 & 1 & \cdots & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & 1 & -2 & 1 \\ 0 & \cdots & \cdots & 1 & -2 \end{pmatrix} \quad (53)$$

Hence, the pricing of the American option defined by ϕ is now reduced to the computation of the bidimensional linear complementarity problem (52), which has been split in two intermediate unidimensional linear complementarity problems involving a tridiagonal matrix.

To summarize, ADI methods

- are competitive in terms of speed of computation in comparison with standard iterative methods used in financial literature for solving linear complementarity problems;
- lead to algorithms that are very easy to implement in the Black–Scholes setting; and
- are unconditionally stable for the L^2 norm [6], which simplifies their implementation in practice.

References

- [1] Barles, G., Daher, Ch. & Romano, M. (1995). Convergence of numerical Schemes for problems arising in Finance theory, *Mathematical Models and Methods in Applied sciences* **5**(1), 125–143.

- [2] Brennan, M.J. & Schwartz, E. (1977). The valuation of the American put option, *Journal of Finance* **32**, 449–462.
- [3] Cottle, R.W., Pang, J.S. & Stone, R.E. (1992). *The Linear Complementarity Problem*, Academic Press.
- [4] Cryer, C.W. (1971). The solution of a quadratic programming problem using systematic overrelaxation, *SIAM Journal on Control and Optimization* **9**, 385–392.
- [5] Cryer, C.W. (1983). The efficient solution of linear complementarity problems for tridiagonal Minkowski matrices, *ACM Transactions on Mathematical Software* **9**, 199–214.
- [6] Hout, K.J. & Welfert, B.D. (2007). Stability of ADI schemes applied to convection-diffusion equations with mixed derivative terms, *Applied Numerical Mathematics* **57**, 19–35.
- [7] Jaillet, P., Lamberton, D. & Lapeyre, B. (1990). Variational inequalities and the pricing of the American options, *Acta Applicandae Mathematicae* **21**, 263–289.
- [8] Peaceman, P.W. & Rachford, H.H. (1955). The numerical solution of Parabolic and Elliptic differential equation, *Journal of the Society for Industrial and Applied Mathematics* **3**, 28–42.
- [9] Villeneuve, S. & Zanette, A. (2002). Parabolic A.D.I. methods for pricing American options on two stocks, *Mathematics of Operation Research* **27**(1), 121–149.

STÉPHANE VILLENEUVE

Conjugate Gradient Methods

In option pricing on the basis of the Black–Scholes model [3], the value of an option is governed by a partial differential equation (PDE) (see **Partial Differential Equations**). Except for special cases, analytic solutions generally do not exist and numerical methods are necessary to approximate the solution. For instance, finite difference methods [20] approximate the solution on a mesh of discrete asset prices. An implicit discretization of the PDE then requires in each time step solving a linear system:

$$Ax = b \quad (1)$$

where x is the solution of next time step, b is the right-hand side, and A is an $N \times N$ matrix (see **Finite Difference Methods for Barrier Options; Finite Difference Methods for Early Exercise Options**). It is interesting to note that matrices arising from option pricing PDEs are often sparse and typically have $O(N)$ nonzeros.

The solution time and storage for solving large linear systems can be very significant. Gaussian elimination is a standard method for solving linear systems, but owing to the issue of fill-in [5], it is usually deemed too expensive in practice when the underlying has more than two assets. Iterative methods [10, 18], for example, Jacobi, Gauss–Seidel, and successive over-relaxation (SOR), on the other hand, are simple to apply. However, their convergence rates, typically depending on the mesh size, are very slow for large problems. The ADI method [16] (see **Alternating Direction Implicit (ADI) Method**) has been used in option pricing. While it can be made efficient for some linear PDEs, it is not clear how it can be easily extended to more general and nonlinear equations.

One way to improve on the classical iterative methods is to use dynamically computed parameters based on the current (and previous) iteration information. The hope is that the appropriately selected parameters would compute “optimal” solutions in some sense. Ideally, it would be desirable to have an iterative method, which (i) is simple to implement, (ii) takes advantage of sparsity structure of A , and (iii) generates solutions whose errors are minimized

in some way. It turns out that such a method can be developed, which is known as the *conjugate gradient (CG) method* [11]. We begin the discussion for symmetric matrices and then generalize the idea to general nonsymmetric cases.

Symmetric Case

Consider the linear system (1) and assume for now that A is symmetric positive definite (i.e., all eigenvalues are positive). To search the solution x in the N -dimensional space $\mathbb{R}^N$ is generally difficult when N is large. A simpler problem would be to search an approximate solution x^k from a low-dimensional subspace S_k where the dimension k is typically much smaller than N .

There are different ways to select x^k from S_k . Intuitively, x^k is “optimal” if $\|x - x^k\|$ is minimized. Geometrically, it is equivalent to saying that the error $e^k \equiv x - x^k$ is orthogonal to S_k , that is, $\langle e^k, s \rangle \equiv \sum_{i=1}^N e_i^k s_i = 0$ for all $s \in S_k$. To enforce this condition, one would need e^k , which is not known. To address this issue, the A -inner product, defined as $\langle u, v \rangle_A \equiv \langle Au, v \rangle$, is used instead. The orthogonality condition then becomes

$$0 = \langle e^k, s \rangle_A = \langle Ae^k, s \rangle = \langle r^k, s \rangle \quad \forall s \in S_k \quad (2)$$

where $r^k \equiv b - Ax^k$ is the residual vector, which is computable. What it does is to minimize $\langle e^k, e^k \rangle_A \equiv \|e^k\|_A^2$, the A -norm of the error.

Different choices of S lead to different methods. For instance, the method of Steepest descent (SD) [9, 18] chooses the one-dimensional subspace $S = \text{span}\{p\}$, where p is the residual vector of the current approximation. A new approximate solution is obtained by enforcing the orthogonality condition (2). The procedure is repeated with another search direction. Note that SD does not increase the dimension of the search subspace S but rather changes S from every iteration. The main drawback of SD is slow convergence in practice, typically because the search directions may repeat and so SD may end up searching in the same direction again and again [10].

Conjugate Gradient

To avoid unnecessary duplicated search effort as in SD, the conjugate gradient (CG) method [11, 18]

is used to find an optimal solution x^k from a set of search directions $\{p^i\}$, which are A -orthogonal; that is, $\langle p^i, p^j \rangle_A = 0$ if $i \neq j$. The A -orthogonality property guarantees that each of the p^k searches in a unique subspace and one never has to look in that subspace again.

The basic idea of CG is to start with an initial guess x^0 and then find an approximate solution in a subspace S . It first begins with a small (dimension 1) subspace and then increases the dimension one by one to obtain better approximation. More precisely, at the k th step, the search subspace is

$$S_k = \text{span}\{p^0, p^1, \dots, p^{k-1}\} \quad (3)$$

where $\{p^i\}_{i=1}^{k-1}$ are search vectors computed from previous steps. CG then looks for the best approximate solution $x^k \in x^0 + S_k$, that is,

$$x^k = x^0 + \sum_{i=0}^{k-1} \alpha_i p^i \quad (4)$$

The orthogonality condition (2) and the A -orthogonality property of $\{p^i\}$ yield $\alpha_i = \langle p^i, r^0 \rangle / \langle p^i, p^i \rangle_A$. Note that α_i does not depend on k . Hence,

$$x^k = x^0 + \sum_{i=0}^{k-2} \alpha_i p^i + \alpha_{k-1} p^{k-1} = x^{k-1} + \alpha_{k-1} p^{k-1} \quad (5)$$

Thus, only the last search direction needs to be stored to update x^k from x^{k-1} .

Once x^k is known, SD would use the residual r^k as the new search direction. CG, however, makes r^k A -orthogonalized against all previous $\{p^i\}$ to obtain p^k , that is,

$$p^k = r^k + \sum_{i=0}^{k-1} \beta_i p^i \quad (6)$$

where $\beta_i = -\langle r^k, p^i \rangle_A / \langle p^i, p^i \rangle_A$ given by the A -orthogonality condition. The new search subspace is then defined as $S_{k+1} = \text{span}\{p^0, \dots, p^k\}$ and the new approximate solution x^{k+1} is computed until it is sufficiently accurate.

A potential drawback of CG is to store all $\{p^i\}$ for computing p^k . Simplification is necessary to make the method practical. An important observation in deriving CG is that S_k is the same as the Krylov subspace [9, 10, 18], defined as

$$\mathcal{K}_k \equiv \text{span}\{r^0, Ar^0, \dots, A^{k-1}r^0\} \quad (7)$$

As a result,

$$\begin{aligned} p^i \in \text{span}\{p^0, \dots, p^i\} &= S_{i+1} = \mathcal{K}_{i+1} \\ &= \text{span}\{r^0, \dots, A^i r^0\} \end{aligned} \quad (8)$$

Hence, $Ap^i \in \text{span}\{Ar^0, \dots, A^{i+1}r^0\} \subset \mathcal{K}_{i+2} = S_{i+2}$. By equation (2), $r^k \perp S_j$, $j = 1, \dots, k$ and hence, $r^k \perp S_{i+2}$ for any $i \leq k-2$. Thus, $\langle Ap^i, r^k \rangle = 0 = \beta_i$, $i \leq k-2$, and equation (6) is simplified to

$$p^k = r^k + \beta_{k-1} p^{k-1} \quad (9)$$

Thus, as for x^k , only the last search vector needs to be stored. Besides, more convenient formulas for α_k and β_k can be derived by applying the various orthogonality properties. Finally, the CG algorithm is given as follows:

Algorithm: Conjugate Gradient

```

 $x^0$  = initial guess
 $r^0 = b - Ax^0$ 
( $p^{-1} = 0$ ,  $\beta_{-1} = 0$ )
for  $k = 0, 1, 2, \dots$ , until convergence
     $p^k = r^k + \beta_{k-1} p^{k-1}$ 
     $\alpha_k = \langle r^k, r^k \rangle / \langle p^k, Ap^k \rangle$ 
     $x^{k+1} = x^k + \alpha_k p^k$ 
     $r^{k+1} = r^k - \alpha_k Ap^k$ 
     $\beta_k = \langle r^{k+1}, r^{k+1} \rangle / \langle r^k, r^k \rangle$ 
end
    
```

Note that the CG algorithm only involves simple vector operations and matrix–vector multiply, and hence the sparsity structure of A can be fully and easily explored. Moreover, two-term recurrence formulas exist for the updates of x^k and other variables. As such, the work and storage for one CG iteration are $O(N)$. Note also that the matrix A is not really needed as long as one can compute the matrix–vector product. This property is particularly useful when A is not available explicitly; see the section Application.

Convergence of CG

Compared to other iterative methods, CG has the desirable property that x^k is the “best” approximation from S_k . Thus, the CG solution x^k improves as the dimension of S_k increases. When $k = N$, then $S_N = \mathbb{R}^N$ and so $e^N = 0$. Hence, CG obtains the exact solution in at most N iterations (known as the *finite termination property* [18]). In practice,

however, CG is often treated as an iterative method in the sense that only a small number of iterations are performed. In many cases, the approximate solution x^k is sufficiently accurate for $k \ll N$.

The number of CG iterations needed depends on the rate of convergence, which is very complex in general. A well-known estimate [9, 18] is given as follows:

$$\|e^k\|_A \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|e^0\|_A \quad (10)$$

where κ is the condition number [8] of A . For a parabolic PDE, with timestep size $O(h)$, $\kappa = O(h^{-1})$, where h is the mesh size of the discrete asset prices. The rates of convergence for Jacobi, Gauss–Seidel, and SD are $1 - O(h)$, whereas the rate for CG, based on the above error bound, is $1 - O(\sqrt{h})$, an order of magnitude improvement. SOR also has the same asymptotic convergence rate as CG, but it would require the knowledge of the optimal over-relaxation parameter. In practice, the CG error bound is often too pessimistic and the actual CG convergence is considerably much faster than the classical iterative methods.

Nonsymmetric Case

CG computes an optimal approximation x^k ($\|e^k\|_A$ is minimized) using short (two-term) recurrence update formulas for x^k and other quantities. Could one do the same for nonsymmetric matrices, such as the discretized Black–Scholes equation? Unfortunately, the answer is provably no [6]. Thus, when generalizing CG to the nonsymmetric case, one keeps some desirable properties of CG and sacrifices the others. There are many different possibilities, which yield numerous methods collectively known as the *Krylov subspace methods* [1]. Here, we describe two of these methods commonly used in practice.

The generalized minimal residual (GMRES) method [17] minimizes the norm of the residual vector:

$$\min_{x^k} \|b - Ax^k\|_2, \quad x^k \in x^0 + \mathcal{K}_k \quad (11)$$

As a result, the residual norm is nonincreasing. Furthermore, convergence is guaranteed for a wide class of matrices. However, it needs to store all the basis vectors for $\mathcal{K}_k$. Thus, work and storage increase

with iteration. A remedy is to restart GMRES every m iterations (m small constant). The convergence of the restarted GMRES, however, may stagnate, that is, $\|r^k\| \approx \|r^{k-1}\|$ for many iterations.

The other method is BiCGSTAB [23], which is derived from the biconjugate gradient (BCG) method [7]. BCG enforces a similar orthogonality condition on r^k as in equation (2) but allows two different Krylov subspaces to be used; more precisely,

$$r^k \perp \tilde{\mathcal{K}}_k \quad \text{and} \quad x^k \in x^0 + \mathcal{K}_k \quad (12)$$

where $\tilde{\mathcal{K}}_k = \text{span}\{\tilde{r}^0, A^T \tilde{r}^0, \dots, (A^T)^{k-1} \tilde{r}^0\}$ for some vector $\tilde{r}^0$. By enforcing the so-called bi-orthogonality condition, which makes the basis vectors $\{v^j\}$ for $\mathcal{K}_k$ and the basis vectors $\{w^j\}$ for $\tilde{\mathcal{K}}_k$ orthogonal to each other, two-term recurrence formulas for updating x^k and other quantities can be found. It has an advantage over GMRES in terms of storage. However, since it does not have the minimization property as GMRES, the residual norm can be very irregular as iterations continue. BiCGSTAB essentially is a “smooth” variant of BCG and its convergence is much more stabilized.

A more comprehensive overview of various Krylov subspace methods can be found in [1]. Similar to CG, the algorithms of the nonsymmetric methods involve only simple vector operations and matrix–vector products.

Preconditioning

CG methods would not have been so popular without the powerful technique called *preconditioning*, which can accelerate convergence drastically. Consider a nonsingular matrix M . The main idea is that the preconditioned system

$$M^{-1}Ax = M^{-1}b \quad (13)$$

is equivalent to equation (1) but the convergence of CG now depends on $M^{-1}A$ instead. The key is to construct a preconditioner M such that $\kappa(M^{-1}A) \ll \kappa(A)$ in order to obtain fast convergence. It is clear that if $M \approx A$, then $M^{-1}A \approx I$, which has condition number 1. On the other hand, M should be simple enough that the matrix–vector product by M^{-1} is easy to compute.

Generally speaking, it is difficult to determine what the optimal M is. Often it is problem specific. A

general class of preconditioners widely used in practice is called *incomplete LU* (ILU) factorization [14, 18]. A full LU would result in Gaussian elimination, which is expensive. An approximate LU factorization trades-off between efficiency and accuracy. Other effective preconditioners include multigrid [15, 22], domain decomposition [19], and sparse approximation inverse [2]. Once a preconditioner is chosen, to incorporate preconditioning into any CG algorithms only requires changing a few lines of code.

Application

CG methods have been used for pricing different options, for instance, American options [12], options with stochastic volatility [25], and options on Lévy driven assets [13]. Comparisons of SOR and CG methods for multiasset problem can be found in [21]. As an example, we consider pricing European options in a rather general exponential Lévy model (cf. [24] for more details in the special case of CGMY). The option value, $V(S, \tau)$, satisfies a partial integro-differential equation (*see Partial Integro-differential Equations (PIDEs)*), which is similar to the Black–Scholes equation [3] but with an extra integral term for the jump process:

$$\begin{aligned} \frac{\partial V}{\partial \tau} = & \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} - rV \\ & + \int_{-\infty}^{\infty} \nu(y) \left[V(Se^y, \tau) - V(S, \tau) \right. \\ & \left. - S(e^y - 1) \frac{\partial V}{\partial S} \right] dy \end{aligned} \quad (14)$$

where S is the value of the underlying asset, τ the time from expiry, σ the volatility, r the interest rate, and $\nu(y)$ a Lévy measure. Let $\{S_i\}_{i=1}^N$ be a set of discrete asset prices. Also, let $V^n = (V_1^n, \dots, V_N^n)$, where V_i^n is an approximation of $V(S_i, \tau^n)$ at time τ^n . Then a fully implicit finite difference discretization requires solving a linear system (1) at each timestep with $x = V^{n+1}$ and $b = V^n$. (The second order Crank–Nicolson discretization results in a similar matrix and right-hand side.) The matrix A can be written as a sum of two matrices: $A = L + B$, where L corresponds to the discretization of the differential term (which is similar to the Black–Scholes matrix) and B corresponds to the discretization of the integral term. While L is sparse, B is not; it has $O(N^2)$

nonzeros. Since A is dense because of B , it is not practical (time and storage) to form A explicitly even for moderate size N . In this case, Gaussian elimination and the classical iterative methods would not be easily applicable.

CG methods, on the other hand, can be used with ease. Note that B has a special convolution structure so that matrix–vector multiply can be computed efficiently using an FFT [4]. Thus, while A may not be available explicitly, the matrix–vector product by A can be computed, which is all CG methods require. In this problem, A is nonsymmetric and BiCGSTAB is used to solve the linear system [24]. Mesh-independent convergence is obtained for the infinite activity and finite variation case and there is a slight increase in iteration numbers for the infinite variation case. In both the cases, it shows an improvement of factor 4 in CPU times over another iterative method on the basis of a fixed point iteration.

Summary

CG methods generate “optimal” approximate solutions by performing simple vector operations and matrix–vector multiply. More importantly, by combining with the right choice of preconditioners, CG methods have been shown to be robust and efficient for solving option pricing PDEs. The discussion has been mainly focused on linear problems. In more general cases, one can apply preconditioned CG methods to the linearized equations from nonlinear problems without making any special modification.

Acknowledgments

The author was supported by the Natural Sciences and Engineering Research Council of Canada.

References

- [1] Barret, R., Berry, M., Chan, T.F., Demmel, J., Donato, J., Dongarra, J., Eijkhout, V., Pozo, R., Romine, C. & Van der Vorst, H. (1994). *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, 2nd Edition, SIAM, Philadelphia.
- [2] Benzi, M., Meyer, C.D. & Tuma, M. (1996). A sparse approximate inverse preconditioner for the conjugate gradient method, *SIAM Journal on Scientific Computing* **17**, 1135–1149.

-
- [3] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.
 - [4] d'Halluin, Y., Forsyth, P.A. & Vetzal, K. (2005). Robust numerical methods for contingent claims under jump diffusion processes, *IMA Journal on Numerical Analysis* **25**, 87–112.
 - [5] Duff, I.S., Erisman, A.M. & Reid, J.K. (1986). *Direct Methods for Sparse Matrices*, Oxford Press, UK.
 - [6] Faber, V. & Manteuffel, T. (1984). Necessary and sufficient conditions for the existence of a conjugate gradient method, *SIAM Journal on Numerical Analysis* **21**, 352–362.
 - [7] Fletcher, R. (1975). Conjugate gradient methods for indefinite systems, in *The Dundee Biennial Conference on Numerical Analysis, 1974*, G.A. Watson, ed, Springer-Verlag, New York, pp. 73–89.
 - [8] Golub, G. & Van Loan, C. (1996). *Matrix Computations*, The Johns Hopkins University Press, Baltimore.
 - [9] Greenbaum, A. (1997). *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia.
 - [10] Hackbusch, W. (1994). *Iterative Solution of Large Sparse Systems of Equations*, Springer-Verlag, New York.
 - [11] Hestenes, M.R. & Stiefel, E.L. (1952). Methods of conjugate gradients for solving linear systems, *Journal of Research of the National Bureau of Standards, Section B* **49**, 409–436.
 - [12] Khaliq, A.Q.M., Voss, D.A. & Kazmi, S.H.K. (2006). A linearly implicit predictor-corrector scheme for pricing American options using a penalty method approach, *Journal of Banking and Finance* **30**, 489–502.
 - [13] Matache, A.M., von Petersdorff, T. & Schwab, C. (2004). Fast deterministic pricing of options on Lévy driven assets, *Mathematical Modelling and Numerical Analysis* **38**, 37–72.
 - [14] Meijerink, J.A. & Van der Vorst, H. (1977). An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix, *Mathematics of Computation* **31**, 148–162.
 - [15] Oosterlee, C.W. (2003). On multigrid for linear complementarity problems with applications to American-style options, *Electronic Transactions on Numerical Analysis* **15**, 165–185.
 - [16] Peaceman, D. & Rachford, H. (1955). The numerical solution of elliptic and parabolic differential equations, *Journal of SIAM* **3**, 28–41.
 - [17] Saad, Y. & Schultz, M.H. (1986). GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM Journal on Scientific and Statistical Computing* **7**, 856–869.
 - [18] Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems*, 2nd Edition, SIAM, Philadelphia.
 - [19] Smith, B., Bjørstad, P. & Gropp, W. (1996). *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge.
 - [20] Strikwerda, J. (2004). *Finite Difference Schemes and Partial Differential Equations*, 2nd Edition, SIAM, Philadelphia.
 - [21] Tavella, D. & Randall, C. (2000). *Pricing Financial Instruments: The Finite Difference Method*, John Wiley & Sons, USA.
 - [22] Trottenbery, U., Oosterlee, C. & Schüller, A. (2001). *Multigrid*, Academic Press.
 - [23] Van der Vorst, H.A. (1992). Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems, *SIAM Journal on Scientific and Statistical Computing* **13**, 631–644.
 - [24] Wang, I.R., Wan, J.W.L. & Forsyth, P.A. (2007). Robust numerical valuation of European and American options under the CGMY process, *Journal of Computational Finance* **10**, 31–69.
 - [25] Zvan, R., Forsyth, P.A. & Vetzal, K.R. (1998). Penalty methods for American options with stochastic volatility, *Journal of Computational and Applied Mathematics* **91**, 199–218.

JUSTIN W.L. WAN

Multigrid Methods

Multigrid Basics

The multigrid iterative solution method has the unique potential of solving partial differential equations (PDEs) discretized with N^d unknowns in $O(N^d)$ work. This property forms the basis for efficiently solving very large computational problems. Initiated by Brandt [2], the development of multigrid has been particularly stimulated by the work done in computational fluid dynamics toward the end of the twentieth century. Introductions to multigrid can be found in [4, 15].

The insights and algorithms developed can be directly transferred to finance, that is, for solving the higher dimensional versions of convection-diffusion-reaction type PDE operators efficiently. These higher dimensional PDEs arise, for example, when dealing with stochastic volatility or with multiasset options under Black–Scholes dynamics. The aim is to solve these discrete PDE problems in just a few multigrid iterations within a split second.

Linear Multigrid

We would like to solve iteratively the discrete problem resulting from a PDE,

$$A_h u_h = f_h, \text{ on grid } G_h \quad (1)$$

for unknown u_h . For any approximation u_h^m , after iteration m , of the solution u_h , we denote the error by $e_h^m := u_h - u_h^m$, and the defect (or residual) by $d_h^m := f_h - A_h u_h^m$. Multigrid methods are motivated by the fact that many iterative methods, such as the well-known pointwise Gauss–Seidel iteration (PGS), have a smoothing effect on e_h^m . A smooth error can be well represented on a coarser grid, containing substantially fewer points, where its approximation is much cheaper. The defect equation,

$$A_h e_h^m = d_h^m \quad (2)$$

represents the error; it is equivalent to the original equation, since $u_h = u_h^m + e_h^m$. Departing from the insight of a smooth error, e_h^m , the idea is to use an

appropriate approximation, A_H , of A_h on a coarser grid, G_H (for instance, a grid with double the mesh size in each direction). The defect equation is then replaced by

$$A_H \widehat{e}_H^m = d_H^m \quad (3)$$

where $A_H : G_H \rightarrow G_H$, $\dim G_H < \dim G_h$ and A_H is invertible. As d_H^m and $\widehat{e}_H^m$ are grid functions on the coarser grid, G_H , we need two transfer operators between the fine and coarse grid. I_h^H is used to restrict d_h^m to G_H , and I_H^h is used to interpolate (or prolongate) the correction, $\widehat{e}_H^m$, back to G_h :

$$d_H^m := I_h^H d_h^m, \quad \widehat{e}_H^m := I_H^h \widehat{e}_H^m \quad (4)$$

This defines an iterative two-grid solution method, the two-grid correction scheme:

1. ν_1 smoothing steps on the fine grid:
 $\widehat{u}_h \leftarrow S^{\nu_1}(u_h^0, f_h)$;
2. computation of fine-grid residuals:
 $d_h := f_h - A_h \widehat{u}_h$;
3. restriction of residuals from fine to coarse:
 $d_H := I_h^H d_h$;
4. solution of coarse-grid problem: $A_H \widehat{e}_H = d_H$;
5. prolongation of corrections from coarse to fine:
 $\widehat{e}_h := I_H^h \widehat{e}_H$;
6. add correction to fine-grid approximation:
 $\widehat{u}_h \leftarrow \widehat{u}_h + \widehat{e}_h$;
7. ν_2 smoothing steps on the fine grid:
 $u_h^1 \leftarrow S^{\nu_2}(\widehat{u}_h, f_h)$.

Steps (1) and (7) are pre and postsmoothing steps, consisting of a few Gauss–Seidel iterations. Steps (2)–(6) form the “coarse-grid correction cycle”. In a well-converging two-grid method, it is not necessary to solve this coarse-grid defect equation exactly. Instead, one can replace $\widehat{e}_H^m$ by a suitable approximation. A natural way is to apply the two-grid idea again to the coarse-grid equation, now employing an even coarser grid than G_H . This can be done recursively; on each grid γ two-grid iteration steps are applied. With $\gamma = 1$, the multigrid V-cycle is obtained.

The smoothness of the error after Gauss–Seidel iterations depends on the discrete PDE under consideration. For nicely elliptic operators such as the Laplacian, errors will become smooth in all grid directions and grid coarsening can take place along every direction. For the convection-diffusion type equations, we obtain similar smooth errors if we

process the grid points according to the directions governed by the convective term [15]. The choice of coarse grid highly depends on the *smoothness of the error in the approximation*. We coarsen in those directions in which this error is smooth. Coarsening is simple for structured Cartesian grids, where one can remove every second grid point to obtain the coarse grid. For irregular grids, the resulting matrix A_h assists in determining the smoothness of the error. The matrix can accordingly be reduced algebraically (algebraic multigrid, AMG, see, e.g., [15]). In geometric multigrid, coarse grids are defined on the basis of a given fine grid, and coarse-grid corrections are computed from the PDE discretized on the coarse grids.

If N^d is the number of unknowns on G_h and N^d/β is the number of nodes on the coarse grid, then the multigrid algorithm requires $\mathcal{O}(N^d)$ storage and, for accuracy commensurable with discretization accuracy, $\mathcal{O}(N^d \log N^d)$ work, if $\gamma < \beta$. To get $\mathcal{O}(N^d)$ work, the multigrid cycles must be preceded by *nested iteration*, also called *full multigrid* [4, 15].

Multiaspect Options; Dealing with Grid Anisotropies in Multigrid

In this section, we discuss the numerical treatment with multigrid for the multiaspect Black–Scholes operator. Because this operator can be transformed to the multidimensional heat equation [11], we focus the discussion on the Laplacian operator for simplicity.

The major hindrance in the numerical solution of multidimensional PDEs is the so-called *curse of dimensionality*, which implies that with growth in the number of dimensions, we have an exponential growth in the number of grid points on tensor-product grids. Although we do not address this issue, in particular, we would like to stress that a way to handle dimensionality is the sparse-grid method [5, 14, 18]. One of the characteristics of these *sparse grids* is that they are essentially *nonequidistant* and, therefore, efficient multigrid solution methods for this type of grid are quite important.

We therefore consider here the discrete 2D Laplacian, discretized by finite differences on a grid with $h_y \gg h_x$, meaning that the matrix elements related to the x -derivative, $\mathcal{O}(h_x^{-2})$, are significantly larger than those for the y -derivative, $\mathcal{O}(h_y^{-2})$. If we apply

a PGS to this discrete operator, we find that its smoothing effect is very poor in the y -direction. The reason is that PGS has a smoothing effect only with respect to the “strong coupling” in the operator, in this case the x -direction. A multigrid method based on pointwise smoothing and grid coarsening in all directions will not converge well, as we coarsen along directions in which the error is nonsmooth.

An algorithmic improvement is to keep the pointwise relaxation for smoothing, but to *change the grid coarsening* according to the one-dimensional smoothness of errors: the coarse grid defined by doubling the mesh size only in that direction in which the errors are smooth will result in an efficient multigrid method. Figure 1 shows an example of semicoarsening along the x -axis.

A second successful approach is to keep the grid coarsening in all directions, but to change the smoothing procedure from a pointwise to *linewise* iteration. Line relaxations are block iterations in which each block of unknowns corresponds to a line. This smoother generates smooth error in all grid directions in the case of grid anisotropies.

These two strategies for excellent convergence, that is, to maintain standard coarsening and change the smoother, or to keep the pointwise smoothing procedure but adapt the coarsening, remain valid for higher dimensional problems. In a 3D problem, this implies that *planewise relaxation* should be employed (in combination with standard coarsening), in which all unknowns lying in the plane of strongly coupled unknowns are relaxed simultaneously. In contrast to line relaxation, which leads to tridiagonal matrices, in plane relaxation we need to solve a discrete 2D problem. A multigrid treatment of high-dimensional PDEs in finance based on hyperplane relaxation has been proposed in [13], while the use of pointwise

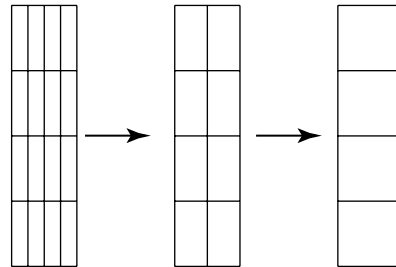


Figure 1 An example of x -semicoarsening using three grids

smoothing and coarsening the grid simultaneously along all dimensions where the errors are strongly coupled by *simultaneous partial grid coarsening* has been employed in [18, 19], until the coarse-grid problem is isotropic to the point where full coarsening is feasible. The resulting multigrid methods are highly efficient.

American Options; Multigrid Treatment of Nonlinear Problems

Next, we discuss multigrid methods for the computation of the value of an American-style option. In [17], it was shown that for American-style options the theory related to free boundary problems, as it was developed in the 1970s [1, 8], applies. It is possible to rewrite the arising free boundary problem as a linear complementarity problem (LCP), of the form

$$Au \leq f_1 \quad \mathbf{x} \in \Omega \quad (5)$$

$$u \geq f_2 \quad \mathbf{x} \in \Omega \quad (6)$$

$$(u - f_2)(Au - f_1) = 0 \quad \mathbf{x} \in \Omega \quad (7)$$

The LCP formulation is beneficial for iterative solution, since the unknown boundary does not appear explicitly and can be obtained in a postprocessing step. The LCP problem is, however, nonlinear, which implies that we have to generalize the linear multigrid algorithm to the nonlinear situation. We can distinguish in solutions of LCPs a so-called active region from an inactive region. In the active region, constraint (6) holds with equality sign, whereas in the inactive region, constraint (5) is valid with equality sign.

Nonlinear Multigrid

The fundamental idea of multigrid for *nonlinear PDEs* of the form

$$N_h u_h = f_h \quad (8)$$

is the same as that for linear equations. First, the errors in the solution have to be smoothed so that they can be approximated on a coarser grid. In the nonlinear case, the fine-grid defect equation is given by

$$N_h(\bar{u}_h^m + e_h^m) - N_h \bar{u}_h^m = d_h^m \quad (9)$$

where $\bar{u}_h^m$ is the approximation of the solution after relaxation in the i th multigrid cycle, e_h^m is the error and d_h^m is the corresponding defect. This equation is approximated on a coarse grid by

$$N_H(\bar{u}_H^m + \hat{e}_H^m) - N_H \bar{u}_H^m = d_H^m \quad (10)$$

Not only is the defect, d_h^m , transferred to the coarse grid by some restriction operator I_h^H but also the relaxed approximation $\bar{u}_h^m$ itself by a restriction operator $\hat{I}_h^H$.

However, as in the linear case, only the coarse-grid corrections, $\hat{e}_H$, are interpolated back to the fine grid, where the fine-grid errors are smoothed again. This forms the basis of the well-known full approximation scheme (FAS) [2]. The nonlinearity of the problems enters in the smoothing operators. If N_h and N_H are linear operators, the FAS method is equivalent to the linear multigrid scheme. For many problems, however, the nonlinearity can also be handled globally, resulting in a sequence of linear problems that can be solved efficiently with linear multigrid.

Multigrid for Linear Complementarity Problems

In 1983, Brandt and Cryer [3] proposed a multigrid method for LCPs arising from free boundary problems. The algorithm is based on the projected successive over-relaxation (SOR) method [7] and is called the *projected full approximation scheme (PFAS)* in [3]. PFAS has been successfully used in the financial community for American options with stochastic volatility in [6, 12].

For the smoothing method in PFAS, one employs a projected version of the PGS, consisting of two partial steps per unknown: In a first step, a Gauss–Seidel iteration is applied to equation (5) at (x_i, y_j) with equality sign. In the second partial step, the solution at (x_i, y_j) is projected, so that constraint (6) is satisfied,

$$\begin{aligned} \bar{u}^m(x_i, y_j) &= \max\{f_2(x_i, y_j), \hat{u}(x_i, y_j)\} \\ \forall (x_i, y_j) &\in G_h \end{aligned} \quad (11)$$

where $\bar{u}^m$ denotes the unknown at (x_i, y_j) after PGS and $\hat{u}$ the unknown after the Gauss–Seidel iteration. A linewise variant of PGS has been applied in [6].

The following LCP holds for $e_h^m := u_h - \bar{u}_h^m$:

$$\begin{aligned} A_h e_h^m &\leq d_h^m \quad \mathbf{x} \in \Omega \\ e_h^m + \bar{u}_h^m &\geq f_{2,h} \quad \mathbf{x} \in \Omega \\ (e_h^m + \bar{u}_h^m - f_{2,h})(A_h e_h^m - d_h^m) &= 0 \quad \mathbf{x} \in \Omega \end{aligned} \quad (12)$$

with defect: $d_h^m = f_{1,h} - A_h \bar{u}_h^m$. A smooth error, e_h^m , can be approximated on a coarse grid without any essential loss of information. The LCP coarse-grid equation for the coarse-grid approximation of the error $\widehat{e}_H^m$ is therefore defined in PFAS by:

$$\begin{aligned} A_H \widehat{e}_H^m &\leq I_H^H d_h^m \\ \widehat{e}_H^m + \widehat{I}_H^H \bar{u}_h^m &\geq f_{2,H} \\ (\widehat{e}_H^m + \widehat{I}_H^H \bar{u}_h^m - f_{2,H})(A_H \widehat{e}_H^m - I_H^H d_h^m) &= 0 \end{aligned} \quad (13)$$

For LCPs, we need to choose “constraint preserving” restriction operators that do not mix information from active and inactive regions on coarse grids. Further, the bilinear interpolation operator I_H^h is applied only to unknowns on the “active” points [3].

We finally mention that in [10], another multigrid variant for LCPs, the so-called monotone multigrid method, has been presented, used in finance in [9].

Multigrid as a Preconditioner

Multigrid as a preconditioner is particularly interesting for *robustness*. An argument for combining multigrid with an acceleration technique is that problems become more and more complex if we treat real-life applications. The fundamental idea of multigrid, to reduce the high-frequency error components by smoothing and to take care of the low frequency error by coarse-grid correction, does not always work optimally if straightforward multigrid approaches are used. In such situations, the combination with Krylov subspace methods, such as conjugate gradient, generalized minimal residual (GMRES), or BiCGSTAB, have the potential to give a substantial convergence acceleration. Often, sophisticated multigrid components may also lead to very satisfactory convergence factors, but they can be difficult to realize and implement.

From the multigrid point of view, multigrid as a preconditioner can also be interpreted as an acceleration of multigrid by iterant recombination. This interpretation easily allows generalizations, for example, to nonlinear problems and to LCPs. Let u_h^0 be an initial approximation for solving $A_h u_h = f_h$, and $d_h^0 = f_h - A_h u_h^0$ its defect. The *Krylov subspace*, K_h^m , is defined by $K_h^m := \text{span}[d_h^0, A_h d_h^0, \dots, A_h^{m-1} d_h^0]$. This subspace can also be represented by $\text{span}[u_h^0 - u_h^m, u_h^1 - u_h^m, \dots, u_h^{m-1} - u_h^m]$, where the u_h^m are previous approximations to the solution. To find an improved approximation $u_{h,\text{acc}}^m$, we now consider a linear combination of the $m+1$ latest approximations:

$$u_{h,\text{acc}}^m = u_h^m + \sum_{i=1}^m \bar{\alpha}_i (u_h^{m-i} - u_h^m) \quad (14)$$

In order to obtain an improved approximation $u_{h,\text{acc}}^m$, the parameters $\bar{\alpha}_i$ are determined in such a way that the defect $d_{h,\text{acc}}^m$ is minimized, for example, with respect to the l_2 -norm $\|\cdot\|_2$. This is a classical defect minimization problem [16]. This technique was generalized to LCPs in [12], where PFAS was used as the method whose iterants were recombined.

Acknowledgments

This research has been partially supported by the Dutch government through the national program BSIK: knowledge and research capacity, in the ICT project BRICKS (<http://www.bsik-bricks.nl>), theme MSV1.

References

- [1] Bensoussan, A. & Lions, J.L. (1982). *Applications des Équations Variationnelles en Contrôle Stochastique*, North-Holland, Dunot, Amsterdam, English Translation 1978.
- [2] Brandt, A. (1977). Multi-level adaptive solutions to boundary-value problems, *Mathematics of Computation* **31**, 333–390.
- [3] Brandt, A. & Cryer, C.W. (1983). Multigrid algorithms for the solution of linear complementarity problems arising from free boundary problems, *SIAM Journal on Scientific Computing* **4**, 655–684.
- [4] Briggs, W.L., Emden Henson, V. & McCormick, S.F. (2000). *A Multigrid Tutorial*, 2nd Edition, SIAM, Philadelphia, PA.
- [5] Bungartz, H.J. & Griebel, M. (2004). Sparse grids, *Acta Numerica* **13**, 147–269.

-
- [6] Clarke, N. & Parrot, K. (1999). Multigrid for American option pricing with stochastic volatility, *Applied Mathematics and Finance* **6**, 177–197.
 - [7] Cryer, C.W. (1971). The solution of a quadratic programming problem using systematic overrelaxation, *SIAM Journal on Control* **9**, 385–392.
 - [8] Friedman, A. (1982). *Variational Principles and Free Boundary Problems*, Wiley, New York.
 - [9] Holtz, M. & Kunoth, A. (2007). B-spline based monotone multigrid methods. *SIAM Journal on Numerical Analysis* **45**, 1175–1199.
 - [10] Kornhuber, R. (1994). Monotone multigrid methods for elliptic variational inequalities I, *Applied Numerical Mathematics* **69**, 167–184.
 - [11] Kwok, Y.K. (1998). *Mathematical Models of Financial Derivatives*, 2nd Edition, Springer, Singapore.
 - [12] Oosterlee, C.W. (2003). On multigrid for linear complementarity problems with application to American-style options. *Electronic Transactions on Numerical Analysis*, **15**, 165–185.
 - [13] Reisinger, C. & Wittum, G. (2004). On multigrid for anisotropic equations and variational inequalities, *Computers and Visual Science*, **7**, 189–197.
 - [14] Reisinger, C. & Wittum, G. (2007). Efficient hierarchical approximation of high-dimensional option pricing problems, *SIAM Journal On Scientific Computing* **29**, 440–458.
 - [15] Trottenberg, U., Oosterlee, C.W. & Schüller, A. (2000). *Multigrid*, Academic Press, London.
 - [16] Washio, T. & Oosterlee, C.W. (1997). Krylov subspace acceleration for nonlinear multigrid schemes. *Electronic Transactions on Numerical Analysis* **6**, 271–290.
 - [17] Wilmott, P., Dewynne, J. & Howison, S. (1993). *Option Pricing*, Oxford Financial Press.
 - [18] bin Zubair, H., Leentvaar, C.C.W. & Oosterlee, C.W. (2007). Efficient d -multigrid preconditioners for sparse-grid solution of high dimensional partial differential equations, *International Journal Of Computer Mathematics* **84**, 1129–1146.
 - [19] bin Zubair, H., Oosterlee, C.W. & Wienands, R. (2007). Multigrid for high dimensional elliptic partial differential equations on non-equidistant grids, *SIAM Journal On Scientific Computing* **29**, 1613–1636.

CORNELIS W. OOSTERLEE

Finite Element Methods

The finite element method (FEM) has been invented by engineers around 1950 for solving the partial differential equations arising in solid mechanics; the main idea was to use the principle of virtual work for designing discrete approximations of the boundary value problems. The most popular reference on FEM in solid mechanics is the book by Zienckiewicz [14]. Generalizations to other fields of physics or engineering have been done by applied mathematicians through the concept of variational formulations and weak forms of the partial differential equations (see, e.g. [4]).

In finance, partial differential equations (PDEs) or partial integro-differential equations (PIDEs) may be used for option pricing. For approximating their solutions, at least four classes of numerical methods can be used:

- *Finite difference methods* are by far the simplest, except when mesh adaptivity is required in which case it is rather difficult to control the numerical error.
- *Finite volume methods* are not really natural, because these methods are better suited for hyperbolic PDEs. In finance, they may be useful, for example, for Asian options when the PDE becomes close to hyperbolic near maturity.
- *Spectral methods* are Galerkin methods with Fourier series of high degree polynomials. They are ideal when the coefficients of the PDE are constant, which is rarely the case in financial engineering. For a very efficient adaptation of spectral methods to finance, see [10].
- *Finite element methods* seem at first glance unnecessarily complex for finance where a large class of problems are one dimensional in space. Yet, they are very flexible for mesh adaptivity and the implementation difficulties are only apparent.

A Simple Example

Take the simplest case, the Black–Scholes model for a put option with a strike K and a maturity T . The price

of the underlying asset is assumed to be a geometric Brownian motion

$$dS_t = S_t(\mu dt + \sigma dW_t) \quad (1)$$

and the volatility σ is allowed to depend on S_t and t . If the volatility function satisfies suitable regularity conditions, the Black–Scholes formula gives the option's price at time $t < T$:

$$P_t = \mathbb{E}^*(e^{-r(T-t)}(K - S_T)^+ | \mathcal{F}_t) \quad (2)$$

where $\mathbb{E}^*(| \mathcal{F}_t)$ stands for the conditional expectation with respect to the risk neutral probability. It can be proved that with $\sigma = \sigma(S_t, T - t)$, then $P_{T-t} = u(S_{T-t}, T - t)$, where u is the solution of

$$\begin{cases} \partial_t u - \frac{\sigma^2 x^2}{2} \partial_{xx} u - rx \partial_x u + ru = 0 \\ \text{for } x > 0, 0 < t \leq T \\ u(x, 0) = (K - x)^+ \end{cases} \quad (3)$$

The variational formulation of equation (3) consists of finding a continuous function u defined on the time interval $[0, T]$ with value in a Hilbert space V (see equation (5)) such that

$$\begin{aligned} \frac{d}{dt}(u, w) + a_t(u, w) &= 0 \\ \text{for a.a. } t \in (0, T) \quad \forall w \in V \\ u(x, 0) &= (K - x)^+ \end{aligned} \quad (4)$$

where

$$\begin{aligned} V &= \left\{ v \in L^2(\mathbb{R}_+) : x \frac{dv}{dx} \in L^2(\mathbb{R}_+) \right\} \\ a_t(u, w) &= \left(\frac{\sigma^2 x^2}{2} \partial_{xx} u, \partial_x w \right) \\ &\quad + (x \partial_x u, (\sigma^2 + x \sigma \partial_x \sigma - r)w) + (ru, w) \\ (u, w) &= \int_0^\infty u(x)w(x) dx \end{aligned} \quad (5)$$

Problem (4) is obtained by multiplying equation (3) by $w(x)$, integrate on $\mathbb{R}^+$, and integrate by part the term containing $\partial_{xx} u$. The precise meaning of equation (4) and the conditions for equation (4) to have a unique solution are subtle. The choice of the space V is dictated by the following: the integrals in equation (4) must exist, and the bilinear form a_t must be continuous on $V \times V$ uniformly in time and

satisfy Gårding's coercivity inequality, that is, that there exist constants c and λ , $c > 0$ and $\lambda \geq 0$ such that

$$a_t(v, v) \geq c\|v\|_V^2 - \lambda\|v\|_{L^2(\mathbb{R}_+)}^2 \quad \forall v \in V \quad (6)$$

Proposition 1 *If σ is a continuous function such that for all x, t , $\sigma(x, t) \in [\sigma_m, \sigma_M]$, $\sigma_m > 0$, and if $x \mapsto x\sigma$ is Lipschitz continuous with a Lipschitz constant independent of t , then equation (4) has a unique solution.*

The simplest finite element methods are discrete versions of equation (4) obtained by replacing V with a finite dimensional space V_h , that is, finding a continuous function u_h defined on the time interval $[0, T]$ with value in the finite dimensional space $V_h \subset V$ such that

$$\frac{d}{dt}(u_h, w_h) + a_t(u_h, w_h) = 0 \quad \text{for a.a. } t \in (0, T) \quad (7)$$

If $\{w^i\}_1^N$ is a basis of V_h , then equation (6) is equivalent to

$$\begin{aligned} \frac{d}{dt} \left(\sum_1^N u_j w^j, w^i \right) + a_t \left(\sum_1^N u_j w^j, w^i \right) &= 0 \\ \forall i &= 1, \dots, N \end{aligned} \quad (8)$$

which is a system of differential equations

$$\begin{aligned} B \frac{dU}{dt} + A(t)U &= 0 \quad \text{where } B_{ij} = (w^j, w^i) \\ A_{ij}(t) &= a_t(w^j, w^i) \end{aligned} \quad (9)$$

A discrete time stepping scheme has still to be applied to equation (9), for instance, the Euler implicit scheme

$$B \frac{U^{m+1} - U^m}{\delta t_m} + A^{m+1}U^{m+1} = 0 \quad (10)$$

where $A^m = A(t_m)$. The easiest choice is to take V_h as a space of piecewise linear functions (linear finite elements), that is,

$$\begin{aligned} V_h &= \{v_h \text{ continuous, linear on each subinterval} \\ &\quad [x_{i-1}, x_i], v_h(L) = 0\} \end{aligned}$$

where $(x_i)_{i=1, \dots, N}$ is an increasing sequence in the interval $[0, L] \approx \mathbb{R}^+$, with $x_1 = 0$ and $x_N = L$. Then a good choice for w^i is to be the (hat) function of V_h , which is equal to one at x_i and zero at x_j , $j \neq i$ (see [2]). With this basis, called the *nodal basis*, the integrals B_{ij} and A_{ij} can be computed exactly.

It is easy to show that the matrices A and B are tridiagonal so that a resolution of the linear system (10) at each time step is best done with a Gaussian elimination method or an LU factorization. In the end the computational complexity is the same as for a finite difference method with either an explicit or an implicit time scheme.

Convergence in the V -norm can be proved to be of the order of $h + \delta t$ ($h^2 + \delta t$ in the L^2 -norm); it is possible to improve it by using Crank–Nicolson scheme and higher order polynomials instead of the linear ones. Most interesting is the following *a posteriori* estimate:

Proposition 2

$$\begin{aligned} [[u - u_{h, \delta t}]](t_n) &\leq c(u_0)\delta t \\ &+ c \frac{\mu}{\sigma_m^2} \left(\sum_{m=1}^n \eta_m^2 + (1 + \rho)^2 \max(2, 1 + \rho) \right. \\ &\quad \times \sum_{m=1}^n \frac{\delta t_m}{\sigma_m^2} \prod_{i=1}^{m-1} (1 - 2\lambda \delta t_i) \sum_{\omega \in \mathcal{T}_{mh}} \eta_{m, \omega}^2 \left. \right)^{1/2} \end{aligned} \quad (11)$$

where μ is the continuity constant of a_t , $\rho = \max_{2 \leq n \leq N} \delta t_n / \delta t_{n-1}$, $\mathcal{T}_{mh}$ is the partition of $(0, L)$ in small intervals ω at time t_m and

$$\begin{aligned} \eta_m^2 &= \delta t_m e^{-2\lambda t_{m-1}} \frac{\sigma_m^2}{2} |u_h^m - u_h^{m-1}|_V^2 \\ \eta_{m, \omega} &= \frac{h_\omega}{x_{\max}(\omega)} \left\| \frac{u_h^m - u_h^{m-1}}{\delta t_m} - r x \frac{\partial u_h^m}{\partial x} + r u_h^m \right\|_{L^2(\omega)} \end{aligned} \quad (12)$$

and h_ω is the diameter of ω .

Note that everything is known after the computation of u_h ; the mesh can then be adapted by using the error indicators η_m and $\eta_{m, \omega}$. Figure 1 shows the computed *a posteriori* error and the actual error for

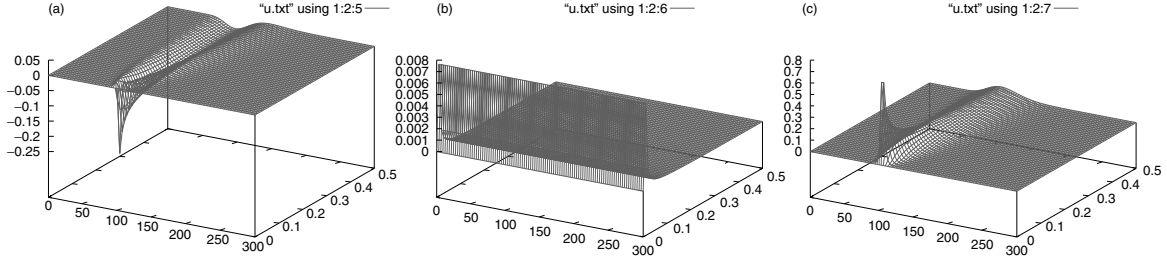


Figure 1 (a) The graph displays the error between the computed solution on a uniform mesh and one computed with Black-Scholes formula. (b) The graph shows $\rho_{\delta t}$; (c) the graphs show $\eta_{n,\omega}$ as a function of x and t . The parameters are $K = 100$, $T = 0.5$, $r = 0$, $\sigma = 0.3$, 50 time steps, and 100 mesh points

a vanilla put with constant volatility and comparison with the Black-Scholes analytic formula.

Stochastic Volatility Models

We now discuss a more involved stochastic volatility model where $dS_t = S_t(\mu dt + \sigma_t dW_t)$, with $\sigma_t = f(Y_t)$ and Y_t is a mean-reverting Ornstein-Uhlenbeck process

$$dY_t = \alpha(m - Y_t) dt + \beta d\hat{Z}_t \quad (13)$$

and $\hat{Z}_t = \rho W_t + \sqrt{1 - \rho^2} Z_t$, W_t and Z_t are two uncorrelated Brownian motions. In the Stein-Stein model, see [12], $f(Y_t) = |Y_t|$.

If $P(S, y, T - t)$ is the pricing function of a put option, it is convenient to introduce

$$u(S, y, t) = P(S, y, T - t) e^{-(1-\eta)(y-m)^2/2v^2} \quad (14)$$

where η is any parameter between 0 and 1 and $v^2 = \beta^2/(2\alpha)$.

A clever financial argument (see [7]) shows that when $\rho = 0$ for simplicity, the new unknown u satisfies, for all $x > 0$, $y \in \mathbb{R}$, $t > 0$

$$\begin{cases} \frac{\partial u}{\partial t} - \frac{1}{2}y^2x^2\frac{\partial^2 u}{\partial x^2} - \frac{1}{2}\beta^2\frac{\partial^2 u}{\partial y^2} - rx\frac{\partial u}{\partial x} + e\frac{\partial u}{\partial y} \\ \quad + fu = 0 \\ u(x, y, 0) = (K - x)^+ \quad \text{for } x > 0, y \in \mathbb{R} \end{cases} \quad (15)$$

whereof

$$e = -(1 - 2\eta)\alpha(y - m) + \beta\gamma$$

$$f = r + 2\frac{\alpha^2}{\beta^2}\eta(1 - \eta)(y - m)^2$$

$$+ 2(1 - \eta)\frac{\alpha}{\beta}(y - m)\gamma - \alpha(1 - \eta) \quad (16)$$

and γ is a *risk premium factor*, possibly a function of x, y, t ; see [1] for the details.

This parabolic equation is degenerate near the axis $y = 0$ because the coefficient in front of $x^2\frac{\partial^2 u}{\partial x^2}$ vanishes on $y = 0$. Following [1], see also [2, 3], let us consider the weighted Sobolev space V on $Q = \mathbb{R}^+ \times \mathbb{R}$:

$$V = \left\{ v : v\sqrt{1 + y^2}, \frac{\partial v}{\partial y}, x|y|\frac{\partial v}{\partial x} \in L^2(Q) \right\} \quad (17)$$

which is a Hilbert space such that all the properties needed by the variational arguments sketched earlier hold.

Theorem 1 *If the data are smooth (see [1, 3]) and if*

$$r(t) \geq r_0 > 0, \quad \alpha^2 > 2\beta^2, \quad 2\alpha^2\eta(1 - \eta) > \beta^2 \quad (18)$$

then equation (15) has a unique weak solution.

Finite Element Solution

As usual, we localize the problem in $\Omega := (0, L_x) \times (-L_y, L_y)$. No boundary condition is needed on the axis $x = 0$ since the PDE is degenerate there. Note also that u is expected to tend to zero as x tends to infinity. For large y , no arguments seem to give a boundary condition. However, if L_y is chosen such that $\alpha(L_y \pm m) \gg \beta^2$, then for $y \sim \pm L_y$, the coefficient of the advection term in the y direction, that is, $\alpha(y - m)$, is much larger in absolute value

than the coefficient of the diffusion in the y direction, that is, $\frac{\beta^2}{2}$. Furthermore, near $y = \pm L_y$, the vertical velocity $\alpha(y - m)$ is directed outward Ω . Therefore, the error caused by an artificial condition on $y = \pm L_y$ is damped away from the boundaries, and localized in boundary layers whose width is of the order of $\frac{\beta^2}{\alpha y}$. We may thus apply a Neumann condition.

Then one must write the problem in variational form: find u in V_0 the subset of V of functions that are zero at $x = L_x$ and

$$\begin{cases} \frac{d}{dt}(u, w) + a(u, w) = 0 & \forall w \in V_0 \\ u(x, y, 0) \text{ given} \end{cases} \quad (19)$$

with

$$\begin{aligned} a(u, w) = & \int_{\Omega} \frac{y^2 x^2}{2} \partial_x u \partial_x w + \frac{\beta^2}{2} \partial_y u \partial_y w \\ & + (y^2 - r)x \partial_x u + (e + \beta \partial_y \beta) \partial_y u + fu)w \end{aligned} \quad (20)$$

Now Ω is triangulated into a set of nonintersecting triangles $\{\omega_k\}_1^K$ in such a way that no vertices lie in the middle of an edge of another triangle. V_0 is approximated by V_{0h} the space of continuous piecewise linear functions on the triangulation, which vanish at $x = L_x$.

The functions of V_{0h} are completely determined by their values at the vertices of the triangles, so if we denote by w^i the function in V_{0h} , which takes the value zero at all vertices except the i th one where it is one, then $\{w^i\}_1^N$ is a basis of V_{0h} if N is the number of vertices not on $x = L_x$ (it is convenient for the presentation to number the nodes on the boundary last).

As in the one-dimensional case

$$u_h^m(x, y) = \sum_1^N u_i^m w^i(x, y) \quad (21)$$

is an approximation of $u(t_m)$ if the u_i^{m+1} are solutions of

$$B \frac{U^{m+1} - U^m}{\delta t_m} + AU^{m+1} = 0 \quad (22)$$

with $B_{ij} = (w^j, w^i)$ and $A_{ij} = a(w^j, w^i)$.

Numerical Implementation

Integrals of polynomial expressions of x, y can be evaluated exactly on triangles (formula (4.19) in [2]). The difficulty is in the strategy to solve the linear systems (22). Multigrid seems to give the best results in term of speed [11], but the computer program may not be so easy to write for an arbitrary triangulation. Since the number of vertices are not likely to exceed a few thousands, fast multifrontal libraries such as SuperLU [6] can be used with a minimum programming effort. Table 1 shows the gain in time over more classical algorithms.

Alternatively one could also use a preconditioned conjugate gradient-like iterative method (see **Conjugate Gradient Methods**). Another way is to use FreeFem++ [9], which is a general two-dimensional PDE solver particularly well suited to these option pricing models, where the coefficients are involved functions and one may want to try to change them often. Results for equation (22) obtained with the following FreeFem++ code are shown in Figure 2.

A similar program has been developed in 3D by S. Delpino: freefem3d. However, mesh refinement is not yet implemented. A basket option has been computed for illustration (Figure 3). The script that drives the program is self-explanatory and gives the values used.

eqf12-005

Table 1 Comparison of CPU time for the LU algorithm and superLU for a product put option on a uniform mesh and 200 step time (computed by N. Lantos)

Mesh size	Gauss-LU (s)	Relative error (%)	SuperLU (s)	Relative error (%)
101 × 101	10.094	3.532	2.39	3.076
126 × 126	14.547	1.338	4.016	1.797
151 × 151	22.313	0.751	6.203	0.489
176 × 176	31.985	1.131	8.735	0.790
201 × 201	43.938	0.432	12.109	0.670

```

real T=1, K=100, r=0.05, alpha=1, beta=1 , m=0.2 , gamma=0, eta=0.5;
int Lx=800, Ly=3, Nmax=50;
real scale=100, dt =T/Nmax;

mesh th = square(50,50,[x*Lx,(2*y-1)*Ly*scale]);
fespace Vh(th,P1);

func u0 = max(K-x,0.)*exp(-(1-eta)*(y/scale-m)^2 * alpha/(beta^2));
func e = beta*gamma -(1-2*eta)*alpha*(y/scale-m);
func f = r + 2*(alpha/beta)^2 * eta*(1-eta)*(y/scale-m)^2
      + 2*(1-eta)*(alpha/beta)*(y/scale-m)*gamma - alpha*(1-eta);

Vh uold=u0,u,v;

problem stein(u,v,init=n) = int2d(th)( u*v*(f+1/dt)
      + dx(u)*dx(v) *(x*y/scale)^2/2 + dy(u)*dy(v) *(beta*scale)^2/2
      + dx(u)*v *x*((y/scale)^2-r) + dy(u)*v *e*scale
      ) - int2d(th)(uold*v/dt) + on(2,u=0) ;

for (int n=0; n<Nmax ; n++){ stein; uold=u; };
// call to keyword adaptmesh not shown..

double N = 25;          double L=200.0;    double T=0.5;
double dt = T / 15 ; double K=100;    double r = 0.02;
double s1 = 0.3;        double s2 = 0.2;   double s3 = 0.25;
double q12 = -0.2*s1*s2;    double q13 = -0.1*s1*s3;
double q23 = -0.2*s2*s3;    double s11 = s1*s1/2;
double s22=s2*s2/2;        double s33=s3*s3/2;

vector n = (N,N,N);
vector a = (0,0,0);
vector b = (L,L,L);
mesh M = structured(n,a,b);
femfunction uold(M) = max(K-x-y-z,0);
femfunction u(M);

double t=0; do {
    solve(u) in M cg(maxiter=900,epsilon=1E-10)
    {
pde(u) (1./dt+r)*u-dx(s11*x*x*dx(u))-dy(s22*y*y*dy(u))-dz(s33*z*z*dz(u))
      - dx(q12*x*y*dy(u)) - dx(q13*x*z*dz(u)) - dy(q23*y*z*dz(u))
      - r*x*dx(u) - r*y*dy(u) - r*z*dz(u) = uold/dt;
      dnu(u)=0 on M;
    }; t = t + dt;
} while(t<T);
save(medit,"u",u,M);          save(medit,"u",M);

```

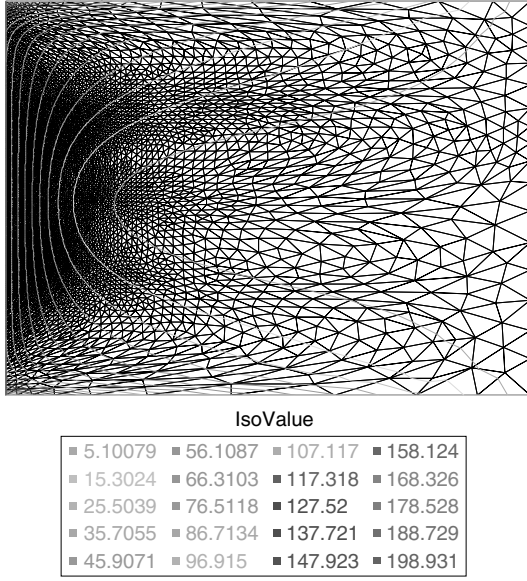



Figure 2 Solution of equation (15) for a put with maturity $T = 1$, strike $K = 100$ with $r = 0.02$, $\beta = \alpha = 1$, and $m = 0.2$. We have used the automatic mesh adaptivity of FreeFem++; the contours and the triangulation are shown

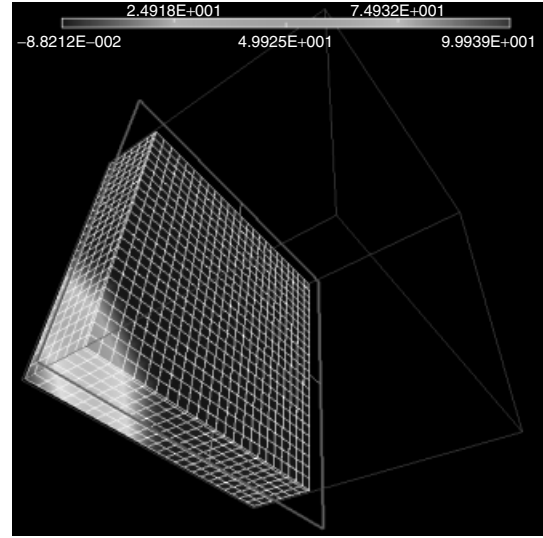


Figure 3 A put on a basket of three assets computed with freefem3d [5] with maturity $T = 0.5$, strike $K = 100$, and so on (see the listing). The visualization is done with medit also public domain [8]

Asian Calls

For the financial modeling of Asian options, we refer to [13] and the references therein.

We consider an Asian put with fixed strike whose payoff is $P_0(S, y) = (y - K)_+$, calling y the average value of the asset in time, $y = 1/t \int_0^t S(\tau) d\tau$. The price of the option is found by solving for all

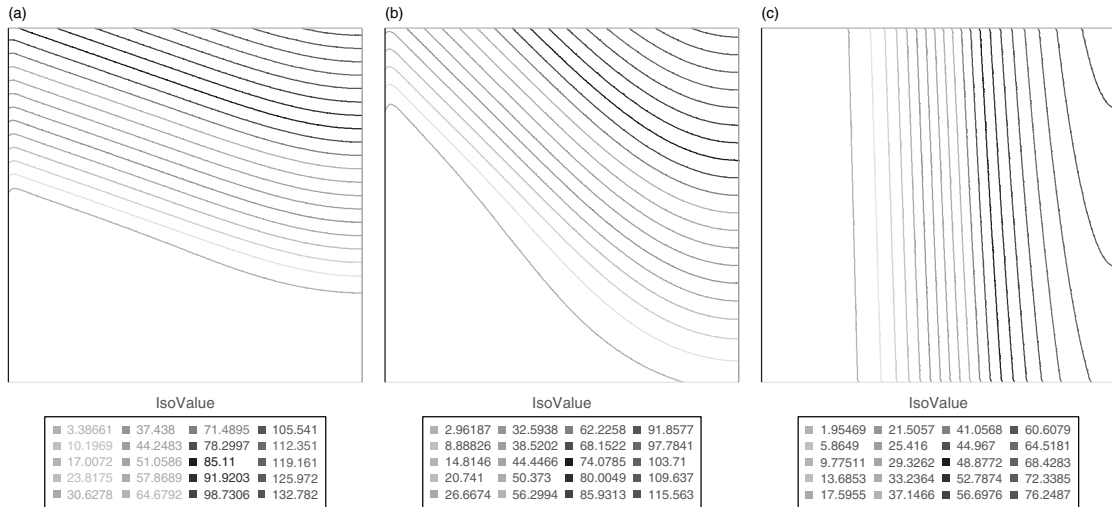


Figure 4 Solution of equation (24) for an Asian option of maturity $T = 4$ and strike $K = 100$ when the interest rate is 0.03 and the volatility 0.3. The figure shows the contours at times 0.96 (a), 1.96 (b), and 3.96 (c)

$$\{x, y, t\} \in \mathbb{R}^+ \times \mathbb{R}^+ \times (0, T]$$

$$\begin{aligned} \partial_t u - \partial_x \left(\frac{\sigma^2 x^2}{2} \partial_x u \right) + (\sigma^2 - r)x \partial_x u \\ + \frac{y-x}{T-t} \partial_y u + ru = 0 \end{aligned} \quad (23)$$

$$\begin{aligned} u(x, y, 0) = (y - K)^+, \quad \partial_x u|_{+\infty, y, t} \\ \approx \begin{cases} \frac{t}{T} e^{-rt} & \text{if } y < \frac{KT}{T-t} \\ \frac{1-re^{-rt}}{Tr} & \text{otherwise} \end{cases} \end{aligned} \quad (24)$$

The difficulties of this problem is that the second-order part of the differential operator is incomplete and the convective velocity $v = ((\sigma^2 - r)x, \frac{y-x}{T-t})^T$ tends to infinity as $t \rightarrow T$; so great care must be taken for the upwinding of these first-order terms. One of the best upwinding is the characteristic finite element method, whereby

$$\partial_t u + v \cdot \nabla u|_{x, y, t_{m+1}} \approx \frac{1}{\delta t} (u^{m+1}(q) - u^m(Q))$$

with

$$Q = q - \delta t v \left(q - v(q) \frac{\delta t}{2} \right) \quad \text{where } q = (x, y)^T \quad (25)$$

The discontinuous P^2 reconstruction with characteristic convection of the gradient and of the value at the center of gravity is applied. The domain is localized to $(0, L) \times (0, L)$, with $L = 250$, that is, greater than twice the contract price $K = 100$, in this example; the interest rate is $r = 3\%$ and the volatility $\sigma = 0.3$. The square domain is triangulated by a uniform 50×50 mesh and we have taken 50 time steps over the four-years period of this example ($T = 4$). The results are shown in Figure 4 at time $t = 0.96, 1.96$, and 3.96 .

References

- [1] Achdou, Y. & Tchou, N. (2002). Variational analysis for the Black and Scholes equation with stochastic volatility, *M2AN Mathematical Modelling and Numerical Analysis* **36**(3), 373–395.
- [2] Achdou, Y. & Pironneau, O. (2005). *Computational Methods For Option Pricing*, *Frontiers in Applied Mathematics*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.
- [3] Achdou, Y. & Pironneau, O. (2009). chapter Partial differential equations for option pricing, *Handbook on Mathematical Finance*, Elsevier, to appear.
- [4] Ciarlet, P.G. (1991). Basic error estimates for elliptic problems, *Handbook of Numerical Analysis*, North-Holland, Amsterdam, Vol. II, pp. 17–351.
- [5] Delpino, S. freefem3d, www.freefem.org.
- [6] Demmel, J.W., Gilbert, J.R. & Li, X.S. (1999). *SuperLU Users' Guide*, LBNL–44289, Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA.
- [7] Fouque, J.P., Papanicolaou, G. & Sircar, K.R. (2000). *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press, Cambridge.
- [8] Frey, P. *Medit a Visualization Software*, www.ann.jussieu.fr/~frey/logiciels/medit.html.
- [9] Hecht, F. & Pironneau, O. *freefem++*, www.freefem.org.
- [10] Huang, X. & Cornelis, W. (2008). *Oosterlee Adaptive integration for multi-factor portfolio credit loss models*, *Finance and Stochastics*, to appear.
- [11] Ikonen, S. & Toivanen, J. (2008). Efficient numerical methods for pricing American options under stochastic volatility, *Numerical Methods for Partial Differential Equations* **24**(1), 104–126.
- [12] Stein, E. & Stein, J. (1991). Stock price distributions with stochastic volatility : an analytic approach, *The Review of Financial Studies* **4**(4), 727–752.
- [13] Wilmott, P., Dewynne, J. & Howison, S. (1993). *Option Pricing*. Oxford Financial Press, Oxford.
- [14] Zienkiewicz, O.C. & Taylor, R.L. (2000). *The Finite Element Method*, 5th Edition, Butterworth-Heinemann, Oxford, Vol. 1, The basis.

YVES ACHDOU & OLIVIER PIRONNEAU

Fourier Transform

Discrete Fourier Transform (DFT)

The n th root of unity is denoted by $z_n := \cos \frac{2\pi}{n} + i \sin \frac{2\pi}{n} = e^{i\frac{2\pi}{n}}$. Geometrically, the numbers z_n^j , with $j = 0, \dots, n-1$, represent n equidistantly spaced points on the unit circle in the complex plane. The point z_n^{j+1} is reached from z_n^j by turning $1/n$ th of the full circle anticlockwise.

Definition 1 Given an n -dimensional vector $a = [a_0, a_1, \dots, a_{n-1}]$, we say that

$$\text{rev}(a) := [a_0, a_{n-1}, \dots, a_1] \quad (1)$$

is a in reverse order. Intuitively, if a is written around the circle in the anticlockwise direction, then $\text{rev}(a)$ is obtained by reading from a_0 in the clockwise direction.

Definition 2 The discrete Fourier transform (DFT) of $a = [a_0, a_1, \dots, a_{n-1}] \in \mathbb{C}^n$ is the vector $b = [b_0, b_1, \dots, b_{n-1}] \in \mathbb{C}^n$ such that

$$b_k = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} a_j z_n^{jk} = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} a_j e^{i\frac{2\pi}{n}jk} \quad (2)$$

We write

$$b = \mathcal{F}(a) \quad (3)$$

Equation (2) represents the forward transform. The inverse transform is given by

$$a_l = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} b_k z_n^{-kl} = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} b_k e^{-i\frac{2\pi}{n}kl} \quad (4)$$

and we write

$$a = \mathcal{F}^{-1}(b) \quad (5)$$

Proposition 1 The following statements hold:

1. The inverse DFT of sequence b is the same as the forward transform of the same sequence in the reverse order:

$$\mathcal{F}^{-1}(b) = \mathcal{F}(\text{rev}(b)) \quad (6)$$

and vice versa

$$\mathcal{F}^{-1}(\text{rev}(b)) = \mathcal{F}(b) \quad (7)$$

2. $\mathcal{F}^{-1}$ is indeed an inverse transformation to $\mathcal{F}$, that is,

$$\mathcal{F}^{-1}(\mathcal{F}(a)) = \mathcal{F}(\mathcal{F}^{-1}(a)) = a \quad (8)$$

Next, we explain the relationship between binomial option pricing and the DFT.

Binomial Option Pricing

Consider a two-period binomial model (see **Binomial Tree**) with a high rate of return of 25% and a low rate of return of -20%. Assuming that the initial stock price is 100, the evolution of the stock price in the two periods ahead is given in Table 1.

Suppose that we wish to price a call option struck at $K = 100$, maturing two periods from now. The risk-free rate is 0%. The intrinsic value of the option at maturity is

$$C(2, :) = [56.25 \quad 0 \quad 0]^\top \quad (9)$$

As per the asset pricing theory, the no-arbitrage price of the option is given recursively as

$$C(t, j) = \frac{q_u C(t+1, j) + q_d C(t+1, j+1)}{R_f} \quad (10)$$

where the risk-neutral probabilities q_u and q_d satisfy

$$q_u = \frac{R_f - R_d}{R_u - R_d} = \frac{1 - 0.8}{1.25 - 0.8} = \frac{4}{9} \quad (11)$$

$$q_d = 1 - q_u = \frac{5}{9} \quad (12)$$

Here, $R_u = S(t+1, j)/S(t, j)$, $R_d = S(t+1, j+1)/S(t, j)$, and $R_f - 1$ is equal to the risk-free rate. Recursive application of equation (10) with terminal value (9) leads to option prices as given in Table 2.

Option Pricing by Circular Convolution

For any two n -dimensional vectors $a = [a_0, a_1, \dots, a_{n-1}]$ and $b = [b_0, b_1, \dots, b_{n-1}]$, we define *circular* (cyclic) convolution of a and b to be a new vector $c = [c_0, c_1, \dots, c_{n-1}]$,

$$c = a \circledast b \quad (13)$$

Table 1 Binomial stock price lattice

Number of low returns j	$S(0, j)$	$S(1, j)$	$S(2, j)$
0	100	125	156.25
1		80	100
2			64

Table 2 Option prices in a binomial lattice

Number of low returns j	$C(0, j)$	$C(1, j)$	$C(2, j)$
0	$11\frac{1}{9}$	25	56.25
1		0	0
2			0

such that

$$c_j = \sum_{k+l=j \text{ or } k+l=j+n} a_k b_l \quad (14)$$

We use the binomial example to illustrate the use of circular convolution in option pricing. At maturity, the option has three values:

$$C(2, :) = [56.25 \quad 0 \quad 0]^\top \quad (15)$$

Let vector q contain the conditional one-period risk-neutral probabilities $q_u = 0.43523$ and $q_d = 0.56477$. Since there are just two states over one period the remaining entries will be padded by zeros:

$$q = [q_u \quad q_d \quad 0]^\top \quad (16)$$

To compute option prices at time $t = 1$, we evaluate

$$C(1, :) = C(2, :) \circledast \text{rev}(q) / R_f = [25 \quad 0 \quad 31.25]^\top \quad (17)$$

Note that we need only the first two prices in equation (17). The last entry is meaningless—it corresponds to the no-arbitrage price of the payoff $[0 \ 56.25]$. Moving backward in time, we have

$$\begin{aligned} C(0, :) &= C(1, :) \circledast \text{rev}(q) / R_f \\ &= \left[11\frac{1}{9} \quad 17\frac{13}{36} \quad 27\frac{7}{9} \right]^\top \end{aligned} \quad (18)$$

Again, only the first entry is meaningful.

This procedure can be generalized to multinomial lattices as follows.

Proposition 2 Consider an N -period model in which the random variables S_{t+1}/S_t are independent, identically distributed under the risk-neutral measure and such that $\ln S_{t+1}/S_t$ takes k distinct, equidistantly spaced values with probabilities $q_1, \dots, q_k$. Then stock price can be represented by a multinomial lattice with $1 + t(k - 1)$ values at time t . Let $C(N, :) \in \mathbb{R}^n$, $n = 1 + N(k - 1)$, be the payoff of a derivative asset as a function of S_N , and define risk-neutral price of the option recursively by setting

$$\begin{aligned} C(t, j) &= \sum_{l=1}^k q_l C(t+1, j+l-1), \\ 1 \leq l \leq 1 + t(k-1) \end{aligned} \quad (19)$$

Let q be a vector with first k entries $q_1, \dots, q_k$ and padded by zeros to dimension n . Then $C(t, :)$ can also be obtained from the following formula:

$$C(t, :) = C(N, :) \circledast \overbrace{\text{rev}(q) \circledast \text{rev}(q) \circledast \dots \circledast \text{rev}(q)}^{(N-t) \text{ times}} \quad (20)$$

The vector $C(t, :)$ computed in this manner has more entries than required (n in total); the useful $1 + t(k - 1)$ entries are at the top end of the vector.

Pricing via Discrete Fourier Transform

In this section, we reformulate the circular pricing formula (20) using the DFT.

Proposition 3 Consider $a, b \in \mathbb{C}^n$. The following equalities hold:

$$\mathcal{F}(a \circledast b) = \sqrt{n} \mathcal{F}(a) \mathcal{F}(b) \quad (21)$$

$$\mathcal{F}^{-1}(a \circledast b) = \sqrt{n} \mathcal{F}^{-1}(a) \mathcal{F}^{-1}(b) \quad (22)$$

Now applying the inverse transform $\mathcal{F}^{-1}$ to both sides of equation (20) and using property (22) on the

right-hand side

$$\mathcal{F}^{-1}(C_0) = \mathcal{F}^{-1}(C_N) \times \left(\sqrt{\text{dimension of } C_N} \mathcal{F}^{-1}(\text{rev}(q)) / R_f \right)^N \quad (23)$$

Recall from equation (7) that $\mathcal{F}^{-1}(\text{rev}(q)) = \mathcal{F}(q)$, and substitute this into equation (23). Finally, applying the forward transform to both sides again, and using equation (8) on the left-hand side, we obtain

$$C_0 = \mathcal{F} \left(\mathcal{F}^{-1}(C_N) \times \left(\sqrt{\text{dimension of } C_N} \mathcal{F}(q) / R_f \right)^N \right) \quad (24)$$

Fast Fourier Transform (FFT)

A naive implementation of the DFT algorithm with an n -dimensional input requires n^2 complex multiplications. An efficient implementation of DFT, known as the *fast Fourier transform (FFT)*, will only require $Kn \ln n$ operations, but one still has to choose n carefully because the constant K can be very large for some choices of n . There are many instances when FFT of length n_1 is faster than FFT of length n_2 even though $n_1 > n_2$. This slightly counterintuitive phenomenon is illustrated in Table 3. In general, it is advisable to use the transform size of the form $n = 2^p 3^q 5^r$, where q and r should be small compared to p . In the context of option pricing, this is always possible by padding the payoff vector C_N with a sufficient number of zeros.

An efficient implementation of FFT for all transform lengths is suggested in [17]. Efficient implementation of mixed 2, 3, 5-radix algorithm is

Table 3 Execution time of FFT algorithm for different input lengths n

n	Factorization	Execution time (seconds)
999 983	999 983	26.3
2097 150	$2 \times 3 \times 5^2 \times 11 \times 31 \times 41$	1.4
2097 152	2^{21}	0.42
2160 000	$2^7 3^3 5^4$	0.40

Intel Core 2, 2 GHz, 2Gb RAM, Matlab 6.5.

due to Temperton [25]. Duhamel and Vetterli [15] survey a wide range of FFT algorithms. Fast implementation of equation (24) for Gauss and Matlab is discussed in [7].

FFT in Continuous-time Models

In the continuous-time limit, the DFT is replaced by the (continuous) Fourier transform: that is, for a contingent claim with payoff $C_T = f(\ln S_T)$, we wish to find coefficients $\psi(v)$ such that

$$f(x) = \int_{\beta-i\infty}^{\beta+i\infty} \psi(v) e^{vx} dv \quad (25)$$

for some real constant β [19]. The recipe for obtaining the coefficients $\psi(v)$ is known—it is given by the inverse Fourier transform:

$$\psi(v) = \frac{1}{2\pi i} \int_{-\infty}^{+\infty} f(x) e^{-vx} dx \quad (26)$$

For example, the payoff of a European call option with strike price e^k corresponds to $f(x) = (e^x - e^k)^+$, whereby, simple integration yields

$$\psi(v) = \frac{e^{(1-v)k}}{2\pi i v(v-1)} \text{ for } \text{Re } v > 1 \quad (27)$$

Substituting for C_T from equation (25) the risk-neutral pricing formula reads

$$\begin{aligned} C_0(\ln S_0) &= e^{-rT} \mathbb{E}[C_T(\ln S_T)] \\ &= e^{-rT} \mathbb{E} \left[\int_{\beta-i\infty}^{\beta+i\infty} \psi(v) e^{v \ln S_T} dv \right] \\ &= e^{-rT} \int_{\beta-i\infty}^{\beta+i\infty} \psi(v) e^{v \ln S_0} \mathbb{E}[e^{v \ln S_T / S_0}] dv \end{aligned} \quad (28)$$

In practice, one is mainly interested in models where the risk-neutral characteristic function of log stock return $\phi(-iv) := \mathbb{E}[e^{v \ln S_T / S_0}]$ is available in the closed form. This is the case in the class of exponential Lévy models with affine stochastic volatility process, discussed in [6] (*see also Time-changed Lévy Process*). This class contains a large number of popular models allowing for excess kurtosis, stochastic volatility, and leverage effects. It includes, among

others, the stochastic volatility models of Heston [18] (see **Heston Model**), Duffie *et al.* [14], and all exponential Lévy models (cf. [16, 23], see **Exponential Lévy Models**). For a detailed description of affine processes, see [13, 20].

Option pricing, therefore, boils down to the evaluation of integrals of the type

$$\begin{aligned} C_0(\ln S_0) &= S_0 e^{-rT} \int_{\beta-i\infty}^{\beta+i\infty} \frac{e^{(v-1)(\ln S_0-k)}}{2\pi v(v-1)} \phi(-iv) dv \\ &= 2S_0 e^{-rT} \int_{\beta+i0}^{\beta+i\infty} \operatorname{Re} \left(\frac{e^{(v-1)(\ln S_0-k)}}{2\pi v(v-1)} \phi(-iv) \right) dv \end{aligned} \quad (29)$$

where both $\psi(v)$ and $\phi(v)$ are known. To evaluate equation (29), one truncates the integral at a high value of $\operatorname{Im}v$ and then uses a numerical quadrature to approximate it by a sum; see [22] for a detailed exposition. This yields an expression of the following type:

$$\begin{aligned} C_0(\ln S_0) &\approx 2\operatorname{Re} \left(S_0 e^{-rT} \sum_{j=0}^{n-1} w_j \frac{e^{(v_j-1)(\ln S_0-k)}}{2\pi v_j(v_j-1)} \phi(-iv_j) \right) \end{aligned} \quad (30)$$

where the integration weights w_j and abscissas v_j depend on the quadrature rule. It is particularly convenient to use Newton–Cotes rules, which employ equidistantly spaced abscissas. For example, a trapezoidal rule yields

$$v_j = \beta + ij\Delta v \quad (31)$$

$$\operatorname{Im}v_{\max} = (n-1)\Delta v \quad (32)$$

$$w_0 = w_n = \frac{1}{2}\Delta v \quad (33)$$

$$w_1 = w_2 = \dots = w_{n-1} = \Delta v \quad (34)$$

In conclusion, if the characteristic function of log returns is known, one needs to evaluate a single sum (30) to find the option price. Consequently, there is *no need* to use FFT if one wishes to evaluate the option price for *one fixed* log strike k .

FFT Option Pricing with Multiple Strikes

The situation is different if we wish to evaluate the option price (30) for many different strikes simultaneously. Let us consider m values of moneyness $\kappa_l = \ln S_0 - k_l$, ranging from $\kappa_{\max} - m\Delta\kappa$ to $\kappa_{\max}$ with increment $\Delta\kappa$

$$\kappa_l = \kappa_{\max} - l\Delta\kappa \quad (35)$$

$$l = 0, \dots, m-1 \quad (36)$$

The idea of using FFT in this context is due to Carr and Madan [5], who noted that with equidistantly spaced abscissas (31) one can write the option pricing equation (30) for different strike values (35 and 36) as a z -transform with $z_l = e^{-i\Delta v \Delta \kappa l}$:

$$C_{0l} = 2S_0 e^{(\beta-1)\kappa_l - rT} \operatorname{Re} \sum_{k=0}^{n-1} e^{i\Delta v \Delta \kappa kl} a_j \quad (37)$$

$$a_j = w_j \frac{e^{ij\Delta v \kappa_{\max}} \phi(-iv_j)}{2\pi v_j(v_j-1)} \quad (38)$$

Setting $\Delta v \Delta \kappa = \frac{2\pi}{n}$, Carr and Madan obtained a DFT in equation (37). Chourdakis [11] points out that one can evaluate equation (37) by means of a fractional FFT with $\Delta v \Delta \kappa \neq \frac{2\pi}{n}$. For the discussion of relative merits of the two strategies, see [7].

Further Reading

Further applications of FFT appear in [1–4, 10, 12, 24]. For the latest developments in option pricing using (continuous) Fourier transform, see [6, 8, 9, 19, 21]. Proofs of Propositions 1 and 3 can be found in [7].

Acknowledgments

I am grateful to Peter Forsyth and Damien Lamberton for their helpful comments.

References

- [1] Albanese, C., Jackson, K. & Wiberg, P. (2004). A new Fourier transform algorithm for value at risk, *Quantitative Finance* **4**, 328–338.
- [2] Andersen, L. & Andreasen, J. (2000). Jump diffusion processes: volatility smile fitting and numerical methods

- for option pricing, *Review of Derivatives Research* **4**(3), 231–261.
- [3] Andreas, A., Engelmann, B., Schwendner, P. & Wystup, U. (2002). Fast Fourier method for the valuation of options on several correlated currencies, in *Foreign Exchange Risk*, J. Hakala & U. Wystup, eds, Risk Publications.
- [4] Benhamou, E. (2002). Fast Fourier transform for discrete Asian options, *Journal of Computational Finance* **6**(1), 49–68.
- [5] Carr, P. & Madan, D.B. (1999). Option valuation using the fast Fourier transform, *Journal of Computational Finance* **2**, 61–73.
- [6] Carr, P. & Wu, L. (2004). Time-changed Lévy processes and option pricing, *Journal of Financial Economics* **71**(1), 113–141.
- [7] Černý, A. (2004). Introduction to FFT in finance, *Journal of Derivatives* **12**(1), 73–88.
- [8] Černý, A. (2007). Optimal continuous-time hedging with leptokurtic returns, *Mathematical Finance* **17**(2), 175–203.
- [9] Černý, A. & Kallsen, J. (2008). Mean–variance hedging and optimal investment in Heston’s model with correlation, *Mathematical Finance* **18**(3), 473–492.
- [10] Chiarella, C. & El-Hassan, N. (1997). Evaluation of derivative security prices in the Heath-Jarrow-Morton framework as path integrals using fast Fourier transform techniques, *Journal of Financial Engineering* **6**(2), 121–147.
- [11] Chourdakis, K. (2004). Option pricing using the fractional FFT, *Journal of Computational Finance* **8**(2), 1–18.
- [12] Dempster, M.A.H. & Hong, S.S.G. (2002). Spread option valuation and the fast Fourier transform, in *Mathematical Finance – Bachelier Congress 2000*, H. Geman, D. Madan, S.R. Pliska & T. Vorst, eds, Springer, pp. 203–220.
- [13] Duffie, D., Filipović, D. & Schachermayer, W. (2003). Affine processes and applications in finance, *The Annals of Applied Probability* **13**(3), 984–1053.
- [14] Duffie, D., Pan, J. & Singleton, K. (2000). Transform analysis and asset pricing for affine jump diffusions, *Econometrica* **68**, 1343–1376.
- [15] Duhamel, P. & Vetterli, M. (1990). Fast Fourier transforms: a tutorial review and a state of the art, *Signal Processing* **19**, 259–299.
- [16] Eberlein, E., Keller, U. & Prause, K. (1998). New insights into smile, mispricing and value at risk: the hyperbolic model, *Journal of Business* **71**(3), 371–405.
- [17] Frigo, M. & Johnson, S.G. (1998). FFTW: an adaptive software architecture for the FFT, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, Vol. 3, pp. 1381–1384.
- [18] Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**, 327–344.
- [19] Hubalek, F., Kallsen, J. & Krawczyk, L. (2006). Variance-optimal hedging for processes with stationary independent increments, *The Annals of Applied Probability* **16**, 853–885.
- [20] Kallsen, J. (2006). A didactic note on affine stochastic volatility models, in *From Stochastic Calculus to Mathematical Finance: The Shiryaev Festschrift*, Y. Kabanov, R. Liptser & J. Stoyanov, eds, Springer, Berlin, pp. 343–368.
- [21] Kallsen, J. & Vierthauer, R. (2009). Variance-optimal hedging in affine stochastic volatility models, *Review of Derivatives Research* **12**(1), 3–27.
- [22] Lee, R.W. (2004). Option pricing by transform methods: extensions, unification and error control, *Journal of Computational Finance* **7**(3), 1–15.
- [23] Madan, D. & Seneta, E. (1990). The variance gamma model for stock market returns, *Journal of Business* **63**(4), 511–524.
- [24] Rebonato, R. & Cooper, I. (1998). Coupling backward induction with monte carlo simulations: a fast Fourier transform (FFT) approach, *Applied Mathematical Finance* **5**(2), 131–141.
- [25] Temperton, C. (1992). A generalized prime factor FFT algorithm for any $n = 2^p 3^q 5^s$, *SIAM Journal on Scientific and Statistical Computing* **13**, 676–686.

Related Articles

Exponential Lévy Models; Fourier Methods in Options Pricing; Wavelet Galerkin Method.

ALĚŠ ČERNÝ

Crank–Nicolson Scheme

In a paper published in 1947 [2], John Crank and Phyllis Nicolson presented a numerical method for the approximation of diffusion equations. Starting from the simplest example

$$\frac{\partial V}{\partial t} = \frac{\partial^2 V}{\partial x^2} \quad (1)$$

a spatial approximation on a uniform grid with spacing h leads to the semidiscrete equations

$$\frac{dV_j}{dt} = h^{-2} \delta_j^2 V_j \quad (2)$$

where $V_j(t)$ is an approximation to $V(x_j, t)$ and $\delta_j^2 V_j \equiv V_{j+1} - 2V_j + V_{j-1}$ is a central second difference. Crank–Nicolson time-marching discretizes this in time with a uniform timestep k using the approximation

$$k^{-1} (V_j^{n+1} - V_j^n) = \frac{1}{2} h^{-2} (\delta_j^2 V_j^{n+1} + \delta_j^2 V_j^n) \quad (3)$$

which can be rearranged to give

$$(1 - \frac{1}{2} k h^{-2} \delta_j^2) V_j^{n+1} = (1 + \frac{1}{2} k h^{-2} \delta_j^2) V_j^n \quad (4)$$

This can be viewed as the $\theta = \frac{1}{2}$ case of the more general θ scheme

$$(1 - \theta k h^{-2} \delta_j^2) V_j^{n+1} = (1 + (1-\theta) k h^{-2} \delta_j^2) V_j^n \quad (5)$$

$\theta=0$ corresponds to explicit Euler time-marching, while $\theta=1$ corresponds to fully implicit Euler time-marching. For $\theta > 0$, the θ -scheme defines a tridiagonal system of simultaneous equations, which can be solved very efficiently using the Thomas algorithm to obtain the values for V_j^{n+1} . The scheme is unconditionally stable in the L_2 norm, meaning that the L_2 norm of the solution does not increase for any value of k , provided $\theta \geq 1/2$. The Crank–Nicolson scheme is thus on the boundary of unconditional stability. It is also special in having a numerical error for smooth initial data, which is $O(h^2, k^2)$, whereas the error is $O(h^2, k)$ for other values of θ . The unconditional L_2 stability means that one can choose to make k proportional to h , and, together with the second-order accuracy, this makes the scheme both accurate

and efficient, and hence a very popular choice for approximating parabolic PDEs (see **Partial Differential Equations**).

Application to Black–Scholes Equation

The Crank–Nicolson method is used extensively in mathematical finance for approximating parabolic PDEs such as the Black–Scholes equation, which can be written in reversed-time form (with $\tau \equiv T - t$ being the time to maturity T) as

$$\frac{\partial V}{\partial \tau} = -rV + rS \frac{\partial V}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} \quad (6)$$

Switching to the new coordinate $x \equiv \log S$ gives the transformed equation

$$\frac{\partial V}{\partial \tau} = -rV + (r - \frac{1}{2} \sigma^2) \frac{\partial V}{\partial x} + \frac{1}{2} \sigma^2 \frac{\partial^2 V}{\partial x^2} \quad (7)$$

and its Crank–Nicolson discretization on a grid with uniform timestep k and uniform grid spacing h is

$$(I + \frac{1}{2} D) V_j^{n+1} = (I - \frac{1}{2} D) V_j^n \quad (8)$$

where the discrete operator D is defined by

$$D = r k - \frac{1}{2} k h^{-1} (r - \frac{1}{2} \sigma^2) \delta_{2j} - \frac{1}{2} k h^{-2} \sigma^2 \delta_j^2 \quad (9)$$

with the central first difference operator δ_{2j} defined by $\delta_{2j} V_j \equiv V_{j+1} - V_{j-1}$.

For an European call option with strike K , the initial data at maturity is $V(S, 0) = \max(S - K, 0)$. Figure 1a shows the numerical solution $V(S, 2)$ for parameter values $r=0.05$, $\sigma=0.2$, $K=1$, and timestep/spacing ratio $\lambda \equiv k/h = 10$. The agreement between the numerical solution and the known analytic solution appears quite good, but b and c show much poorer agreement for the approximations to $\Delta \equiv \partial V / \partial S$ and $\Gamma \equiv \partial^2 V / \partial S^2$ (see **Delta Hedging**; **Gamma Hedging**) obtained by central differencing of the numerical solution V_j^n . In particular, note that the maximum error in the computed value for Γ occurs at $S=1$, which is the location of the discontinuity in the first derivative of the initial data. Figure 2a–c show the behavior of the maximum error as the computational grid is refined, keeping fixed the ratio $\lambda \equiv k/h$. It can be seen that for

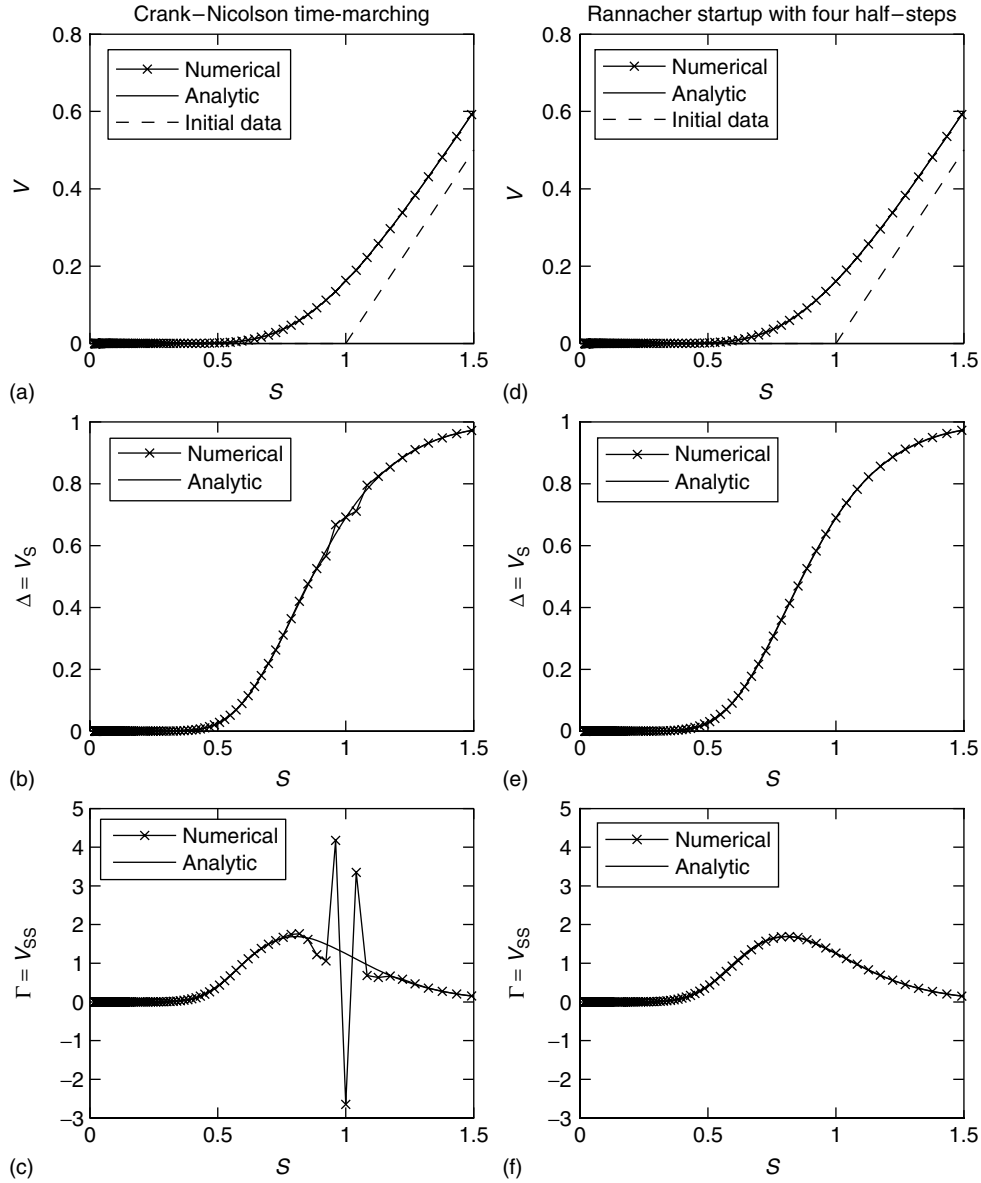


Figure 1 V , Δ and Γ for a European call option, with $\lambda \equiv k/h = 10$

largest value of λ the numerical solution V_j exhibits first-order convergence, while the discrete approximation to Δ does not converge, and the approximation to Γ diverges. For smaller values of λ , it appears that the convergence is better, but, in fact, the asymptotic behavior is exactly the same except that it becomes evident only on much finer grids.

At first sight, this is a little surprising as textbooks almost always describe the Crank–Nicolson method as unconditionally stable and second-order accurate. The key is that it is only unconditionally stable in the L_2 norm, and this only ensures convergence in the L_2 norm for initial data, which has a finite L_2 norm [9]. Furthermore, the order of convergence may be less than second order for initial data, which is not

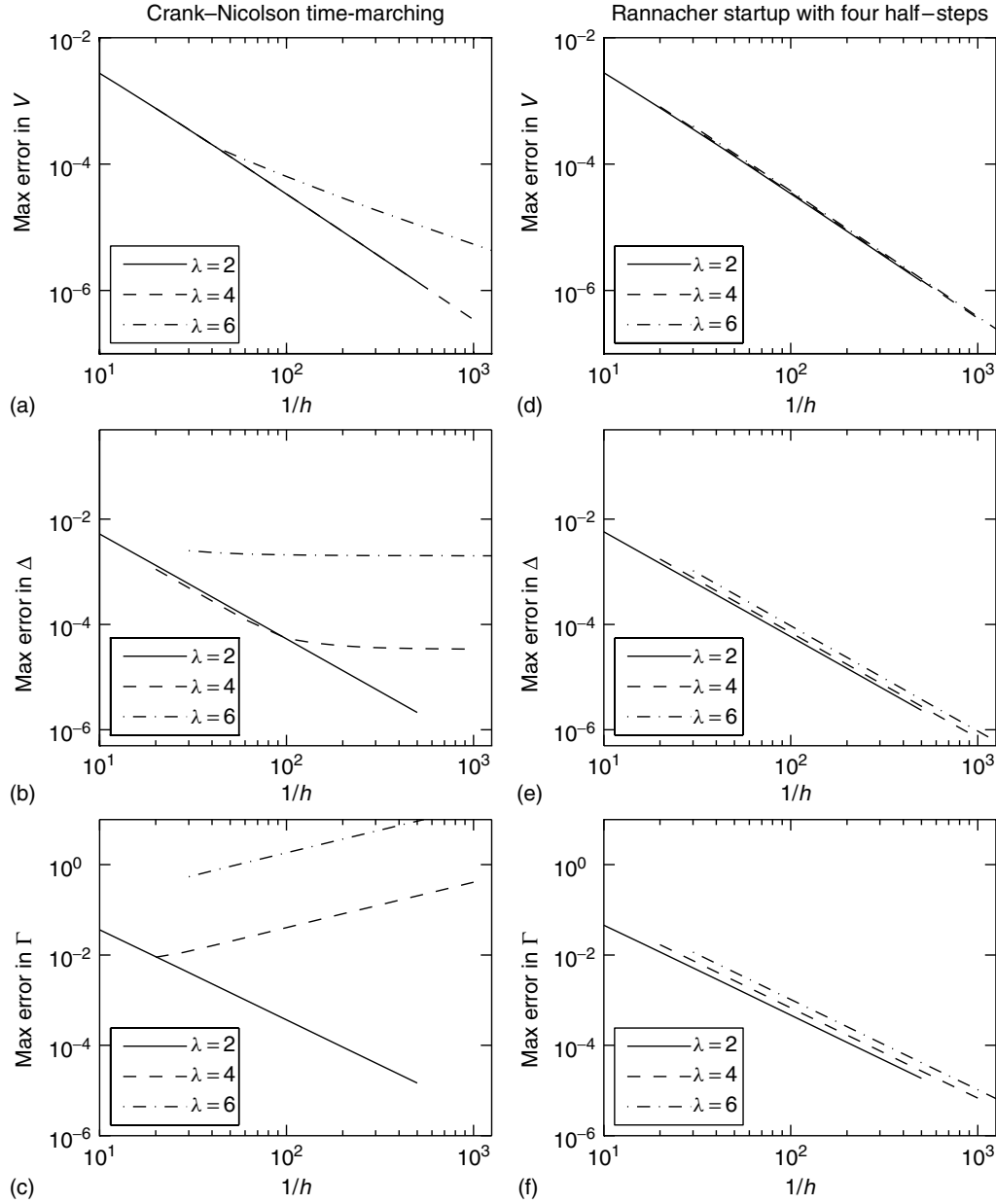


Figure 2 Grid convergence for a European call option, with fixed $\lambda \equiv k/h$

smooth; for example, the L_2 order of convergence for discontinuous initial data is $1/2$. With the European call Black–Scholes application, the initial data for V lies in L_2 , as does its first derivative, but the second derivative is the Dirac delta function, which does not lie in L_2 . This is the root cause of the observed failure

to converge as the grid is refined. Furthermore, it is the maximum error, the L_∞ error, which is most relevant in financial applications.

One solution to this problem is to use an alternative second-order backward difference method, but these methods require special start-up procedures

because they require more than one previous time level, and they are usually less accurate than the Crank–Nicolson method for the same number of timesteps. Better alternatives are higher order backward difference methods [5] or the Rannacher start-up procedure described in the next section.

Rannacher Start-up Procedure

Rannacher analyzed this problem of poor L_2 convergence of convection–diffusion approximations with discontinuous initial data [8], and recovered second-order convergence by replacing the Crank–Nicolson approximation for the very first timestep by two half-timesteps of implicit Euler time integration, and by using a finite element projection of the discontinuous initial data onto the computational grid. This technique, often referred to as *Rannacher timestep-ping*, has been used with success in approximations of the Black–Scholes equations [6, 7], with the half-timestep implicit Euler discretization given by

$$\left(I + \frac{1}{2}D\right) V_j^{n+1/2} = V_j^n \quad (10)$$

The problem has been further investigated by Giles and Carter [4] who analyzed the maximum errors in finance applications and proved that it is necessary to go further and replace the first two Crank–Nicolson timesteps by four half-timesteps of implicit Euler to achieve second-order accuracy in the L_∞ norm for V , Δ and Γ for put, call, and digital options. The improved accuracy is demonstrated by (d–f) in Figures 1 and 2.

Nonlinear and Multifactor Extensions

The use of a nonlinear penalty function in approximating American options (see **Finite Difference Methods for Early Exercise Options**) leads to a nonlinear discretization of the form [3]

$$\left(I + \frac{1}{2}D\right) V_j^{n+1} = \left(I - \frac{1}{2}D\right) V_j^n + P\left(V_j^{n+1}\right) \quad (11)$$

where the nonlinear penalty term $P(V_j^{n+1})$ is negligible in the region where the option is not exercised, and elsewhere ensures that V_j^{n+1} is approximately equal to the exercise value.

This nonlinear system of equations can be solved using a Newton iteration, starting with $V_j^{n+1,0} = V_j^n$

and defining the $m+1^{th}$ iterate to be $V_j^{n+1,m+1} = V_j^{n+1,m} + \Delta V_j$ with the correction ΔV_j given by the linear equations

$$\begin{aligned} & \left(I + \frac{1}{2}D - \frac{\partial P}{\partial V}\right) \Delta V_j \\ &= -\left(I + \frac{1}{2}D\right) V_j^{n+1,m} \\ &+ \left(I - \frac{1}{2}D\right) V_j^n + P\left(V_j^{n+1,m}\right) \end{aligned} \quad (12)$$

Alternatively, one can use just one step of the Newton iteration, in which case one has $V_j^{n+1} = V_j^n + \Delta V_j$ with the change ΔV_j given by

$$\left(I + \frac{1}{2}D - \frac{\partial P}{\partial V}\right) \Delta V_j = -D V_j^n + P(V_j^n) \quad (13)$$

In one dimension, the linear equations are a tridiagonal system that can be solved very efficiently. In higher dimensions, the direct solution cost is much greater and alternative approaches are usually adopted. One is to use an **Alternating Direction Implicit (ADI) Method** approximate factorization into a product of operators, each of which involves differences in only one direction [9]. To maintain second-order accuracy, it is necessary to use the Craig–Sneyd treatment for any cross-derivative term [1]. Another approach is to use a preconditioned iterative solver such as BiCGStab with ILU preconditioning (see **Conjugate Gradient Methods**).

References

- [1] Craig, I.J.D. & Sneyd, A.D. (1988). An alternating-direction implicit scheme for parabolic equations with mixed derivatives, *Computers and Mathematics with Applications* **16**(4), 341–350.
- [2] Crank, J. & Nicolson, P. (1947). A practical method for numerical integration of solutions of partial differential equations of heat-conduction type. *Proceedings Cambridge Philosophical Society* **43**, 50.
- [3] Forsyth, P.A. & Vetzal, K.R. (2002). Quadratic convergence for valuing American options using a penalty method, *SIAM Journal on Scientific Computing* **23**(6), 2095–2122.
- [4] Giles, M.B. & Carter, R. (2006). Convergence analysis of Crank–Nicolson and Rannacher time-marching, *Journal of Computational Finance* **9**(4), 89–112.
- [5] Khaliq, A.Q.M., Voss, D.A., Yousuf, R. & Wendland, W. (2007). Pricing exotic options with L-stable Padé schemes, *Journal of Banking and Finance* **31**(11), 3438–3461.

- [6] Pooley, D.M., Forsyth, P.A. & Vetzal, K.R. (2003). Numerical convergence properties of option pricing PDEs with uncertain volatility, *IMA Journal of Numerical Analysis* **23**, 241–267.
- [7] Pooley, D.M., Vetzal, K.R. & Forsyth, P.A. (2003). Convergence remedies for non-smooth payoffs in option pricing, *Journal of Computational Finance* **6**(4), 25–40.
- [8] Rannacher, R. (1984). Finite element solution of diffusion problems with irregular data, *Numerische Mathematik* **43**, 309–327.
- [9] Richtmyer, R.D. & Morton, K.W. (1967). *Difference Methods for Initial-value Problems*, 2nd Edition, John

Wiley & Sons. Reprint Edition (1994), Krieger Publishing Company, Malabar.

Related Articles

Alternating Direction Implicit (ADI) Method; Conjugate Gradient Methods; Finite Difference Methods for Barrier Options; Finite Difference Methods for Early Exercise Options; Finite Element Methods; Partial Differential Equations.

MICHAEL B. GILES

Monotone Schemes

Monotone numerical schemes are particularly relevant for partial differential equations (PDEs) occurring in finance, for example, in the pricing of American options, exotic options, and in portfolio problems [4]. Although naive numerical schemes for such problems may not converge, or if they converge, may converge to an incorrect solution, monotone schemes converge to the correct solution under very weak assumptions on the schemes, equations, and boundary conditions. The drawback is that these schemes may converge slowly, as they are only first- or second-order accurate.

result. Finally, we discuss error estimates for convex problems.

Viscosity Solutions

The concept of viscosity solution was introduced by Crandall and Lions in 1983. A general reference on the subject is the User's Guide [14]. Several books have also been written on the subject, see for example, [1–3, 19, 21]. Viscosity solutions are particularly important for nonlinear or degenerate PDEs of first and second order. There are many such problems in finance, we show some examples taken from [4]. Lower indices denote partial derivatives.

$$\begin{cases} -u_t - \frac{1}{2}\sigma^2 S^2 u_{SS} - rSu_S + ru - \frac{S}{T}u_Z = 0 & \text{in } \mathbb{R}^+ \times \mathbb{R}^+ \times (0, T) \\ u(S, Z, T) = (Z - S)^+ & \text{in } \mathbb{R}^+ \times \mathbb{R}^+ \end{cases} \quad (1)$$

$$\begin{cases} \min\{-u_t - \frac{1}{2}\sigma^2 S^2 u_{SS} - rSu_S + ru, u - (K - S)^+\} = 0 & \text{in } \mathbb{R}^+ \times (0, T) \\ u(S, T) = (K - S)^+ & \text{in } \mathbb{R}^+ \end{cases} \quad (2)$$

$$\begin{cases} \min \left\{ \inf_{c \geq 0} \left[-v_t - \frac{1}{2}\sigma^2 y^2 u_{yy} - (rx - c)u_x - byv_y - U(c) + \beta v \right], \right. \\ \left. -v_y + (1 + \lambda)v_x, -(1 - \mu)v_x + v_y \right\} = 0 \\ \text{in } \{(x, y) : x + (1 - \mu)y \geq 0, x + (1 + \lambda)y \geq 0\} \times (0, T) \end{cases} \quad (3)$$

Convergence results and error estimates for monotone schemes can be obtained using the concept of *viscosity solution*. The main goal of this survey is to introduce these methods and describe the results they lead to. The main results are (i) a general convergence theorem and (ii) error estimates for convex problems.

Viscosity solutions are (very) weak solutions of linear and nonlinear first- and second-order equations. They are closely related to the maximum (comparison) principle from which uniqueness of solutions follows. Viscosity solutions are useful for proving convergence of numerical schemes, because they are stable under very weak continuity assumptions: if a sequence of equations and their viscosity solutions converge locally uniformly, then the limit solution is a viscosity solution of the limit equation. The general convergence results is an extension of this statement.

We first motivate and define viscosity solutions and discuss basic facts about them. Then we discuss monotone schemes and give the general convergence

and in the last case we also impose state-constrained boundary conditions. The first two problems are related to the pricing of Asian and American options. The first equation is degenerate in the Z -direction and at $S = 0$ (meaning there is no diffusion here), the second equation is an obstacle problem with degeneracy at $S = 0$. The last one, related to an investment-consumption model by Tourin and Zariphopoulou, is nonlinear, degenerate, and has difficult boundary conditions.

Note that to solve these problems numerically, we must reduce to bounded domains and hence we need to impose further boundary conditions. All of these equations are *degenerate elliptic* equations that can be written in the following abstract form:

$$F(x, u(x), Du(x), D^2u(x)) = 0 \quad \text{in } \Omega \quad (4)$$

for some domain (open connected set) Ω in $\mathbb{R}^N$ and function $F(x, r, p, X)$ on $\Omega \times \mathbb{R} \times \mathbb{R}^N \times \mathcal{S}_N$ where $\mathcal{S}_N$ is the space of all real symmetric $N \times N$ matrices,

and where F satisfies the condition

$$F(x, r, p, X) \leq F(x, s, p, Y) \quad \text{whenever } r \leq s \quad \text{and } X \geq Y \quad (5)$$

Here, $X \geq 0$ means that X is a positive semidefinite matrix. The assumption that F is nonincreasing in X is called *degenerate ellipticity*. Note that assumption (5) rules out many quasi-linear equations like, for example, conservation laws, and that we represent both time-dependent and time-independent problems in the form (4). In the time-dependent case, we take $x = (t, x')$ for $t \geq 0$, $x' \in \mathbb{R}^{N-1}$. It is an instructive exercise to check that the above equations satisfy equation (5). This abstract formulation will help us formulate results in an economical way.

A *classical solution* of equation (4) is a function u in $C^2(\Omega)$ (twice continuous differentiable functions on Ω) satisfying equation (4) in every point in Ω . We now define viscosity solution for equation (4) starting with the following observation: if u is a classical solution of equation (4), ϕ belongs to $C^2(\Omega)$, and $u - \phi$ has local maximum at $x_0 \in \Omega$, then

$$Du(x_0) = D\phi(x_0) \quad \text{and} \quad D^2u(x_0) \leq D^2\phi(x_0) \quad (6)$$

and hence by equations (4) and (5) we must have

$$F(x_0, u(x_0), D\phi(x_0), D^2\phi(x_0)) \leq 0 \quad (7)$$

On the other hand, if x_0 is a local minimum point of $u - \phi$, then we get the opposite inequality. If these inequalities hold for *all test functions* ϕ and *all maximum/minimum points* x_0 of $u - \phi$ and the function u belongs to C^2 , then it easily follows that u is a classical solution of equation (4).

This second definition of classical solutions can be used to define viscosity solutions simply by relaxing the regularity assumption on u from C^2 to continuous. Note that in this definition, only test functions need to be differentiated.

Definition 1 A viscosity solution of equation (4) is a continuous function u on Ω satisfying

$$F(x_0, u(x_0), D\phi(x_0), D^2\phi(x_0)) \leq 0 \quad (8)$$

for all C^2 test functions ϕ and all maximum points x_0 of $u - \phi$, and

$$F(x_1, u(x_1), D\phi(x_1), D^2\phi(x_1)) \geq 0 \quad (9)$$

for all C^2 test functions ϕ and all minimum points x_1 of $u - \phi$.

By the previous discussion, we see that all classical solutions are viscosity solutions and that all C^2 viscosity solutions are classical solutions.

A problem that is often encountered when you study nonlinear equations is that classical solutions may not exist and weak solutions are not unique [2, 18]. To pick out the physically relevant solution (and solve the nonuniqueness problem), we often have to require that additional (entropy) inequalities are satisfied by the solution. One of the main strengths of viscosity solutions is that they are unique under very general assumptions—in some sense the additional constraints are built into the definition.

Example 1 Consider the following initial value problem,

$$\begin{aligned} u_t + |u_x| &= 0 \quad \text{in } \mathbb{R} \times (0, +\infty), \\ u(0, x) &= |x| \quad \text{in } \mathbb{R} \end{aligned} \quad (10)$$

It has *no* classical solutions, *infinitely many* generalized solutions (functions satisfying the equation a.e.), and one *unique* viscosity solution. Two generalized solutions are $|x| - t$ and $(|x| - t)^+$, and the last one is also the viscosity solution.

Now we explain how to impose boundary conditions for degenerate equations satisfying equation (5). Consider

$$\begin{cases} F(x, u(x), Du(x), D^2u(x)) = 0 & \text{in } \Omega \\ G(x, u, Du) = 0 & \text{in } \partial\Omega \end{cases} \quad (11)$$

where G gives the boundary conditions. Dirichlet and Neumann boundary conditions are obtained by choosing

$$\begin{aligned} G(x, r, p) &= r - g(x) \\ \text{and} \quad G(x, r, p) &= p \cdot n(x) - h(x) \end{aligned} \quad (12)$$

respectively, where $n(x)$ is the exterior unit normal vector field of $\partial\Omega$. The problem here is that under assumption (5), the equation may be degenerate on all or a part of the boundary $\partial\Omega$. This part of the boundary may not be *regular* with respect to the

equation, meaning that boundary conditions imposed here do not influence the solution of equation (11) in Ω . In Ω , a solution of equation (11) is determined only by the equation and the boundary conditions on $\partial\Omega_{\text{reg}}$, the regular part of the boundary. Imposing boundary conditions in $\partial\Omega_{\text{irreg}} = \partial\Omega \setminus \partial\Omega_{\text{reg}}$ therefore makes the solution *discontinuous* in $\partial\Omega_{\text{irreg}}$ in general. We refer to [5, 14, 28] for a more detailed discussion.

Note that the continuous extension $\tilde{u}$ of the solution u from Ω to $\bar{\Omega}$, satisfies by continuity $F = 0$ also in $\partial\Omega_{\text{irreg}}$ (under suitable assumptions). Hence at any boundary point, $\tilde{u}$ satisfies *either* the boundary condition *or* the equation.

Now we give the precise definition of discontinuous viscosity solutions. This concept is crucial for the main convergence result of this survey, and as we have just seen, the boundary value problem (11) is well posed in general only if solutions can be discontinuous on the boundary. To this end, we define the upper and lower semicontinuous envelopes of u ,

$$u^*(x) = \limsup_{\Omega \ni y \rightarrow x} u(y) \quad \text{and} \quad u_*(x) = \liminf_{\Omega \ni y \rightarrow x} u(y) \quad (13)$$

A function u is upper semicontinuous, lower semicontinuous, or continuous at $x \in \Omega$ if and only if $u(x) = u^*(x)$, $u(x) = u_*(x)$, or $u(x) = u^*(x) = u_*(x)$, respectively. At a boundary point, the definition requires that either the boundary condition or the equation holds.

Definition 2 A discontinuous viscosity subsolution of equation (11) is a locally bounded function u on $\bar{\Omega}$ satisfying for all $\phi \in C^2(\bar{\Omega})$ and all local maximum points $x_0 \in \bar{\Omega}$ of $u^* - \phi$,

$$\begin{cases} F(x_0, u^*(x_0), D\phi(x_0), D^2\phi(x_0)) \leq 0 & \text{if } x_0 \in \Omega \\ \min \left\{ F(x_0, u^*(x_0), D\phi(x_0), D^2\phi(x_0)), \right. \\ \quad \left. G(x_0, u^*(x_0), D\phi(x_0)) \right\} \leq 0 & \text{if } x_0 \in \partial\Omega \end{cases} \quad (14)$$

A discontinuous viscosity supersolution of equation (11) is a locally bounded function u on $\bar{\Omega}$ satisfying for all $\phi \in C^2(\bar{\Omega})$ and all local minimum points

$x_0 \in \bar{\Omega}$ of $u_* - \phi$,

$$\begin{cases} F(x_0, u_*(x_0), D\phi(x_0), D^2\phi(x_0)) \geq 0 & \text{if } x_0 \in \Omega \\ \max \left\{ F(x_0, u_*(x_0), D\phi(x_0), D^2\phi(x_0)), \right. \\ \quad \left. G(x_0, u_*(x_0), D\phi(x_0)) \right\} \geq 0 & \text{if } x_0 \in \partial\Omega \end{cases} \quad (15)$$

A discontinuous viscosity solution of equation (11) is sub and supersolution at the same time.

Viscosity solutions are unique under very weak assumptions and stable under continuous perturbations of the equation. From the strong comparison result (a uniqueness result) and Perron's method, interior continuity and existence follows. We refer to [14] for precise statements and a wider discussion.

A classical way to compute the solution of degenerate boundary value problem is via elliptic/parabolic regularization. This method is called the *vanishing viscosity method*: if u^ϵ , $\epsilon > 0$, are classical (or viscosity) solutions of

$$\begin{cases} -\epsilon \Delta u + F(x, u, Du, D^2u) = 0 & \text{in } \Omega \\ G(x, u, Du) = 0 & \text{in } \partial\Omega \end{cases} \quad (16)$$

then u_ϵ converge pointwise as $\epsilon \rightarrow 0$ to the discontinuous viscosity solution of equation (11) under mild assumptions [14]. In the regularized problem, boundary conditions are assumed continuously, but as $\epsilon \rightarrow 0$ there will be formation of boundary layers near $\partial\Omega_{\text{irreg}}$, and in the limit the solution will be discontinuous here.

Example 2 Let $\epsilon > 0$ and consider the boundary value problem

$$\begin{cases} -\epsilon u''(x) + u'(x) - 1 = 0, & x \in (0, 1) \\ u(x) = 0, & x \in \{0, 1\} \end{cases} \quad (17)$$

Here, the unique (classical) solution u_ϵ converges pointwise (but not uniformly) as $\epsilon \rightarrow 0$ to a discontinuous function u :

$$u_\epsilon(x) = x - \frac{1 - e^{-\frac{1}{\epsilon}x}}{1 - e^{-\frac{1}{\epsilon}}} \xrightarrow{\epsilon \rightarrow 0} u = \begin{cases} 0, & x = 0 \\ x - 1, & x \in (0, 1] \end{cases} \quad (18)$$

By formally taking the limit in equation (17), we get the following boundary value problem,

$$\begin{cases} u'(x) - 1 = 0, & x \in (0, 1) \\ u(x) = 0, & x \in \{0, 1\} \end{cases} \quad (19)$$

We write this problem in the form (11) by taking $F(p) \equiv p - 1$ and $G(r) \equiv r$. Then, since $u^* \equiv u$ and $u_* \equiv x - 1$ in $[0, 1]$, it is easy to see that in the viscosity sense,

$$\begin{aligned} & \begin{cases} F\left(\frac{du^*}{dx}\right) \leq 0, & x \in (0, 1), \\ G(u^*) \leq 0, & x = 0, \\ G(u^*) \leq 0, & x = 1, \end{cases} \\ \text{and} \quad & \begin{cases} F\left(\frac{du_*}{dx}\right) \geq 0, & x \in (0, 1) \\ F\left(\frac{du_*}{dx}\right) \geq 0, & x = 0 \\ G(u_*) \geq 0, & x = 1 \end{cases} \end{aligned} \quad (20)$$

Hence, u is a viscosity solution of equation (19) according to Definition 2.

To have an even more compact notation, we define

$$H(x, r, p, X) = \begin{cases} F(x, r, p, X) & \text{if } x \in \Omega \\ G(x, r, p) & \text{if } x \in \partial\Omega \end{cases} \quad (21)$$

and note that H^* , H_* equals H in Ω and $\max(F, G)$, $\min(F, G)$, respectively, on $\partial\Omega$. Hence, by the above discussion, u is viscosity solution of equation (11), or equivalently, of

$$H(x, u, Du, D^2u) = 0 \quad \text{in } \bar{\Omega} \quad (22)$$

if the following inequalities hold in the viscosity sense:

$$H_*(x, u, Du, D^2u) \leq 0 \quad \text{in } \bar{\Omega} \quad (23)$$

$$H^*(x, u, Du, D^2u) \geq 0 \quad \text{in } \bar{\Omega} \quad (24)$$

Monotone Schemes and Convergence

Monotone schemes, or schemes of *positive type*, were introduced by Motzkin and Wasow [26] for linear equations and later extended to nonlinear equations (see [27]). Monotone schemes satisfy the discrete maximum principle (under natural assumptions) and the principal error term in the formal error expansion is elliptic. Hence, the schemes produce “smooth” and

monotone approximations to a regularized version of the original problem.

The main advantage of monotone schemes is that they “always” converge to the physically relevant solution [10]. For nonlinear or degenerate elliptic/parabolic equations, weak solutions are not unique in general, and extra conditions are needed to select the physically relevant solution. Nonmonotone schemes do not converge in general [27], and can even produce nonphysical solutions [29].

The main *disadvantage* of monotone schemes is the low order of convergence: first order for first-order PDEs and at most second order for second-order PDEs [27].

The main result of this section is the general convergence result of Barles and Souganidis for monotone, consistent, and stable schemes. We write a numerical scheme for the boundary value problem (22) or (11) as

$$S(h, x, u_h(x), u_h) = 0 \quad \text{on } \Omega_h \quad (25)$$

where S is a real-valued function defined on $\mathbb{R}^+ \times \Omega_h \times \mathbb{R} \times B(\Omega_h)$ where $B(\Omega_h)$ is the set of bounded functions on Ω_h . Typically, $\{u_h(x), u_h\}$ is the stencil of the method and u_h denotes the values at the neighbors of x . The “grid” Ω_h satisfies

$\Omega_h \subset \Omega$ is closed and

$$\lim_{h \rightarrow 0} \Omega_h = \{x : \exists \{x_h\}, x_h \in \Omega_h, x_h \rightarrow x\} = \bar{\Omega} \quad (26)$$

that is, any point in $\bar{\Omega}$ can be reached by a sequence of grid points. This assumption is satisfied for any natural family of grids, and it is necessary to have convergence. The grid Ω_h may be discrete or continuous ($\Omega_h = \bar{\Omega}$) depending on the scheme. We assume that the scheme (25) is

Monotone: For any $h > 0$, $x \in \Omega_h$, r , and $u, v \in B(\Omega_h)$

$$u \geq v \implies S(h, x, r, u) \leq S(h, x, r, v) \quad (27)$$

Consistent: For any smooth function ϕ ,

$$\begin{aligned} & \liminf_{\substack{h \rightarrow 0, \xi \rightarrow 0 \\ \Omega_h \ni x_h \rightarrow x}} S(h, x_h, \phi(x_h) + \xi, \phi + \xi) \\ & \geq H_*(x, \phi(x), D\phi(x), D^2\phi(x)) \quad \text{and} \end{aligned}$$

Table 1 Monotone explicit and implicit finite-difference schemes

Equation	Scheme	CFL
$u_t = \sigma^2 u_{xx}$ $\frac{2\sigma^2 \Delta t}{\Delta x^2} \leq 1$ $-$	$\begin{cases} u_m^{n+1} = u_m^n + \frac{\sigma^2 \Delta t}{\Delta x^2} [u_{m+1}^n - 2u_m^n + u_{m-1}^n] \\ u_m^{n+1} = u_m^n + \frac{\sigma^2 \Delta t}{\Delta x^2} [u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}] \end{cases}$	
$u_t = H(u_x)$	$u_m^{n+1} = u_m^n + \Delta t \left[H \left(\frac{u_{m+1}^n - u_{m-1}^n}{\Delta x} \right) + \ H'\ \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\Delta x} \right]$	$\frac{2\ H'\ \Delta t}{\Delta x} \leq 1$

$$\begin{aligned} & \limsup_{\substack{h \rightarrow 0, \xi \rightarrow 0 \\ \Omega_h \ni x_h \rightarrow x}} S(h, x_h, \phi(x_h) + \xi, \phi + \xi) \\ & \leq H^*(x, \phi(x), D\phi(x), D^2\phi(x)) \end{aligned} \quad (28)$$

Stable (L^∞ stable): For any $h > 0$, there is a solution u_h of equation (25). Moreover, u_h is uniformly bounded, that is, there is a constant $K \geq 0$ such that for any $h > 0$,

$$|u_h| \leq K \quad \text{in} \quad \Omega_h \quad (29)$$

Example 3 Monotone, consistent, and stable schemes are given in Table 1. Note that an explicit scheme is monotone (and stable) only when a CFL condition hold.

Here, $u_m^n \approx u(x_m, t_n)$ and $\|\cdot\| = \|\cdot\|_{L^\infty}$. The second equation is a Hamilton–Jacobi equation and the corresponding scheme is called the *Lax-Friedrich scheme* [15].

We assume that the boundary value problem (22) satisfies the following:

Strong Comparison Result: If $u, v \in B(\bar{\Omega})$ are upper and lower semicontinuous respectively, u is a subsolution of equation (22), and v is a supersolution of equation (22), then

$$u \leq v \quad \text{in} \quad \Omega \cup \Gamma \quad \text{where} \quad \Gamma \subset \partial\Omega \quad (30)$$

Under the above assumptions, we have the following convergence result:

Theorem 1 *The solution u_h of equation (25) converge locally uniformly (uniformly on compact subsets) in $\Omega \cup \Gamma$ to the unique viscosity solution of equation (22).*

The result is due to Barles and Souganidis and it gives sufficient conditions for locally uniform convergence of approximations. It applies to a very large class of equations, initial/boundary conditions, and (monotone) schemes, see for example, [4, 10, 19]. The only regularity required of the approximating sequence is uniform boundedness. The proof makes use of the Barles-Perthame method of “half relaxed limits” and can be found in [10].

Outline of proof: Define

$$\begin{aligned} \underline{u}(x) &= \liminf_{\substack{h \rightarrow 0 \\ \Omega_h \ni x_h \rightarrow x}} u_h(x_h) \\ \text{and} \quad \bar{u}(x) &= \limsup_{\substack{h \rightarrow 0 \\ \Omega_h \ni x_h \rightarrow x}} u_h(x_h) \end{aligned} \quad (31)$$

These functions are upper and lower semicontinuous in $\bar{\Omega}$, and by monotonicity and consistency they are respectively sub- and supersolutions of the limiting equation (22). Then, by the strong comparison result it follows that $\bar{u} \leq \underline{u}$ in $\Omega \cup \Gamma$. But by definition $\bar{u} \geq \underline{u}$ in $\bar{\Omega}$, and hence we have

$$\bar{u} = \underline{u} \quad \text{in} \quad \Omega \cup \Gamma \quad (32)$$

This immediately implies pointwise convergence. Locally, uniform convergence is followed by a variation of Dini’s theorem.

Remark 1 In Theorem 1 and the strong comparison result, the set Γ always contains all regular points on the boundary ($\partial\Omega_{\text{reg}} \subset \Gamma$) and may equal $\emptyset$, $\partial\Omega$, or a proper subset of $\partial\Omega$. It is equal to $\partial\Omega$ for Neumann problems or when $\partial\Omega_{\text{reg}} = \partial\Omega$. For Dirichlet problems, we refer to papers given in Table 2 below for the precise definition of Γ . When

$\Gamma \neq \partial\Omega$, then the solution of equation (22) may be discontinuous in $\partial\Omega \setminus \Gamma$, see the discussion in the section Viscosity Solutions. In this case, there will be formation of boundary layers in the numerical solutions that prevents (locally) uniform convergence in $\partial\Omega \setminus \Gamma$.

Example 4 A boundary value problem for a heat equation on $(0, 1)$ and its finite-difference approximation may be written in the forms (22) and (25) choosing

$$\begin{aligned} G(u, u_t, u_{xx}) \\ = \begin{cases} u_t - \sigma^2 u_{xx} + bu_x, & t > 0, x \in (0, 1) \\ u, & t > 0, x \in \{0, 1\} \end{cases} \end{aligned} \quad (33)$$

where $b \geq 0$, and for $\Delta x = 1/N$, $(0, 1)_{\Delta x} = \{m\Delta x\}_{m=1}^{N-1}$, and $\Delta t = c\Delta x^2$

$$\begin{aligned} S(\Delta x, u_m^{n+1}, \{u_m^n, u_{m\pm 1}^n\}) \\ = \begin{cases} \frac{u_m^{n+1} - u_m^n}{c\Delta x^2} - \sigma^2 \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{\Delta x^2} \\ + b \frac{u_m^n - u_{m-1}^n}{\Delta x}, & x_m \in (0, 1)_{\Delta x} \\ u_m^{n+1}, & x_m \in \{0, 1\} \end{cases} \end{aligned} \quad (34)$$

The (explicit) scheme is monotone if $c \leq (1/2\sigma^2 + b\Delta x)$, and consistent:

$$\begin{aligned} |G[\phi] - S[\phi]| \leq \\ \begin{cases} \left[\|\phi_t\|c + \frac{\sigma^2}{12} \|\phi_{xxx}\| \right] \Delta x^2 \\ + \frac{b}{2} \|\phi_{xx}\| \Delta x, & \text{at } x_m \in (0, 1)_{\Delta x} \\ 0, & \text{at } x_m \in \{0, 1\} \end{cases} \end{aligned} \quad (35)$$

The scheme is also stable if $c \leq (1/2\sigma^2 + b\Delta x)$ and $\max_m |u_m^{n+1}| \leq \max_m |u_m^0|$. If $\sigma \neq 0$, $b \geq 0$ are constants, $u(x, 0)$ is continuous, and $u(0, 0) = u(1, 0)$, then the strong comparison result holds in $[0, 1]$ [13]. Hence by Theorem 1, the finite-difference solution converge locally uniformly as $\Delta x \rightarrow 0$ to the solution of the equation.

Example 5 (Degeneracy, boundary layers). In Example 4, we replace σ, b by functions, either

$$\begin{aligned} \text{(i) } b = 0, \sigma(x) = \begin{cases} \frac{1}{4} - x, & x \in (0, \frac{1}{4}) \\ 0, & x \in (\frac{1}{4}, \frac{3}{4}) \\ x - \frac{3}{4}, & x \in (\frac{3}{4}, 1) \end{cases} \\ \text{or } \text{(ii) } b = 1, \sigma(x) = x \end{aligned} \quad (36)$$

In both cases, the strong comparison result follows from [13], see also [5, 14].

In case (i), the equation degenerates for $x \in (\frac{1}{4}, \frac{3}{4})$ and the solution is not more than continuous here. But there are no degeneracies at the boundary (it is regular) so the comparison result holds on $[0, 1]$ and Theorem 1 implies uniform convergence of the numerical solution in $[0, 1]$.

In case (ii), the equation degenerates only at the boundary $x = 0$. At this point, the exact solution will be discontinuous for $t > 0$, and the strong comparison result holds only on $(0, 1]$. This is also where the numerical solution will converge by Theorem 1. *Uniform* convergence in $x = 0$ is not possible because of formation of boundary layers as the numerical solution is refined, see also Example 2.

Remark 2 (The strong comparison result). A main difficulty in applying Theorem 1 is that the boundary value problem must satisfy the strong comparison result. There are general results that cover most (but probably not all) applications in finance, see Table 2.

In particular, the results of [13] cover Dirichlet (and state constrained) problems for linear and convex second-order PDEs (with or without time) when the domain satisfies an outer ball condition. This includes, for example, all box and convex polyhedral domains in $\mathbb{R}^n$.

Remark 3 (Assumptions on the scheme). See also the discussion in [4, 27].

Monotonicity: This condition is analogous to ellipticity of the equation [4, 10, 27], see the discussion at the beginning of this section. For approximations

Table 2 Strong comparison results

BC	Equation	Domain	Paper
Dirichlet	Linear in second derivatives	Smooth	[5]
	Convex fully nonlinear	Smooth	[8]
	(includes linear equations)	Nonsmooth	[13]
	Second-order quasilinear	Smooth	[9]
Neumann/ Robin	Second-order quasi/fully nonlinear	—	see [14]

of stationary linear equations on a grid $\{x_m\}_m$, monotonicity means [26]:

$$u_m = \sum_{k \neq 0} a_k u_{m+k} \quad \text{for } a_k \geq 0 \quad (u_m \approx u(x_m)) \quad (37)$$

Consistency: The strange formulation above is necessary since we consider the equation and the boundary conditions at the same time, see Example 4.

Stability: The type of stability required here is L^∞ stability and it is more restrictive than L^2 or von Neumann stability. For example, the Crank–Nicolson scheme is unconditionally von Neumann stable but not L^∞ stable in general [7].

Generally speaking, stability (in L^∞) follows if $S(h, x, r, v)$ defined in equation (25) is monotone and strictly increasing in r . This is typically the case for approximations of

- parabolic problems;
- degenerate elliptic problems where F def. in equation (4) is strictly increasing in u ;
- uniformly elliptic problems.

Some very general stability results can be found in [7, 27] for whole space problems, while [17, 25, 26] deal with more particular problems on domains.

Error Estimates for Convex Equations

For linear and nondegenerate problems satisfying condition (5), error estimates follow from classical (L^2) methods that can be found in most advanced textbooks on numerical solutions of PDEs. For degenerate and/or nonlinear problems these methods do not

apply, and there exists no general theory today. For equations (4) satisfying condition (5) there are results in the following cases:

1. General first-order equations [15];
2. Convex or concave second-order equations [7, 17];
3. Nondegenerate second-order equations [12].

In the first case, and to some degree also in the second case, there are rather satisfactory and extensive results. In the rest of this section, we concentrate on the case (ii), since most PDEs in finance belong to this category (including linear ones). The first result in this direction came in two papers of Krylov [22, 23], and his ideas have been developed and improved by several authors since, we refer to [7, 17] for the most general results at this time.

In what follows, we rewrite the available results in a framework inspired by Barles and Jakobsen [6, 7] and present them in the context of the following possibly degenerate, fully nonlinear, convex model equation:

$$u_t + \sup_{\vartheta \in \Theta} \left\{ -\text{tr}[\sigma^\vartheta \sigma^{\vartheta T} D^2 u] - b^\vartheta Du + c^\vartheta u + f^\vartheta \right\} = 0 \quad \text{in } \mathbb{R}^N \times (0, T) \quad (38)$$

where σ (matrix), b (vector), c , f are functions of t, x, ϑ , and $^T/\text{tr}$ denotes transpose/trace of matrices. Note that equation (38) is linear if Θ is a singleton. All the results in this section requires that the initial value problem for equation (38) has a unique solution, which is Lipschitz continuous in x uniformly in t . This is the case [7] if, for example,

$$|u(\cdot, 0)|_1 + |\sigma^\vartheta|_1 + |b^\vartheta|_1 + |c^\vartheta|_1 + |f^\vartheta|_1 \leq K \quad \text{for some } K \text{ independent of } \vartheta \quad (39)$$

where $|\phi|_1 = \sup_{x,t} |\phi(x, t)| + \sup_{x \neq y, t} (|\phi(x, t) - \phi(y, t)|/|x - y|)$. Without loss of generality we also assume that $c^\vartheta \geq 0$.

We approximate equation (38) by a scheme (25), which we assume to be as follows:

Monotone and parabolic: $\phi(t) \in C^1$, $u \leq v$ in Ω_h

$$\begin{aligned} \implies S(h, t, x, r + \phi(t), u + \phi) &\geq S(h, t, x, r, v) \\ &+ \phi'(t) - \frac{Kh^2}{2} \|\phi''\|_\infty \end{aligned} \quad (40)$$

Continuous: $S(h, t, x, r, u)$ uniformly continuous in r uniformly in t, x .

Consistent: $\phi(t, x) \in C^\infty$

$$\begin{aligned} \implies & |F(t, x, \phi, \partial_t \phi, D_x \phi, D_x^2 \phi) \\ & - S(h, t, x, \phi(t, x), \phi)| \\ & \leq \sum_{i,j} K_{i,j} \|\partial_t^i D_x^j \phi\|_\infty h^{\alpha_{i,j}} \end{aligned} \quad (41)$$

Here $K, K_{ij}, \alpha_{ij} \geq 0$ are constants independent of h and (t, x) . Under all of the previously mentioned assumptions, we have the following upper bound on the error:

Theorem 2 (Upper bound).

$$u - u_h \leq C_u \min_{\epsilon > 0} \left\{ \epsilon + \sum_{i,j} K_{i,j} \epsilon^{1-2i-j} h^{\alpha_{i,j}} \right\} \quad \text{in } \Omega_h \quad (42)$$

This result was first proved in [23], and we refer to [7] for a discussion on the present formulation. For the most common monotone finite-difference schemes, this result produces the upper bound $Kh^{1/2}$ [20, 24], which is optimal [16] in this setting.

To get a lower bound, we need additional assumptions as follows:

Convexity: $S(h, t, x, r, u)$ is convex in (r, u) and commutes with translations in x .

Approximation and regularity: For h small enough and $0 \leq \epsilon < 1$, there is unique solution u_h^ϵ of the scheme

$$\max_{0 \leq s \leq \epsilon^2, |e| \leq \epsilon} S(h, t + s, x + e, u_h^\epsilon(x), u_h^\epsilon) = 0 \quad \text{on } \Omega_h \quad (43)$$

where $u_h := u_h^0$ solves equation (25), and there is a constant C such that for all s, t, x, y, h, ϵ ,

$$\begin{aligned} |u_h^\epsilon(t, x) - u_h^\epsilon(s, y)| &\leq C(|t - s|^{1/2} + |x - y|), \\ |u_h^0(t, x) - u_h^\epsilon(t, x)| &\leq C\epsilon \end{aligned} \quad (44)$$

Under all of the above assumptions, we have the first lower bound:

Theorem 3 (Lower bound I).

$$u - u_h \geq -C_{l_1} \min_{\epsilon > 0} \left\{ \epsilon + \sum_{i,j} K_{i,j} \epsilon^{1-2i-j} h^{\alpha_{i,j}} \right\} \quad \text{in } \Omega_h \quad (45)$$

Alternatively, we may replace the last two assumptions on the scheme (25) by slightly stronger assumptions on the equation (38):

Continuity in ϑ : Θ is a separable metric space, and σ, b, c, f are continuous in ϑ for every (t, x) .

Then we have the second lower bound:

Theorem 4 (Lower bound II).

$$u - u_h \geq -C_{l_2} \min_{\epsilon > 0} \left\{ \epsilon^{1/3} + \sum_{i,j} K_{i,j} \epsilon^{1-2i-j} h^{\alpha_{i,j}} \right\} \quad \text{in } \Omega_h \quad (46)$$

The typical lower bounds produced by Theorems 3 and 4 are $Kh^{1/2}, Kh^{1/5}$, respectively. The first bound is again optimal, but the result applies only to particular schemes and equations [6, 20, 22, 24]. The second result is not optimal in general, but it applies to *any* consistent monotone scheme, see [7] for the most general results and a wider discussion. Theorem 3 was (essentially) stated in the present general form in [6, 20] and follows from arguments of [22, 23]. Theorem 4 was stated and proved in [6].

Remark 4 (Approximation and regularity). Under quite general assumptions, see [24], it is possible to show that the ‘‘Approximation and regularity’’ assumption of Theorem 3 holds for any $\epsilon \in [0, 1)$ whenever it holds for $\epsilon = 0$, that is, what we need is a uniform in h Hölder estimate on the solution of u_h of the scheme (25).

Remark 5 (Proofs). There are 3–4 main ideas.

1. Mollification of the equation produce a smooth subsolution by convexity. An upper bound on the error then follows from classical L^∞ -argument using monotonicity and consistency [22].
2. The method of shaking the coefficients allows to treat general problems with variable coefficients [23].
3. The lower bound. Either you (i) interchange the role of the scheme and the equation in

part (1) to get bound I, or (ii) you introduce additional approximations to avoid working with the scheme and get a type II bound [7, 23]. In case (i), you need a uniform Lipschitz bound on the solutions of the scheme, which is very difficult to obtain in general.

Remark 6 (Extensions). Stationary problems have been considered in several papers, including some boundary value problems, *see* [7, 17]. There have been papers treating more general equations like parabolic obstacle problems, impulse control problems, and integro-differential problems (*see* **Partial Integro-differential Equations (PIDEs)**). When solutions are less regular (Hölder continuous), lower rates have been obtained in [6, 7] and in some cases when solutions are more regular, higher order of convergence can be obtained [16].

Example 6 Consider a special case of equation (38),

$$u_t + \sup_{\vartheta \in \Theta} \left\{ - \sum_{\beta} \left[(\bar{\sigma}_{\beta}^{\vartheta})^2 D_{\beta}^2 + \bar{b}^{\vartheta} D_{\beta} \right] u + c^{\vartheta} u + f^{\vartheta} \right\} = 0 \quad \text{in } \mathbb{R}^N \times (0, T) \quad (47)$$

where $\bar{b} \geq 0$, $\bar{\sigma}$ are scalar functions, $D_{\beta} = \beta \cdot D$ is a directional derivative, and $\{\beta\}$ is a finite collection of vectors in $\mathbb{Z}^N$. We approximate on a uniform grid $\Omega_h = h\mathbb{Z} \times ch^2\mathbb{Z}^+$ by an implicit finite-difference scheme proposed in [11, 24]

$$u_{\alpha}^{n+1} = u_{\alpha}^n + \Delta t \sup_{\theta \in \Theta} \left\{ - \sum_{\beta} \left[(\bar{\sigma}_{\beta}^{\theta})^2 \Delta_{h\beta} + \bar{b}^{\theta} \delta_{\beta h}^+ \right] \times u_{\alpha}^{n+1} + c^{\theta} u_{\alpha}^{n+1} - f^{\theta} \right\} = 0 \quad (48)$$

for $n \in \mathbb{Z}^+$, $\alpha \in \mathbb{Z}^n$, and where

$$\Delta_{h\beta} w(x) = \frac{w(x + h\beta) - 2w(x) + w(x - h\beta)}{h^2 |\beta|^2}$$

and $\delta_{h\beta}^+ w(x) = \frac{w(x + \beta h) - w(x)}{h |\beta|} \quad (49)$

This scheme is obviously monotone and a Taylor expansion shows that

$$|F(t, x, \phi, \partial_t \phi D\phi, D^2 \phi) - S(h, t, x, \phi(t, x), \phi)| \leq C(ch^2 |\phi_{tt}|_0 + h |D^2 \phi|_0 + h^2 |D^4 \phi|_0) \quad (50)$$

If $\bar{\sigma}, \bar{b}, c, f, u(0, \cdot)$ are uniformly x -Lipschitz, then u_h is also uniformly x -Lipschitz [24], and by Theorems 2 and 3 we get

$$|u - u_{\alpha}^n| \leq Ch^{1/2} \quad (51)$$

From a practical or probabilistic point of view, $\bar{\sigma}$ need not be Lipschitz. In this case, Theorems 2 and 4 yield a worse bound:

$$|u - u_{\alpha}^n| \leq Ch^{1/5} \quad (52)$$

Note that we have imposed the CFL condition $\Delta t = c\Delta x^2$ ($\Delta x = h$). If this condition is not satisfied, then the rates will be reduced. We refer to [7] for more general explicit and implicit schemes of this kind.

References

- [1] Bardi, M. & Capuzzo-Dolcetta, I. (1997). *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkhäuser.
- [2] Bardi, M., Crandall, M.G., Evans, L.C., Soner, H.M. & Souganidis, P.E. (1997). *Viscosity solutions and applications, Lecture Notes in Mathematics*, Springer-Verlag, Berlin, pp. 1660.
- [3] Barles, G. (1994). *Solutions de Viscosité des Equations de Hamilton-Jacobi, Mathématiques & Applications*, Springer-Verlag, Paris, p. 17.
- [4] Barles, G. (1997). Convergence of numerical schemes for degenerate parabolic equations arising in finance theory, in *Numerical Methods in Finance*, Newton Institute, Cambridge University Press, Cambridge, pp. 1–21.
- [5] Barles, G. & Burdeau, J. (1995). The Dirichlet problem for semilinear second-order degenerate elliptic equations and applications to stochastic exit time control problems. *Communications in Partial Differential Equations* **20**(1–2), 129–178.
- [6] Barles, G. & Jakobsen, E.R. (2002). On the convergence rate of approximation schemes for Hamilton–Jacobi–Bellman equations, *M2AN Mathematical Modelling and Numerical Analysis* **36**(1), 33–54.
- [7] Barles, G. & Jakobsen, E.R. (2007). Error bounds for monotone approximation schemes for parabolic Hamilton–Jacobi–Bellman equations, *Mathematics of Computation* **76**(260), 1861–1893.
- [8] Barles, G. & Rouy, E. (1998). A strong comparison result for the Bellman equation arising in stochastic exit time control problems and its applications, *Communications in Partial Differential Equations* **23**(11–12), 1995–2033.
- [9] Barles, G., Rouy, E. & Souganidis, P.E. (1999). Remarks on the Dirichlet problem for quasilinear elliptic and parabolic equations, in *Stochastic Analysis, Control, Optimization and Applications*, W.M. McEneaney, G.G. Yin & Q. Zhang, eds, *Systems & Control Foundations & Applications*, Birkhäuser, Boston, pp. 209–222.

- [10] Barles, G. & Souganidis, P.E. (1991). Convergence of approximation schemes for fully nonlinear second order equations, *Asymptotic Analysis* **4**(3), 271–283.
- [11] Bonnans, F. & Zidani, H. (2003). Consistency of generalized finite difference schemes for the stochastic HJB equation, *SIAM Journal of Numerical Analysis* **41**(3), 1008–1021.
- [12] Caffarelli, L.A. & Souganidis, P.E. (2008). A rate of convergence for monotone finite difference approximations to fully nonlinear, uniformly elliptic PDEs. *Communications on Pure Applied Mathematics* **61**(1), 1–17.
- [13] Chaumont, S. (2004). Uniqueness to elliptic and parabolic Hamilton–Jacobi–Bellman equations with non-smooth boundary, *C. R. Mathematical Academy of Science, Paris* **339**(8), 555–560.
- [14] Crandall, M.G., Ishii, H. & Lions, P.-L. (1992). User’s guide to viscosity solutions of second order partial differential equations, *Bulletin of the American Mathematical Society (N.S.)* **27**(1), 1–67.
- [15] Crandall, M.G. & Lions, P.-L. (1984). Two approximations of solutions of Hamilton–Jacobi equations, *Mathematics of Computation* **43**(167), 1–19.
- [16] Dong, H. & Krylov, N.V. (2005). Rate of convergence of finite-difference approximations for degenerate linear parabolic equations with C^1 and C^2 coefficients, *Electronic Journal of Differential Equations* **2005**(102), 1–25.
- [17] Dong, H. & Krylov, N.V. (2007). The rate of convergence of finite-difference approximations for parabolic Bellman equations with Lipschitz coefficients in cylindrical domains, *Applied Mathematics and Optimization* **56**(1), 37–66.
- [18] Evans, L.C. (1998). *Partial Differential Equations, Graduate Studies in Mathematics*, American Mathematical Society, Providence, p. 19.
- [19] Fleming, W.H. & Soner, H.M. (1993). *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York.
- [20] Jakobsen, E.R. (2003). On the rate of convergence of approximation schemes for Bellman equations associated with optimal stopping time problems, *Mathematical Models and Methods in Applied Science (M3AS)* **13**(5), 613–644.
- [21] Koike, S. (2004). *A Beginner’s Guide to the Theory of Viscosity Solutions, MSJ Memoirs*, Mathematical Society of Japan, Tokyo, p. 13.
- [22] Krylov, N.V. (1997). On the rate of convergence of finite-difference approximations for Bellman’s equations, *St. Petersburg Mathematical Journal* **9**(3), 639–650.
- [23] Krylov, N.V. (2000). On the rate of convergence of finite-difference approximations for Bellman’s equations with variable coefficients, *Probability Theory and Related Fields* **117**, 1–16.
- [24] Krylov, N.V. (2005). On the rate of convergence of finite-difference approximations for Bellman equations with Lipschitz coefficients, *Applied Mathematics and Optimization* **52**(2), 365–399.
- [25] Kuo, H. & Trudinger, N.S. (1995). Local estimates for parabolic difference operators, *Journal of Differential Equations* **122**(2), 398–413.
- [26] Motzkin, T.S. & Wasow, W. (1953). On the approximation of linear elliptic differential equations by difference equations with positive coefficients, *Journal of Mathematical Physics* **31**, 253–259.
- [27] Oberman, A.M. (2006). Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton–Jacobi equations and free boundary problems, *SIAM Journal of Numerical Analysis* **44**(2), 879–895.
- [28] Oleinik, O.A. & Radkevich, E.V. (1973). *Second Order Equations with Nonnegative Characteristic Form*, Plenum Press, New York–London.
- [29] Pooley, D.M., Forsyth, P.A. & Vetzal, K.R. (2003). Numerical convergence properties of option pricing PDEs with uncertain volatility, *IMA Journal of Numerical Analysis* **23**(2), 241–267.

Further Reading

- Kushner, H.J. & Dupuis, P. (2001). *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, New York.

ESPEN R. JAKOBSEN

Sparse Grids

The sparse grid method is a general *numerical discretization* technique for multivariate function representation, integration and partial differential equations. This approach, first introduced by the Russian mathematician Smolyak in 1963 [26], constructs a multidimensional multilevel basis by a special truncation of the tensor product expansion of a one-dimensional multilevel basis (see Figure 1 for an example of a sparse grid).

Discretizations on sparse grids involve only $O(N(\log N)^{d-1})$ degrees of freedom, where d is the problem dimension and N denotes the number of degrees of freedom in one coordinate direction. The accuracy obtained this way is comparable to one using a full tensor product basis involving $O(N^d)$ degrees of freedom, if the underlying problem is smooth enough, that is, if the solution has bounded mixed derivatives.

This way, the *curse of dimension*, that is, the exponential dependence of conventional approaches on the dimension d , can be overcome to a certain extent. This makes the sparse grid approach particularly attractive for the numerical solution of moderate and higher dimensional problems. Still, the classical sparse grid method is not completely independent of the dimension due to the above logarithmic term in the complexity.

Sparse grid methods are known under various names, such as hyperbolic cross points, discrete blending, Boolean interpolation, or splitting extrapolation. For a comprehensive introduction to sparse grids, see [5].

In computational finance, sparse grid methods have been employed for the valuation of *multiasset options* such as basket [24] (see **Basket Options**) or outperformance options [12], various types of *path-dependent derivatives*, due to the high dimension of the arising partial differential equations, or integration problems.

One-dimensional Multilevel Basis

The first ingredient of a sparse grid method is a one-dimensional multilevel basis. In the classical sparse grid approach, a hierarchical basis based on standard

hat functions,

$$\phi(x) := \begin{cases} 1 - |x| & \text{if } x \in [-1, 1], \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

is used. Then, a set of equidistant grids Ω_l of level l on the unit interval $\bar{\Omega} = [0, 1]$ and mesh width 2^{-l} is considered. The grid points $x_{l,i}$ are given by

$$x_{l,i} := i \cdot h_l, 0 \leq i \leq 2^l \quad (2)$$

The standard hat function is then taken to generate a family of basis functions $\phi_{l,i}(x)$ having support $[x_{l,i} - h_l, x_{l,i} + h_l]$ by dilation and translation, that is,

$$\phi_{l,i}(x) := \phi\left(\frac{x - i \cdot h_l}{h_l}\right) \quad (3)$$

Thereby, the index i indicates the location of a basis function or a grid point. This basis is usually termed as *nodal basis* or *Lagrange basis* (see Figure 2, bottom). These basis functions are then used to define function spaces V_l consisting of piecewise linear functions^a

$$V_l := \text{span} \{ \phi_{l,i} : 1 \leq i \leq 2^l - 1 \} \quad (4)$$

With these function spaces, the hierarchical increment spaces W_l ,

$$W_l := \text{span} \{ \phi_{l,i} : i \in I_l \} \quad (5)$$

using the index set

$$I_l = \{ i \in \mathbb{N} : 1 \leq i \leq 2^l - 1, i \text{ odd} \} \quad (6)$$

are defined. These increment spaces satisfy the relation

$$V_l = \bigoplus_{k \leq l} W_k \quad (7)$$

The basis corresponding to W_l is *hierarchical basis* (see Figure 2, top) and any function $u \in V_l$ can be uniquely represented as

$$u(x) = \sum_{k=1}^l \sum_{i \in I_k} v_{k,i} \cdot \phi_{k,i}(x) \quad (8)$$

with coefficient values $v_{k,i} \in \mathbf{R}$. Note that the supports of all basis functions $\phi_{k,i}$ spanning W_k are mutually disjoint.

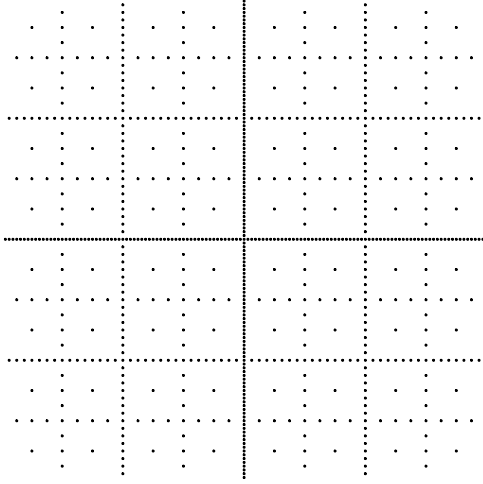


Figure 1 A regular two-dimensional sparse grid of level 7

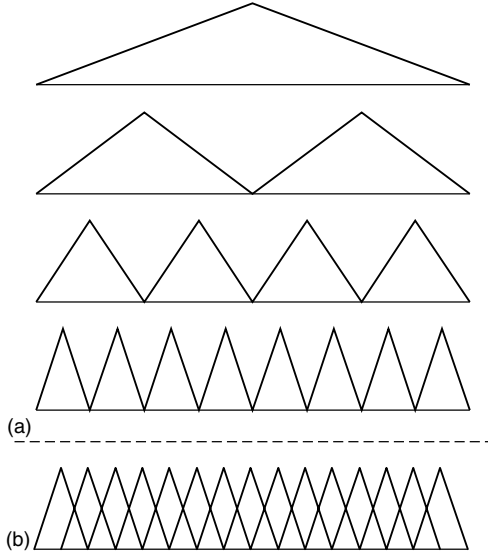


Figure 2 Piecewise linear hierarchical basis (a) versus nodal basis (b) of level 4

Tensor Product Construction

From this one-dimensional hierarchical basis, a multi-dimensional basis on the d -dimensional unit cube $\bar{\Omega} := [0, 1]^d$ is obtained by a tensor product construction. With the multiindex $\mathbf{l} = (l_1, \dots, l_d) \in \mathbf{N}^d$, which indicates the level in a multivariate sense, the set of d -dimensional standard rectangular grids $\Omega_{\mathbf{l}}$ on $\bar{\Omega}$ with mesh size $\mathbf{h}_{\mathbf{l}} := (h_{l_1}, \dots, h_{l_d}) := 2^{-\mathbf{l}}$ are

considered. Each grid $\Omega_{\mathbf{l}}$ is equidistant with respect to each individual coordinate direction, but, in general, may have varying mesh sizes in the different directions. The grid points $\mathbf{x}_{\mathbf{l}, \mathbf{i}}$ of the grid $\Omega_{\mathbf{l}}$ are the points

$$\mathbf{x}_{\mathbf{l}, \mathbf{i}} := (x_{l_1, i_1}, \dots, x_{l_d, i_d}), \quad \mathbf{1} \leq \mathbf{i} \leq 2^{\mathbf{l}} - \mathbf{1} \quad (9)$$

where for the above multiindices, all arithmetic operations are to be understood component-wise.

Then, for each grid point $\mathbf{x}_{\mathbf{l}, \mathbf{i}}$, an associated piecewise d -linear basis function $\phi_{\mathbf{l}, \mathbf{i}}(\mathbf{x})$ (see Figure 3) is defined as the product of the one-dimensional basis functions

$$\phi_{\mathbf{l}, \mathbf{i}}(\mathbf{x}) := \prod_{j=1}^d \phi_{l_j, i_j}(x_j) \quad (10)$$

Each of the multidimensional (nodal) basis functions $\phi_{\mathbf{l}, \mathbf{i}}$ has a support of size $2 \cdot \mathbf{h}_{\mathbf{l}}$. These basis functions are again used to define function spaces $V_{\mathbf{l}}$ consisting of piecewise d -linear functions, which are 0 on the boundary of $\bar{\Omega}$,

$$V_{\mathbf{l}} := \text{span} \{ \phi_{\mathbf{l}, \mathbf{i}} : \mathbf{1} \leq \mathbf{i} \leq 2^{\mathbf{l}} - \mathbf{1} \} \quad (11)$$

Similar to the one-dimensional case, the hierarchical increments $W_{\mathbf{l}}$ are defined by

$$W_{\mathbf{l}} := \text{span} \{ \phi_{\mathbf{l}, \mathbf{i}} : \mathbf{i} \in \mathbf{I}_{\mathbf{l}} \} \quad (12)$$

with the index set

$$\mathbf{I}_{\mathbf{l}} := \{ \mathbf{i} \in \mathbf{N}^d : \mathbf{1} \leq \mathbf{i} \leq 2^{\mathbf{l}} - \mathbf{1}, \\ i_j \text{ odd for all } 1 \leq j \leq d \} \quad (13)$$

This way, the hierarchical increment spaces $W_{\mathbf{l}}$ are related to the nodal spaces $V_{\mathbf{l}}$ by

$$V_{\mathbf{l}} = \bigoplus_{\mathbf{k} \leq \mathbf{l}} W_{\mathbf{k}} \quad (14)$$

Again, the supports of all multidimensional hierarchical basis functions $\phi_{\mathbf{l}, \mathbf{i}}$ spanning $W_{\mathbf{l}}$ are mutually disjoint. Also, again each function $u \in V_{\mathbf{l}}$ can uniquely be represented by

$$u_{\mathbf{l}}(\mathbf{x}) = \sum_{\mathbf{k}=1}^{\mathbf{l}} \sum_{\mathbf{i} \in \mathbf{I}_{\mathbf{k}}} v_{\mathbf{k}, \mathbf{i}} \cdot \phi_{\mathbf{k}, \mathbf{i}}(\mathbf{x}) \quad (15)$$

with hierarchical coefficients $v_{\mathbf{k}, \mathbf{i}} \in \mathbf{R}$.

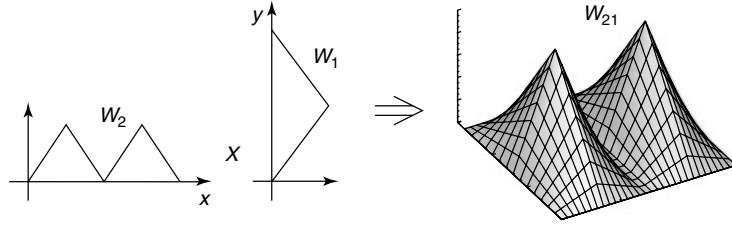


Figure 3 Tensor product approach to generate the piecewise bilinear basis functions $\phi_{(2,1),(1,1)}$ and $\phi_{(2,1),(1,1)}$ from the one-dimensional basis functions $\phi_{2,1}$, $\phi_{2,2}$ and $\phi_{1,1}$

Classical Sparse Grids

The classical sparse grid construction arises from a cost-to-benefit analysis in function approximation. Thereby, functions $u : \Omega \rightarrow \mathbf{R}$ which have bounded mixed second derivatives

$$D^\alpha u := \frac{\partial^{|\alpha|_1} u}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} \quad (16)$$

for $|\alpha|_\infty \leq 2$ are considered. These functions belong to the Sobolev space $H_2^{\text{mix}}(\bar{\Omega})$ with

$$H_2^{\text{mix}}(\bar{\Omega}) := \{u : \bar{\Omega} \rightarrow \mathbf{R} : D^\alpha u \in L_2(\Omega), \\ \times |\alpha|_\infty \leq 2, u|_{\partial\Omega} = 0\} \quad (17)$$

Here, the two norms $|\alpha|_1$ and $|\alpha|_\infty$ for multiindices are defined by

$$|\alpha|_1 := \sum_{j=1}^d \alpha_j \text{ and } |\alpha|_\infty := \max_{1 \leq j \leq d} \alpha_j \quad (18)$$

For functions $u \in H_2^{\text{mix}}(\bar{\Omega})$, the hierarchical coefficients $v_{1,i}$ decay as

$$|v_{1,i}| = O(2^{-2|\mathbf{l}_1|}) \quad (19)$$

On the other hand, the size (i.e., the number of degrees of freedom) of the subspaces W_1 is given by

$$|W_1| = O(2^{|\mathbf{l}_1|}) \quad (20)$$

An optimization with respect to the number of degrees of freedom and the resulting approximation accuracy directly leads to *sparse grid* spaces $\hat{V}_n$ of level n defined by

$$\hat{V}_n := \bigoplus_{|\mathbf{l}_1| \leq n+d-1} W_1 \quad (21)$$

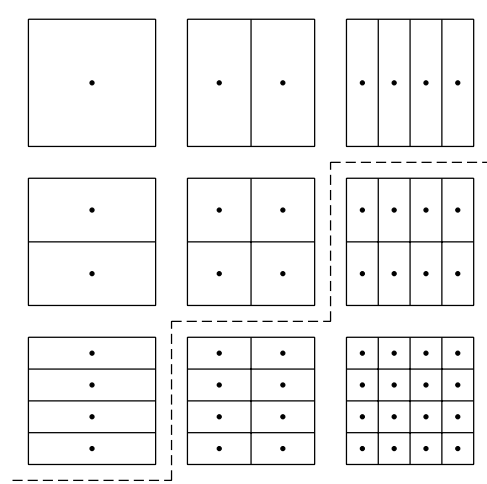


Figure 4 All subspaces W_1 for levels $|\mathbf{l}|_\infty \leq 3$ which together form the full grid space V_3 . The corresponding sparse grid space $\hat{V}_3$ consists of all subspaces above the dashed line ($|\mathbf{l}|_1 \leq 4$)

In comparison to the standard *full grid* space

$$V_n := V_{(n,\dots,n)} = \bigoplus_{|\mathbf{l}|_\infty \leq n} W_1 \quad (22)$$

which corresponds to cubic sectors of subspaces, sparse grids use triangular or simplicial sectors, see Figure 4. The dimension of the space $\hat{V}_n$, that is, the number of degrees of freedom or grid points is given by

$$|\hat{V}_n| = \sum_{i=0}^{n-1} 2^i \cdot \binom{d-1+i}{d-1} \\ = O(h_n^{-1} \cdot |\log_2 h_n|^{d-1}) \quad (23)$$

This shows the order $O(2^n n^{d-1})$, which is a significant reduction of the number of degrees of freedom

and, thus, of the computational and storage requirement compared to the order $O(2^{nd})$ of the dimension of the full grid space $|V_n|$.

On the other hand, the approximation accuracy of the sparse grid spaces for functions $u \in H_2^{\text{mix}}(\bar{\Omega})$ is in the L_p norms for $1 \leq p \leq \infty$ given by

$$\|u - \hat{u}_n\|_p = O(h_n^2 \cdot n^{d-1}) \quad (24)$$

For the corresponding full grid spaces, the accuracy is

$$\|u - u_n\|_p = O(h_n^2) \quad (25)$$

This shows the crucial advantage of the sparse grid space $\hat{V}_n$ in comparison with the full grid space V_n : the number of degrees of freedom is significantly reduced, whereas the accuracy is only slightly deteriorated. This way, the curse of dimensionality can be overcome, at least to some extent. The dimension still enters through logarithmic terms both in the computational cost and the accuracy estimate as well as in the constants hidden in the order notation.

Extensions and Applications

The classical sparse grid concept has been generalized in various ways. First, there are special sparse grids, which are optimized with respect to the energy seminorm [4]. These energy-based sparse grids are further sparsified and possess a cost complexity of $O(h_n^{-1})$ for an accuracy of $O(h_n)$. Thus, the dependence on the dimension d in the order is completely removed (but is still present in the hidden constants [8]). A generalization to sparse grids, which is optimal with respect to other Sobolev norms, can be found in [13]. In case the underlying space is not known *a priori*, dimension-adaptive methods [11] can be applied to find optimized sparse grids.

The sparse grid approach based on piecewise linear interpolation can be generalized to higher order polynomial [5] or wavelet discretizations (e.g., interpolants or prewavelets) [15, 25], which allows to utilize additional properties (such as higher polynomial exactness or vanishing moments) of the basis.

Furthermore, sparse grid methods can be applied to nonsmooth problems by using spatially adaptive refinement methods [2], see Figure 5 for an adaptively refined sparse grid. Spatial adaptivity helps if the original smoothness conditions for the sparse grid approach are not fulfilled. This is the case in nearly

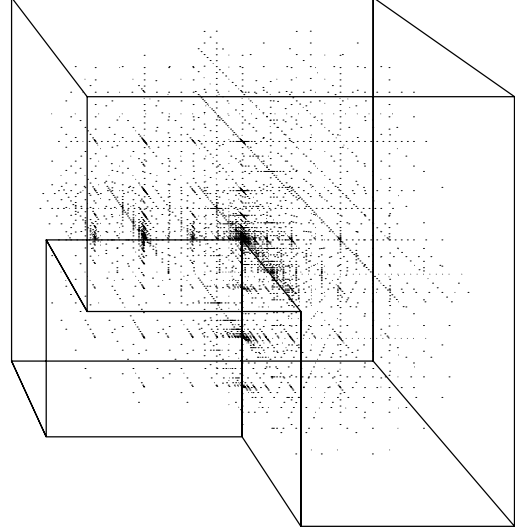


Figure 5 An at-a-corner singularity adaptively refined three-dimensional sparse grid

all option pricing problems since they lead to discontinuities in the initial conditions, which can, in some cases, extend into the interior of the domain. Here, adaptive refinement methods can often attain the same convergence rates as for smooth problems, which can be shown using approximation theory in Besov spaces [20]. Additionally, transformations that align areas of discontinuities with coordinate axes can significantly enhance the efficiency of sparse grid methods, as was shown in [24].

Sparse grids have been applied for the solution of different kinds of low- and moderate-dimensional partial differential equations, such as elliptic [5, 27], parabolic [1, 14], and hyperbolic [17] problems. In this context, finite element methods [2], finite difference methods [7], and finite volume methods [19] have been used in the discretization process.

For the solution of partial differential equations, often the so-called combination technique [9] is employed. Here, a sparse grid solution is obtained by a combination of anisotropic full grid solutions according to the combination formula

$$\hat{u}_n(\mathbf{x}) = \sum_{n \leq |\mathbf{l}|_1 \leq n+d-1} (-1)^{n+d-|\mathbf{l}|_1-1} \binom{d-1}{|\mathbf{l}|_1-n} u_{\mathbf{l}}(\mathbf{x}) \quad (26)$$

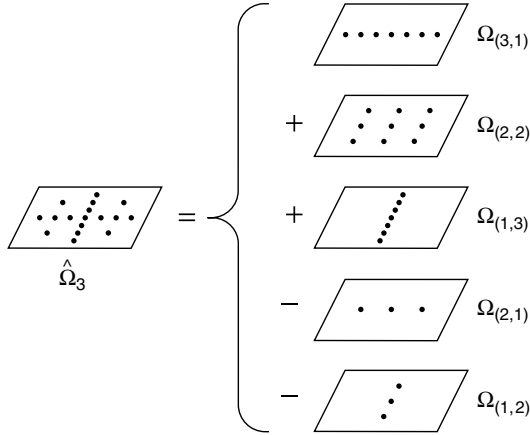


Figure 6 The combination technique in two dimensions for level $n = 3$: combine coarse full grids Ω_l , $|\mathbb{I}_l| \in \{3, 4\}$, with mesh widths 2^{-l_1} and 2^{-l_2} to get a sparse grid $\hat{\Omega}_n$ corresponding to $\hat{V}_n$

where $u_l(\mathbf{x})$ is a full grid solution on an anisotropic grid with mesh width 2^{-l} , see Figure 6 for a two-dimensional example. The combination technique can be further optimized with respect to the underlying differential operator [18].

The sparse grid approach can also be used for numerical integration, for example, for the computation of expectations [10, 23]. Thereby, the classical sparse grid construction starts with a sequence of one-dimensional quadrature formulas $Q_l f$ using n_l points to integrate a function f on the unit interval $[0, 1]$,

$$Q_l f := \sum_{i=1}^{n_l} w_{li} \cdot f(x_{li}) \quad (27)$$

Using the difference quadrature formulas

$$\Delta_k f := (Q_k - Q_{k-1})f \text{ with } Q_0 f := 0 \quad (28)$$

the sparse grid quadrature formula $\hat{Q}_n f$ of level n for a d -dimensional function f on the cube $[0, 1]^d$ is then defined by

$$\hat{Q}_n f := \sum_{|\mathbb{I}_l| \leq n+d-1} (\Delta_{l_1} \otimes \cdots \otimes \Delta_{l_d}) f \quad (29)$$

Again, this construction can be improved by using spatially adaptive or dimension-adaptive refinement [3, 11].

The sparse grid methodology has also been successfully applied to the solution of integral equations

[16], interpolation and approximation [21], and data analysis [6, 22].

End Notes

^aIn order to simplify this exposition, we assume that the functions in V_l are 0 on the boundary of $\bar{\Omega}$. This restriction can be overcome by adding appropriate boundary basis functions.

References

- [1] Balder, R. & Zenger, C. (1996). The solution of the multidimensional real Helmholtz equation on sparse grids, *SIAM Journal on Scientific Computing* **17**, 631–646.
- [2] Bungartz, H. (1992). An adaptive Poisson solver using hierarchical bases and sparse grids, in *Iterative Methods in Linear Algebra*, R. Beauwens, ed, North-Holland, pp. 293–310.
- [3] Bungartz, H. & Dimstorfer, S. (2003). Multivariate quadrature on adaptive sparse grids, *Computing* **71**, 89–114.
- [4] Bungartz, H. & Griebel, M. (1999). A note on the complexity of solving Poisson's equation for spaces of bounded mixed derivatives, *Journal of Complexity* **15**, 1–121.
- [5] Bungartz, H. & Griebel, M. (2004). Sparse grids, *Acta Numerica* **13**, 147–269.
- [6] Garcke, J., Griebel, M. & Thess, M. (2001). Data mining with sparse grids, *Computing* **67**, 225–253.
- [7] Griebel, M. (1998). Adaptive sparse grid multilevel methods for elliptic PDEs based on finite differences, *Computing* **61**, 151–179.
- [8] Griebel, M. (2006). Sparse grids and related approximation schemes for higher dimensional problems, in *Proceedings of FoCM05*, L. Pardo, A. Pinkus, E. Sulis & M. Todd, eds, Cambridge University Press.
- [9] Griebel, M., Schneider, M. & Zenger, C. (1992). A combination technique for the solution of sparse grid problems, in *Iterative Methods in Linear Algebra*, P. de Groen & R. Beauwens, eds, Elsevier, pp. 263–281.
- [10] Gerstner, T. & Griebel, M. (1998). Numerical integration using sparse grids, *Numerical Algorithms* **18**, 209–232.
- [11] Gerstner, T. & Griebel, M. (2003). Dimension-adaptive tensor-product quadrature, *Computing* **71**, 65–87.
- [12] Gerstner, T. & Holtz, M. (2008). Valuation of performance-dependent options, *Applied Mathematical Finance* **15**, 1–20.
- [13] Griebel, M. & Knappek, S. (2000). Optimized tensor-product approximation spaces, *Constructive Approximation* **16**, 525–540.
- [14] Griebel, M. & Oeltz, D. (2007). A sparse grid space-time discretization scheme for parabolic problems, *Computing* **81**, 1–34.

- [15] Griebel, M. & Oswald, P. (1995). Tensor product type subspace splitting and multilevel iterative methods for anisotropic problems, *Advances in Computational Mathematics* **4**, 171–206.
- [16] Griebel, M., Oswald, P. & Schiekofer, T. (1999). Sparse grids for boundary integral equations, *Numerische Mathematik* **83**, 279–312.
- [17] Griebel, M. & Zumbusch, G. (1999). Adaptive sparse grids for hyperbolic conservation laws, in *Hyperbolic Problems: Theory, Numerics, Applications*, M. Fey & R. Jeltsch, eds, Birkhäuser, pp. 411–422.
- [18] Hegland, M., Garcke, J. & Challis, V. (2007). The combination technique and some generalisations, *Linear Algebra and Its Applications* **420**, 249–275.
- [19] Hemker, P. (1995). Sparse-grid finite-volume multigrid for 3D-problems, *Advances in Computational Mathematics* **4**, 83–110.
- [20] Hochmuth, R. (2001). Wavelet characterizations of anisotropic Besov spaces, *Applied and Computational Harmonic Analysis* **12**, 179–208.
- [21] Klimke, A. & Wohlmuth, B. (2005). Algorithm 847: Spinterp: piecewise multilinear hierarchical sparse grid interpolation in MATLAB, *ACM Transactions on Mathematical Software* **31**, 561–579.
- [22] Laffan, S., Nielsen, O., Silcock, H. & Hegland, M. (2005). Sparse grids: a new predictive modelling method for the analysis of geographical data, *International Journal of Geographical Information Science* **19**, 267–292.
- [23] Novak, E. & Ritter, K. (1996). High dimensional integration of smooth functions over cubes, *Numerische Mathematik* **75**, 79–97.
- [24] Reisinger, C. & Wittum, G. (2007). Efficient hierarchical approximation of high-dimensional option pricing problems, *SIAM Journal on Scientific Computing* **29**, 440–458.
- [25] Schwab, C. & Todor, R. (2003). Sparse finite elements for stochastic elliptic problems: higher order moments, *Computing* **71**, 43–63.
- [26] Smolyak, S. (1963). Interpolation and quadrature formulas for the classes W_s^a and E_s^a , *Soviet Mathematics Doklady* **4**, 240–243.
- [27] Zenger, C. (1991). Sparse grids, in *Parallel Algorithms for Partial Differential Equations*, W. Hackbusch, ed, Vieweg, pp. 241–251.

Related Articles

Finite Difference Methods for Barrier Options; Finite Difference Methods for Early Exercise Options; Finite Element Methods; Wavelet Galerkin Method.

THOMAS GERSTNER & MICHAEL GRIEBEL

Optimization Methods

No area of computational mathematics plays a greater role in the support of financial decision making and strategy development than numerical optimization. The full gamut of optimization methodologies are applied to this end—linear and quadratic, nonlinear and global, stochastic and deterministic, discrete and continuous—and applications propagate throughout the front office, back office, analysis, and trading operations. The use of optimization is deep, pervasive, and growing.

We organize our presentation in three “levels”. The top level in our presentation is the management of portfolios based on quadratic programming, the second level is stochastic programming for portfolio optimization, and the lowest (but perhaps the most important) level is model calibration (*see* **Model Calibration**).

While some of the methodological issues that arise are specific to the different levels, there are three dominant themes that cut across all levels: speed, robustness, and quality of solution. Speed often dominates thinking in financial circles since rapid informed decision making (sometimes, “automatic”) can translate into capital gains (conversely, lack of sufficient speed can lead to losses). Nevertheless, a “wrong” answer computed at record speed is not of much value (and could be quite disastrous). A solution can be wrong in several ways. For example, if the computed solution is not robust, then the resultant strategy may be a very poor strategy under slight tweaking of the parameters defining the problem. This is a serious practical problem because problem parameters are almost always determined in an approximate manner (i.e., they are not known exactly). Other solution quality issues can arise. For example, some optimization problems are too hard to solve in reasonable time and so approximation schemes must be used. But how good is the approximate solution?

As we discuss some of the optimization challenges that arise under the organizing levels mentioned above, we make particular note of these three unifying numerical concerns.

Portfolio Management: Quadratic Programming

The most famous optimization application in finance is the mean–variance portfolio optimization problem, first introduced by Markowitz [19]; *see also* **Risk–Return Analysis**. The question that is addressed by “mean–variance portfolio optimization” is both easy to understand and practical: how should one distribute, across a given set of financial instruments, a finite investment in order to balance (according to the investors preference) risk and expected return? In its pure form, this question can be formulated as a positive definite quadratic programming problem.

Let $\mu \in \Re^n$ be a vector of expected returns for n assets, and an n -by- n matrix Q be the covariance matrix of asset returns. Assume that the vector $x \in \Re^n$ denotes the percentage of asset holdings. Then the mean-variance portfolio optimization problem can be formulated as

$$\begin{aligned} \min_x \quad & -\mu^T x + \lambda x^T Q x \\ \text{subject to} \quad & \sum_{i=1}^n x_i = 1 \end{aligned} \quad (1)$$

where $\lambda \geq 0$ is a risk aversion parameter. Additional (linear) constraints can be imposed, for example, no short selling constraints correspond to $x \geq 0$. There are many good algorithms, and codes (e.g., MOSEK and LOQO [24]) to solve positive definite quadratic programming problems, but the situation becomes more complex (and more interesting) as financial (and numerical) concerns are introduced.

One complication arises in the equity setting: portfolios held in many firms can be quite large—in several thousands—and typically have a dense matrix Q ; nevertheless, there is serious need to determine a solution rapidly. Moreover, because many of the subsets of instruments in the portfolio behave in a highly correlated way the matrix Q can be ill-conditioned. This means numerical algorithms can have a hard time computing accurate answers, and small changes in the input data can lead to very different proposed strategies and portfolios. One approach to address these difficulties is to use a factor model. However, algorithm implementation needs to exploit this special structure of the covariance matrix for optimal computational efficiency.

Another complication is the need to use additional terms, for example, to capture transaction costs (see **Transaction Costs**). These additional terms yield a more realistic model; however, they may also change the objective function into a more nonlinear function. This apparent small change has a big impact—general nonlinear codes (even with linear constraints) are more complex and have fewer “guarantees” relative to quadratic objective functions.

Recently, there has been considerable attention paid to the fact that the objective function in equation (1) is just an approximation to reality. Estimation of expected returns is notoriously difficult; the expected return parameter may be closer to wishful thinking than reality. Specifically, return μ and covariance matrix Q are supposed to represent the return and risk going forward in time but are typically, in fact, the return and risk going backward. The question is, how well does our chosen portfolio (i.e., the optimization solution based on these estimated parameters) perform as reality rolls forward under real conditions? Unfortunately, the answer is that it may not do very well at all; see for example, [5, 6].

There is now considerable attention being paid to this very practical concern, generally under the label “robust optimization”; see for example, [14, 15, 23], and **Robust Portfolio Optimization**. The goal of robust optimization is to guarantee the best performance in the worse case. Since the support for an uncertain parameter may be infinite, a robust portfolio is typically determined by considering optimal performance in the worst case within some uncertainty sets for model parameters. For example, the min–max robust formulation for equation (1) can be expressed as

$$\begin{aligned} \min_x \quad & \max_{\mu \in S_\mu, Q \in S_Q} -\mu^T x + \lambda x^T Q x \\ \text{subject to} \quad & \sum_{i=1}^n x_i = 1 \end{aligned} \quad (2)$$

where S_μ and S_Q are uncertainty sets for the expected return μ and covariance matrix Q , respectively. The uncertainty sets are often either intervals or ellipsoids (typically corresponding to some confidence intervals). Efficient computational methods for conic optimization and semidefinite programming can be used to solve some of these robust optimization problems; see, for example, [11].

There is need for still more research here though since much of the work to date takes an unduly conservative—and expensive—point of view: given a range to capture the possibility values of the parameters, solve the problem in the worst case. This solution provides protection but is certainly on the extreme risk-averse side. Recently, a conditional Value-at-Risk (CVaR) robust formulation is considered in [25] to address uncertainty in the parameters for mean–variance portfolio optimization.

CVaR Minimization and Optimal Executions: Stochastic Programming

Optimal financial decisions often need to be made using uncertain parameters which describe the optimization problems. This view leads to stochastic programming problems.

Even in a single-period portfolio optimization framework, if instrument values (e.g., options) depend nonlinearly on the risk factors, a different risk measure, instead of standard deviation, for example, Value-at-Risk (VaR) (see **Value-at-Risk**) or CVaR (see **Expected Shortfall**), needs to be used. Both these measures quantify near-worst case losses and both present interesting optimization challenges.

VaR is essentially a quantile of a loss distribution. For a confidence level β , for example, 95%, the VaR of a portfolio is the loss in the portfolio’s market value over a specified time horizon that is exceeded with probability $1 - \beta$. When VaR is used as a risk measure, the portfolio optimization problem is, in general, a nonconvex programming problem. Computing a global minimizer remains a computationally challenging task.

An alternative risk measure to VaR is CVaR. When the distribution of the portfolio loss is continuous, for a given time horizon and a confidence level β , CVaR is the conditional expectation of the loss exceeding VaR. In contrast to VaR, CVaR provides additional information on the magnitude of the excess loss. It has been shown that CVaR is a coherent risk measure; see for example, [3, 21]. In addition, minimizing CVaR typically leads to a portfolio with a small VaR.

Assume that $L(x)$ is the random variable denoting loss of a portfolio $x \in \mathcal{R}^n$ within a given time horizon. If x is a vector of instrument holdings and δV is (random) change in the instrument values, then $L(x) = -x^T(\delta V)$. For a given confidence level,

CVaR is given by

$$\begin{aligned} & \text{CVaR}_\beta(L(x)) \\ &= \min_{\alpha} \left(\alpha + (1 - \beta)^{-1} \mathbf{E}((L(x) - \alpha)^+) \right) \end{aligned} \quad (3)$$

where $(L(x) - \alpha)^+ = \max(L(x) - \alpha, 0)$ and $\mathbf{E}(\cdot)$ denotes the expectation of a random variable. When the loss distribution is continuous, the above relation follows directly from the optimality condition [22]. Unlike VaR, the CVaR portfolio optimization problem,

$$\min_{x, \alpha} \left(\alpha + (1 - \beta)^{-1} \mathbf{E}((L(x) - \alpha)^+) \right) \quad (4)$$

is a convex optimization problem [22].

Assume that $\{(\delta V)_i\}_{i=1}^m$ are independent samples of the change in the instrument values over the given horizon. Then the following is a scenario CVaR optimization problem, which approximates the above continuous CVaR optimization problem:

$$\min_{(x, \alpha)} \left(\alpha + \frac{1}{m(1 - \beta)} \sum_{i=1}^m [-(\delta V)_i^T x - \alpha]^+ \right) \quad (5)$$

This piecewise linear optimization problem has an equivalent linear programming formulation, which can be solved using standard linear programming methods. The resulting linear program has $O(m + n)$ variables and $O(m + n)$ constraints, where m is the number of Monte Carlo samples and n is the number of instruments. Note that any additional linear constraints can easily be included. Although this linear programming problem can be solved using standard linear programming software, a smoothing technique is proposed in [1]; this smoothing method is shown to be significantly more computationally efficient when the number of instruments and scenarios become large.

Frequently, portfolio optimal decision problem is also a multistage dynamic programming problem. Recently, there has been much interest in the optimal execution of a portfolio of large trades under the market impact consideration; see, for example, [2, 4, 13].

The optimal execution problem can be formulated as a continuous time stochastic control problem. We illustrate the problem here in the discrete setting. Suppose that a financial institution wants to sell a

large number of shares $\bar{S} \in \mathbb{R}^m$ in m assets by trading at $t_0 = 0 < t_1 < \dots < t_N = T$, where $t_{i+1} - t_i = \tau = \frac{T}{N}$ and $T > 0$ is the time horizon. Let the trades between t_{k-1} and t_k be denoted by vectors n_k , $k = 1, 2, \dots, N$.

Let us assume that, at time t_k , $k = 0, 1, \dots, N$, the vector P_k are the prices per share of assets that are publicly available in the market and $\tilde{P}_k$ are the execution prices of one unit of the assets. The *execution cost* (see **Execution Costs**) of the trades is often defined as $P_0^T \bar{S} - \sum_{k=1}^N n_k^T \tilde{P}_k$. Owing to uncertainties in price movements and in realized prices, this implementation cost is a random variable. Hence, the mean-variance formulation of the execution cost problem with a risk-aversion parameter $\lambda \geq 0$ is

$$\begin{aligned} & \min_{n_1, n_2, \dots, n_N} \mathbf{E} \left(P_0^T \bar{S} - \sum_{k=1}^N n_k^T \tilde{P}_k \right) + \\ & \quad \lambda \cdot \mathbf{Var} \left(P_0^T \bar{S} - \sum_{k=1}^N n_k^T \tilde{P}_k \right) \\ & \text{s.t.} \quad \sum_{k=1}^N n_k = \bar{S} \\ & \quad n_k \geq 0, \quad k = 1, 2, \dots, N \end{aligned} \quad (6)$$

where $\mathbf{E}(\cdot)$ and $\mathbf{Var}(\cdot)$ denote the expectation and the variance of a random variable, respectively.

The complexity level of equation (6) depends on assumptions on the price dynamics and the impact functions. In [2], the price vector P_k is assumed to follow the dynamics:

$$P_k = P_{k-1} + \tau^{1/2} \Sigma \xi_k - \tau g \left(\frac{n_k}{\tau} \right) \quad (7)$$

where $\xi_k^T \in \mathbb{R}^l$ represents an l -vector of independent standard normals, and Σ is an $m \times l$ volatility matrix of the asset prices. The m -vector function $g(\cdot)$ measures the permanent price impact (see **Price Impact**), which is, in general, relatively small. The execution prices are given by

$$\tilde{P}_k = P_{k-1} - h \left(\frac{n_k}{\tau} \right) \quad (8)$$

where the m -vector nonlinear function $h(\cdot)$ describes the temporary price impact.

Even in this simple price dynamic and market impact models, there are many interesting and important issues for the optimal execution problem. Price impact functions represent the *expected* price depression caused by trading assets at a unit rate. Estimating both temporary and permanent impact functions can incur large estimation errors. The sensitivity of the optimal execution strategy to the estimation error in the impact matrices has recently been studied in [20]. In addition, if the price dynamics and impact functions depend on additional state variables as considered in [4], solving a portfolio execution problem with many assets is computationally challenging, especially when no short selling constraints are imposed.

Nonlinear Programming

One of the most active roles optimization plays in finance is the calibration of models (*see Model Calibration*) yield curve construction (*see Yield Curve Construction*) and statistical estimation problems (*see Generalized Method of Moments (GMM); Entropy-based Estimation; Simulation-based Estimation*). Mathematical models are used to represent the behavior of financial instruments, and portfolios of such instruments, and such models almost always require parameters to be estimated. These parameters can be scalars, vectors, matrices, tensors, lines, curves, and surfaces. The estimation processes can lead to linear, nonlinear, convex, and nonconvex optimization problems (see examples in **Model Calibration**).

The usual situation leads to a data-fitting problem: given a model with unknown parameters, and given some real data (say, market prices), determine the “best” value for the parameters. An important class of such problems is the option model calibration problem in which one determines a model so that model values best fit market prices. Such problems are known as *inverse* problems and there is a significant literature on the creation and solution of inverse problems in engineering.

To illustrate, assume that a family of models are described by the model parameters x in a feasible set Ω . The feasible set constraints (such as non-negativity, upper-bound constraints) can be used to impose certain conditions on the model parameters.

Calibration problems determine the best fit to the market option prices; the best fitting parameters can be determined by solving the following nonlinear least-squares problem:

$$\min_{x \in \Omega} \frac{1}{2} \sum_{j=1}^m (V_0(K_j, T_j; x) - V_0^{\text{mkt}}(K_j, T_j))^2 \quad (9)$$

where $V_0^{\text{mkt}}(K_j, T_j)$ denote today’s market prices for standard options with strike K_j and expiry T_j , $j = 1, \dots, m$, and $\{V_0(K_j, T_j; x), j = 1, \dots, m\}$ denote today’s model values corresponding to model parameters x .

Let $F(x) : \Re^n \rightarrow \Re^m$ denote the residual vector

$$F(x) \stackrel{\text{def}}{=} \begin{bmatrix} V_0(K_1, T_1; x) - V_0^{\text{mkt}}(K_1, T_1) \\ \vdots \\ V_0(K_m, T_m; x) - V_0^{\text{mkt}}(K_m, T_m) \end{bmatrix} \quad (10)$$

The calibration problem is a *nonlinear least-squares* problem

$$\min_{x \in \Omega} \frac{1}{2} \|F(x)\|_2^2$$

There are a host of numerical challenges and issues that arise in the calibration setting but we only mention a few of them here. The foremost, without a doubt, is the reliability of the data (and the volume of data to be used). Data reliability can lead to preprocessing steps such as filtering, and, in some cases, choosing an optimization formulation that is relatively insensitive to data errors (e.g., least-squares minimization is much more sensitive to (erroneous) outliers than absolute-value minimization).

Avoiding overfitting is also a major issue. In order for a model to calibrate some market information, for example, market option prices, one needs to consider a family of sufficiently complex models. For example, it is well known that the classical Black–Scholes model is inadequate to calibrate equity option prices and more complex models such as local volatility function models (*see Local Volatility Model*), jump models (*see Exponential Lévy Models*), and stochastic volatility models (*see Heston Model*) have been proposed. When a family of complex models such as local volatility function models are considered, it is crucial to avoid overfitting data; see for example, [8]. Even when

a family of models are described by a few model parameters, the question of whether there exists sufficient information to robustly determine model parameters still remains; see, for example, jump model calibration problems [10, 18]. See also **Tikhonov Regularization** for additional discussion on regularization techniques.

In addition, option model calibration problems face computational challenges. The problem is often nonconvex and it is possible for this calibration optimization problem to have multiple local minimizers; see for example, [17]. Also note that each initial model value $V_0(K_j, T_j; x)$ is a complex nonlinear function of the model parameters x . The Levenberg–Marquardt method or Gauss–Newton method can be used to solve the nonlinear least-squares problem; see, for example, [12]. If the calibration problem has additional bound constraints, an interior point trust region method [7] can be applied. Genetic algorithms have also been used for the calibration problem; see, for example, [17]. Optimization software for this nonlinear least-squares problem requires repeated evaluation of each initial model value $V_0(K_j, T_j; x)$, which is typically done through numerical computation methods for partial differential equations or using Monte Carlo simulations. A good initial guess for the model parameters can also be crucial in ensuring success in obtaining a solution. We note that automatic differentiation may also be a useful computational tool in accurately computing the Jacobian matrices ∇F , which are often required by an optimization software. For more information on automatic differentiation, see, for example, [9, 16].

References

- [1] Alexander, S., Coleman, T.F. & Li, Y. (2004). Derivative portfolio hedging based on CVaR, in *New Risk Measures in Investment and Regulation*, Szego, G., ed., Wiley, pp. 339–363.
- [2] Almgren, R. & Chriss, N. (2000/2001). Optimal execution of portfolio transactions, *Journal of Risk* **3**, 3.
- [3] Artzner, P., Delbaen, F., Eber, J.M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**, 203–228.
- [4] Bertsimas, D. & Lo, A.W. (1998). Optimal execution costs, *Journal of Financial Markets* **1**, 1–50.
- [5] Best, M.J. & Grauer, R.R. (1991). On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results, *The Review of Financial Studies* **4**, 315–342.
- [6] Broadie, M. (1993). Computing efficient frontiers using estimated parameters, *Annals of Operations Research* **45**, 21–58.
- [7] Coleman, T.F. & Li, Y. (1996). An interior, trust region approach for nonlinear minimization subject to bounds, *SIAM Journal on Optimization* **6**(2), 418–445.
- [8] Coleman, T.F., Li, Y. & Verma, A. (1999). Reconstructing the unknown local volatility function, *The Journal of Computational Finance* **2**(3), 77–102.
- [9] Coleman, T.F. & Verma, A. (2000). ADMIT-1: automatic differentiation and matlab interface toolbox, *ACM Transactions on Mathematical Software* **26**, 150–175.
- [10] Cont, R. & Tankov, P. (2004). Nonparametric calibration of jump-diffusion option pricing models, *The Journal of Computational Finance* **7**(3), 1–49.
- [11] Cornuejols, G. & Tütüncü, R.H. (2007). *Optimization methods in finance*, Cambridge.
- [12] Dennis, J.E. & Schnabel, R.B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall Series in Computational Mathematics, Prentice-Hall.
- [13] Engle, R. & Ferstenberg, R. (2006). *Execution Risk*. Technical report, National Bureau of Economic Research, Cambridge, MA.
- [14] Garlappi, L., Uppal, R. & Wang, T. (2007). Portfolio selection with parameter and model uncertainty: A multi-prior approach, *Review of Financial Studies* **20**, 41–81.
- [15] Goldfarb, D. & Iyengar, G. (2003). Robust portfolio selection problems, *Mathematics of Operations Research* **28**(1), 1–38.
- [16] Griewank, A. & Corliss, G. (eds) (1991). *Automatic Differentiation of Algorithms: Theory, Implementation and Applications*. SIAM Proceedings Series, SIAM.
- [17] Hamida, S.B. & Cont, R. (2005). Recovering volatility from option prices by evolutionary optimization, *Journal of Computational Finance* **8**, 1–34.
- [18] He, C., Kennedy, J.S., Coleman, T.F., Forsyth, P.A., Li, Y. & Vetzal, K. (2006). Calibration and hedging under jump diffusion, *Review of Derivative Research* **9**, 1–35.
- [19] Markowitz, H.M. (1959). *Portfolio Selection: Efficient Diversification of Investments*, John Wiley, New York.
- [20] Moazeni, S., Coleman, T.F. & Li, Y. (2007). *Optimal Portfolio Execution Strategies and Sensitivity to Price-impact Parameters' Perturbations*. Technical report, David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada.
- [21] Pflug, G.Ch. (2000). Some remarks on the value-at-risk and the conditional value-at-risk, in *Probabilistic Constrained Optimization: Methodology and Applications*, S. Uryasev, ed., Kluwer Academic Publishers.
- [22] Rockafellar, R.T. & Uryasev, S. (2000). Optimization of conditional value-at-risk, *Journal of Risk* **2**(3), 21–41.
- [23] Tütüncü, R.H. & Koenig, M. (2004). Robust asset allocation, *Annals of Operations Research* **132**(1), 157–187.

- [24] Vanderbei, R.J. (1999). LOQO: an interior-point method code for quadratic programming, *Optimization Codes and Software* **11**, 451–484.
- [25] Zhu, L., Coleman, T.F. & Li, Y. (2007). *Min-max Robust and CVaR Robust Mean-variance Portfolios*. Technical report, David R. School of Computer Science, University of Waterloo, Waterloo, Canada.

Related Articles

**Model Calibration; Risk–Return Analysis;
Stochastic Control; Tikhonov Regularization.**

THOMAS F. COLEMAN & YUYING LI

Lattice Methods for Path-dependent Options

Path-dependent options are options whose payoffs depend on some specific function F of the entire trajectory of the asset price $F_t = F(t, (S_u)_{u \leq t})$. The most well-known examples are the lookback options and Asian options. In a lookback option, the payoff function is dependent on the realized maximum or minimum price of the asset over a certain period within the life of the option. The Asian options are also called *average options* since the payoff depends on a prespecified form of averaging of the asset price over a certain period. Consider an arithmetic average Asian option that is issued at time 0 and expiring at $T > 0$, its terminal payoff is dependent on the arithmetic average A_T of the asset price process S_t over period $[0, T]$. The running average value A_t is defined as

$$A_t = \frac{1}{t} \int_0^t S_u du \quad (1)$$

with $A_0 = S_0$. We are interested in the correlated evolution of the path function with the asset price process. In the above example of arithmetic averaging, the law of evolution of A_t is given as

$$dA_t = \frac{1}{t} (S_t - A_t) dt \quad (2)$$

A variant of the lattice tree methods (binomial/trinomial methods), called the *forward shooting grid* (FSG) approach, has been successfully applied to price a wide range of strong path-dependent options, such as the lookback options, Asian options, convertible bonds with reset feature and Parisian feature, reset strike feature in shout options, and so on. The FSG approach is characterized by augmenting an auxiliary state vector at each node in the usual lattice tree, which serves to capture the path-dependent feature of the option. Under the discrete setting of lattice tree calculations, let G denote the function that describes the correlated evolution of F with S over the time step Δt , which can be expressed as

$$F_{t+\Delta t} = G(t, F_t, S_{t+\Delta t}) \quad (3)$$

For example, let A^n denote the discretely observed arithmetic average defined as

$$A^n = \frac{\sum_{i=0}^n S^i}{n+1} \quad (4)$$

where S^i is the observed asset price at time t_i , $i = 0, 1, \dots, n$. The correlated evolution of A^{n+1} with S^{n+1} is seen to be

$$A^{n+1} = A^n + \frac{S^{n+1} - A^n}{n+2} \quad (5)$$

Another example is provided by the correlated evolution of the realized maximum price M_t and its underlying asset price process S_t . Recall $M_t = \max_{0 \leq u \leq t} S_u$ so that

$$M_{t+\Delta t} = \max(M_t, S_{t+\Delta t}) \quad (6)$$

In the construction of the auxiliary state vector, it is necessary to know the number of possible values that can be taken by the path-dependent state variable. For the lookback feature, the realized maximum asset price is necessarily one of the values taken by the asset price in the lattice tree. However, the number of possible values for the arithmetic average grows exponentially with the number of time steps. To circumvent the problem of dealing with exceedingly large number of nodal values, the state vector is constructed such that it contains a set of predetermined nodal values that cover the range of possible values of arithmetic averaging. Since the realized arithmetic average does not fall on these nodal values in general, we apply interpolation between the nodal values as an approximation.

The FSG approach is pioneered by Hull and White [4] and Ritchken *et al.* [10] for pricing American and European style Asian and lookback options. Theoretical studies on the construction and convergence analysis of the FSG schemes are presented by Barraquand and Pudet [1], Forsyth *et al.* [3], and Jiang and Dai [5]. A list of various applications of the FSG approach in lattice tree algorithms for pricing strongly path-dependent options/derivative products is given as follows:

- options whose underlying asset price follows various kinds of GARCH processes [11];
- path-dependent interest rate claims [9];
- Parisian options, alpha-quantile options, and strike reset options [6];

- soft call requirement in convertible bonds [7];
- target redemption notes [2]; and
- employee stock options with repricing features [8].

In this article, we illustrate the application of the FSG lattice tree algorithms for pricing options with path-dependent lookback and Asian features, convertible bonds with the soft call requirement (Parisian feature), and call options with the strike reset feature.

Lookback Options

Let the risk neutral probabilities of upward, zero, and downward jump in a trinomial tree be represented by p_u , p_0 , and p_d , respectively. In the FSG approach for capturing the path dependence of the discrete asset price process, we append an augmented state vector at each node in the trinomial tree and determine the appropriate grid function that models the discrete correlated evolution of the path dependence. Let $V_{j,k}^n$ denote the numerical option value of the path-dependent option at the n th-time level and j upward jumps from the initial asset value S_0 . Here, k denotes the numbering index for the values assumed by the augmented state vector at the (n, j) th node in the trinomial tree. Let u and d denote the proportional upward and downward jump of the asset price over one time step Δt , with $ud = 1$. Let $g(k, j)$ denote the grid function that characterizes the discrete correlated evolution of the path-dependent state variable F_t and asset price process S_t . When applied to the trinomial tree calculations, the FSG scheme takes the following form:

$$V_{j,k}^n = e^{-r\Delta t} \left[p_u V_{j+1, g(k, j+1)}^{n+1} + p_0 V_{j, g(k, j)}^{n+1} + p_d V_{j-1, g(k, j-1)}^{n+1} \right] \quad (7)$$

where $e^{-r\Delta t}$ denotes the discount factor over one time step (Figure 1).

We consider the floating strike lookback option whose terminal payoff depends on the realized maximum of the asset price, namely, $V(S_T, M_T, T) = M_T - S_T$. The corresponding discrete analogy of the correlated evolution of M_t and S_t is given by the following grid function (equation 6):

$$g(k, j) = \max(k, j) \quad (8)$$

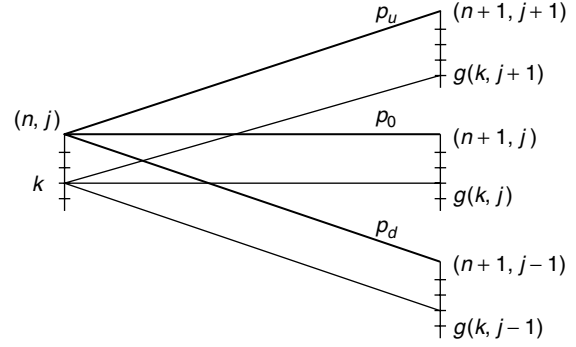


Figure 1 The discrete correlated evolution of the path-dependent state variable F_t and asset price process S_t is characterized by the grid function $g(k, j)$

As in usual trinomial calculations, we apply the backward induction procedure, starting with the lattice nodes at maturity. Suppose that there are a total of N time steps in the trinomial tree so that the maximum value of the discrete asset price process is $S_0 u^N$, corresponding to N successive jumps from the initial value S_0 . The possible range for realized maximum asset price would be $\{S_0, S_0 u, \dots, S_0 u^N\}$. When these possible values of the path-dependent state variable are indexed by k , then k assumes values from $0, 1, \dots$, to N . The terminal option value at the (N, j) th node and k th value in the state vector is given as

$$\begin{aligned} V_{j,k}^N &= S_0 u^k - S_0 u^j, \\ j &= -N, -N+1, \dots, N \text{ and} \\ k &= \max(j, 0), \max(j, 0) + 1, \dots, N \end{aligned} \quad (9)$$

Applying backward induction over one time step from expiry, the option values at the $(N-1)$ th time level are given as

$$\begin{aligned} V_{j,k}^{N-1} &= e^{-r\Delta t} \left[p_u V_{j+1, \max(k, j+1)}^N + p_0 V_{j, \max(k, j)}^N + p_d V_{j-1, \max(k, j-1)}^N \right] \\ j &= -N+1, -N+2, \dots, N-1, \\ k &= \max(j, 0) + 1, \dots, N-1 \end{aligned} \quad (10)$$

where the terminal option values are defined in equation (9). The backward induction procedure is then repeated to obtain numerical option values at the lattice nodes at earlier time levels. Note that the range

of the possible values assumed by the path-dependent state variable narrows as we proceed backward in a stepwise manner until we reach the tip of the trinomial tree.

Asian Options

Recall that the asset price S_j^n at the (n, j) th node in the trinomial tree is given as

$$S_j^n = S_0 u^j = S_0 e^{j\Delta W}, \quad j = -n, -n+1, \dots, n \quad (11)$$

where $u = e^{\Delta W}$ with $\Delta W = \sigma\sqrt{\Delta t}$. Here, σ is the volatility of the asset price. The average asset price at the n th time level must lie between $\{S_0 u^{-n}, S_0 u^n\}$. We take $\rho < 1$ and let $\Delta Y = \rho\Delta W$. Let $\text{floor}(x)$ denote the largest integer less than or equal to x and $\text{ceil}(x) = \text{floor}(x) + 1$. We set the possible values taken by the average asset price to be

$$A_k^n = S_0 e^{k\Delta Y}, \quad k = \text{floor}\left(-\frac{n}{\rho}\right), \dots, \text{ceil}\left(\frac{n}{\rho}\right) \quad (12)$$

The earlier FSG schemes choose ρ to be a sufficiently small number that is independent of Δt . The larger the value chosen for $1/\rho$, the finer the quantification of the average asset price. In view of numerical convergence of the FSG schemes, Forsyth *et al.* [3] propose to choose ρ to depend on $\sqrt{\Delta t}$ (say, $\rho = \lambda\sqrt{\Delta t}$, where λ is independent of Δt), though this would result in an excessive amount of computation in actual implementation. Further details on numerical convergence of various versions of the FSG schemes are presented later.

Suppose that the average is A_k^n and the asset price moves upward from S_j^n to S_{j+1}^{n+1} , then the new average is given as (equation 5)

$$A_{k^+}^{n+1} = A_k^n + \frac{S_{j+1}^{n+1} - A_k^n}{n+2} \quad (13)$$

Next, we set $A_{k^+}^{n+1}$ to be $S_0 e^{k^+(j)\Delta Y}$ for some value $k^+(j)$, that is,

$$k^+(j) = \frac{\ln A_{k^+}^{n+1}/S_0}{\Delta Y} \quad (14)$$

Note that $k^+(j)$ is not an integer in general, so $A_{k^+}^{n+1}$ does not fall onto one of the preset values for the average. Recall that $\text{floor}(k^+(j))$ is

the largest integer less than or equal to $k^+(j)$ and $\text{ceil}(k^+(j)) = \text{floor}(k^+(j)) + 1$. By the above construction, $A_{\text{floor}(k^+(j))}^{n+1}$ and $A_{\text{ceil}(k^+(j))}^{n+1}$ now fall onto the set of preset values. Similarly, we define

$$\begin{aligned} A_{k^-}^{n+1} &= A_k^n + \frac{S_{j-1}^{n+1} - A_k^n}{n+2} \\ A_{k^0}^{n+1} &= A_k^n + \frac{S_j^{n+1} - A_k^n}{n+2} \end{aligned} \quad (15)$$

corresponding to the new average at the $(n+1)$ th time level when the asset price experiences a downward jump and zero jump, respectively. In addition, $\text{floor}(k^-(j))$, $\text{ceil}(k^-(j))$, $\text{floor}(k^0(j))$, and $\text{ceil}(k^0(j))$ are obtained in a similar manner.

Let V_{j,k^+}^n denote the Asian option value at node (n, j) with the averaging state variable A_t assuming the value $A_{k^+}^n$, and assuming similar notation for $V_{j,\text{floor}(k^+(j))}^n$, and so on. In the lattice tree calculations, numerical option values for $V_{j,k}^n$ are obtained only for the case when k is an integer. Since $k^+(j)$ assumes a noninteger value in general, V_{j,k^+}^n is approximated through interpolation using option values at the neighboring nodes. Suppose that linear interpolation is adopted; we approximate V_{j,k^+}^n by the following interpolation formula:

$$V_{j,k^+}^n = \epsilon_{j,k}^+ V_{j,\text{ceil}(k^+(j))}^n + (1 - \epsilon_{j,k}^+) V_{j,\text{floor}(k^+(j))}^n \quad (16)$$

where

$$\epsilon_{j,k}^+ = \frac{\ln A_{k^+}^n - \ln A_{\text{floor}(k^+(j))}^n}{\Delta Y} \quad (17)$$

The FSG algorithm with linear interpolation for pricing an Asian option can be formulated as follows (Figure 2):

$$\begin{aligned} V_{j,k}^n &= e^{-r\Delta t} \left(p_u V_{j,k^+}^{n+1} + p_0 V_{j,k^0}^{n+1} + p_d V_{j,k^-}^{n+1} \right) \\ &= e^{-r\Delta t} \left\{ p_u \left[\epsilon_{j,k}^+ V_{j,\text{ceil}(k^+(j))}^{n+1} + (1 - \epsilon_{j,k}^+) V_{j,\text{floor}(k^+(j))}^{n+1} \right] \right. \\ &\quad + p_0 \left[\epsilon_{j,k}^0 V_{j,\text{ceil}(k^0(j))}^{n+1} + (1 - \epsilon_{j,k}^0) V_{j,\text{floor}(k^0(j))}^{n+1} \right] \\ &\quad + p_d \left[\epsilon_{j,k}^- V_{j,\text{ceil}(k^-(j))}^{n+1} + (1 - \epsilon_{j,k}^-) V_{j,\text{floor}(k^-(j))}^{n+1} \right] \left. \right\} \end{aligned} \quad (18)$$

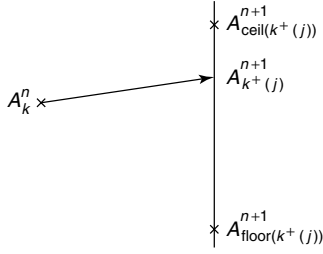


Figure 2 The average value A_k^n at the n th time step changes to $A_{k^+}^{n+1}$ at the $(n+1)$ th time step upon an upward move of the asset price from S_j^n to S_{j+1}^{n+1} . The option value at node $(n+1, j+1)$ with asset price average $A_{k^+}^{n+1}$ is approximated by linear interpolation between the option values with asset price average $A_{\text{floor}(k^+)}^{n+1}$ and $A_{\text{ceil}(k^+)}^{n+1}$

Numerical Convergence of FSG Schemes

Besides linear interpolation between two neighboring nodal values, other forms of interpolation can be adopted (say, quadratic interpolation between 3 neighboring nodal values or nearest node point interpolation). Forsyth *et al.* [3] remark that the FSG algorithm using ρ that is independent of Δt and the nearest node point interpolation may exhibit large errors as the number of time steps increases. They also prove that this choice of ρ in the FSG algorithm together with linear interpolation converges to the correct solution plus a constant error term. Unfortunately, the error term cannot be reduced by decreasing the size of the time step. To ensure convergence of the FSG calculations to the true Asian option price, they propose to use ρ that depends on $\sqrt{\Delta t}$, though this would lead to a large number of nodes in the averaging direction. More precisely, if ρ is independent of $\sqrt{\Delta t}$, then the complexity of the FSG method is $O(n^3)$, but convergence cannot be guaranteed. If we set $\rho = \lambda\sqrt{\Delta t}$, which guarantees convergence, then the complexity becomes $O(n^{7/2})$.

Soft Call Requirement in Callable Convertible Bonds

Most convertible bonds contain the call provision that allows the issuer to have the flexibility to manage the debt–equity ratio in the company’s capital structure. To protect the conversion premium paid upfront by the bondholders to be called away too soon, the bond indenture commonly contains the hard call protection

clause that prevents the issuer from initiating a call during the early life of the convertible bond. In addition, the soft call clause further requires the stock price to stay above the trigger price (typically 30% higher than the conversion price) for a consecutive or cumulative period before initiation of issuer’s call. The purpose of the soft call clause is to minimize the potential of market manipulation by the issuer.

The path-dependent feature that models the phenomenon of the asset price staying above some threshold level for a certain period of time is commonly called the *Parisian feature*. Let B denote the trigger price and the “Parisian” clock starts counting (cumulatively or consecutively) when the asset price stays above B . In the discrete trinomial evolution of the asset price, we construct the grid function $g_{\text{cum}}(k, j)$ that models the correlated evolution of the discrete asset price process and the cumulative counting of the number of time steps that $S_j \geq B$. Given that k is the cumulative counting of the number of time steps that the asset price has been staying above B , the index k increases its value by 1 when $S_j \geq B$. Then we have

$$g_{\text{cum}}(k, j) = k + \mathbf{1}_{\{S_j \geq B\}} \quad (19)$$

where $\mathbf{1}_{\{S_j \geq B\}}$ denotes the indicator function associated with the event $\{S_j \geq B\}$. In a similar manner, the grid function $g_{\text{con}}(k, j)$ that models the consecutive counting of the number of time steps that $S_j \geq B$ is defined as

$$g_{\text{con}}(k, j) = (k + 1)\mathbf{1}_{\{S_j \geq B\}} \quad (20)$$

Using the FSG approach, the path dependence of the soft call requirement can be easily incorporated into the pricing algorithm for a convertible bond with call provision [7]. Suppose that the number of cumulative time steps required for activation of the call provision is K ; then the dynamic programming procedure that enforces the interaction of the game option of holder’s optimal conversion and issuer’s optimal call is applied at a given lattice grid only when the condition $g_{\text{cum}}(k, j) \geq K$ is satisfied.

Call Options with Strike Reset Feature

Consider a call option with strike reset feature where the option’s strike price is reset to the prevailing asset price on a preset reset date if the option is out of the money on that date. Let $t_i, i = 1, 2, \dots, M$, denote

the reset dates and X_i denote the strike price specified on t_i based on the above reset rule. Write X_0 as the strike price set at initiation; then, X_i is given as

$$X_i = \min(X_0, X_{i-1}, S_{t_i}) \quad (21)$$

where S_{t_i} is the prevailing asset price at reset date t_i . Note that the strike price at expiry of this call option is not fixed since its value depends on the realization of the asset price at the reset dates. When we apply the backward induction procedure in the trinomial calculations, we encounter the difficulty in defining the terminal payoff since the strike price is not yet known. These difficulties can be resolved easily using the FSG approach by tracking the evolution of the asset price and the reset strike price through an appropriate choice of the grid function [6].

Recall that S_0 is the asset price at the tip of the trinomial tree and the asset price after j net upward jumps is $S_0 u^j$. In our notation, the index k is used as the one-to-one correspondence to the asset price level $S_0 u^k$. Say, suppose that the original strike price X_0 corresponds to the index k_0 , this would mean $X_0 = S_0 u^{k_0}$. For convenience, we may choose the proportional jump parameter u such that k_0 is an integer. In terms of these indexes, the grid function that models the correlated evolution between the reset strike price and asset price is given as (see equation 21)

$$g_{\text{reset}}(k, j) = \min(k, j, k_0) \quad (22)$$

where k denotes the index that corresponds to the strike price reset in the last reset date and j is the index that corresponds to the prevailing asset price at the reset date.

Since the strike price is reset only on a reset date, we perform the usual trinomial calculations for those time levels that do not correspond to a reset date while the augmented state vector of strike prices are adjusted according to the grid function $g_{\text{reset}}(k, j)$ for those time levels that correspond to a reset date. The FGS algorithm for pricing the reset call option is given as

$$V_{j,k}^n = \begin{cases} p_u V_{j+1,k}^{n+1} + p_0 V_{j,k}^{n+1} + p_d V_{j-1,k}^{n+1} & \text{if } (n+1)\Delta t \neq t_i \text{ for some } i \\ p_u V_{j+1, g_{\text{reset}}(k, j+1)}^{n+1} + p_0 V_{j, g_{\text{reset}}(k, j)}^{n+1} & \\ \quad + p_d V_{j-1, g_{\text{reset}}(k, j-1)}^{n+1} & \\ \text{if } (n+1)\Delta t = t_i \text{ for some } i \end{cases} \quad (23)$$

Lastly, the payoff values along the terminal nodes at the N th time level in the trinomial tree are given as

$$V_{j,k}^N = \max(S_0 u^j - S_0 u^k, 0), \quad j = -N, -N+1, \dots, N \quad (24)$$

and k assumes values that are taken by j and k_0 .

References

- [1] Barraquand, J. & Pudet, T. (1996). Pricing of American path-dependent contingent claims, *Mathematical Finance* **6**, 17–51.
- [2] Chu, C.C. & Kwok, Y.K. (2007). Target redemption note, *Journal of Futures Markets* **27**, 535–554.
- [3] Forsyth, P., Vetzal, K.R. & Zvan, R. (2002). Convergence of numerical methods for valuing path-dependent options using interpolation, *Review of Derivatives Research* **5**, 273–314.
- [4] Hull, J. & White, A. (1993). Efficient procedures for valuing European and American path dependent options, *Journal of Derivatives* **1**(Fall), 21–31.
- [5] Jiang, L. & Dai, M. (2004). Convergence of binomial tree method for European/American path-dependent options, *SIAM Journal of Numerical Analysis* **42**(3), 1094–1109.
- [6] Kwok, Y.K. & Lau, K.W. (2001). Pricing algorithms for options with exotic path dependence, *Journal of Derivatives* **9**, 28–38.
- [7] Lau, K.W. & Kwok, Y.K. (2004). Anatomy of option features in convertible bonds, *Journal of Futures Markets* **24**(6), 513–532.
- [8] Leung, K.S. & Kwok, Y.K. (2008). Employee stock option valuation with repricing features, *Quantitative Finance*, to appear.
- [9] Ritchken, P. & Chuang, I. (2000). Interest rate option pricing with volatility humps, *Review of Derivatives Research* **3**, 237–262.
- [10] Ritchken, P.L., Sankarasubramanian, L. & Vijn, A.M. (1993). The valuation of path dependent contract on the average, *Management Science* **39**, 1202–1213.
- [11] Ritchken, P. & Trevor, R. (1999). Pricing option under generalized GARCH and stochastic volatility processes, *Journal of Finance* **54**(1), 377–402.

Related Articles

Asian Options; Binomial Tree; Convertible Bonds; Lookback Options; Quantization Methods; Tree Methods.

YUE-KUEN KWOK

Wavelet Galerkin Method

Wavelet methods in finance are a particular realization of the finite element method (see **Finite Element Methods**) that provides a very general PDE-based numerical pricing technique. The methods owe their name to the choice of a wavelet basis in the finite element method. This particular choice of basis allows the method to solve partial integro-differential equations (PIDEs) arising from a very large class of market models. Therefore, wavelet-based finite element methods are well suited for the analysis of model risk and pricing in multidimensional and exotic market models. Since wavelet methods are mesh-based methods they allow for the efficient calculation of Greeks and other model sensitivities.

As for any finite element method, the general setup for wavelet methods can be described as follows. Consider a basket of $d \geq 1$ assets whose log returns X_t are modeled by a Lévy or, more generally, a Feller process with state space $\mathbb{R}^d$ and $X_0 = x$. By the fundamental theorem of asset pricing, the arbitrage-free price u of a European contingent claim with payoff $g(\cdot)$ on these assets is given by the conditional expectation

$$u(x, t) = \mathbb{E} [g(X_t) \mid X_0 = x] \quad (1)$$

under an *a priori* chosen equivalent martingale measure (see **Exponential Lévy Models**). Provided sufficient smoothness of u , the price can be obtained as the solution of a PIDE (see **Partial Differential Equations; Partial Integro-differential Equations (PIDEs)**):

$$\begin{aligned} \frac{\partial u}{\partial t} + \mathcal{A}u &= 0 \\ u(\cdot, 0) &= g \end{aligned} \quad (2)$$

where $\mathcal{A}$ denotes the infinitesimal generator of the process X . For the Galerkin-based finite element implementation, equation (2) is converted into variational form on a suitable test space V (cf., e.g., [11, 20] and the references therein). The variational pricing equation then reads: find $u \in V$ such that

$$\left\langle \frac{d}{dt}u, v \right\rangle + \mathcal{E}(u, v) = 0 \quad \text{for all } v \in V \quad (3)$$

where $\mathcal{E}(u, v) := \langle \mathcal{A}u, v \rangle_{L^2 \times L^2}$ denotes the bilinear form associated to the process X . Note that this bilinear form is the central object of discretization in wavelet methods. Owing to the variational formulation (3), wavelet methods are also referred to as *variational methods*.

In one dimension, wavelet methods have been introduced by Matache *et al.* [20, 21]. They were successively applied to American-type contracts (cf. [19, 27]) as well as stochastic volatility models (cf. [14]). In [11], this approach was extended to multidimensional market models based on the sparse tensor product approach and wavelet compression techniques as described in [23, 26] and the references therein.

Admissible Pricing Models

Wavelet-based finite element methods are applicable whenever the variational equation (3) admits a unique solution. Some admissible market models are the multidimensional Black–Scholes model, local volatility models, Kou’s model (see **Kou Model**), stochastic volatility models (see **Barndorff-Nielsen and Shephard (BNS) Models; Heston Model; Hull–White Stochastic Volatility Model**), one-dimensional Lévy models (see **Variance-gamma Model; Jump-diffusion Models; Time-changed Lévy Process; Exponential Lévy Models**), multidimensional Lévy copula models (see [11]), as well as models based on time inhomogeneous or nonstationary processes (see [6]). The particular choice of market model determines the particular form of the bilinear form $\mathcal{E}(\cdot, \cdot)$ in equation (3). In all the above market models, the bilinear form is governed by the characteristic triplet $(\gamma, \mathcal{Q}, \nu)$ of the underlying Lévy process. It consists of a drift vector $\gamma \in \mathbb{R}^d$, a covariance matrix $\mathcal{Q} \in \mathbb{R}^{d \times d}$, and a Lévy-type measure ν that is assumed to be absolutely continuous with density $k(z)dz = \nu(dz)$. Then, $\mathcal{E}(\cdot, \cdot)$ is of the abstract form

$$\begin{aligned} \mathcal{E}(u, v) &= \langle \gamma \nabla u, v \rangle + \frac{1}{2} \langle \mathcal{Q} \nabla u, \nabla v \rangle \\ &\quad - \int \int_{\mathbb{R}^d \times \mathbb{R}^d} (u(x + f(z)) - u(x) \\ &\quad - f(z) \cdot \nabla u(x)) v(x) \nu(dz) dx \end{aligned} \quad (4)$$

Note that in case the underlying process X is not stationary or time inhomogeneous, the parameters $(\gamma, \mathcal{Q}, \nu) = (\gamma(x, t), \mathcal{Q}(x, t), \nu(x, t))$ as well as the function $f(\cdot)$ may depend on space and time (e.g., in models based on Sato processes [6]). Furthermore, the parameters are allowed to degenerate (e.g., in stochastic volatility models). Under rather weak conditions, wavelet-based finite elements methods are still applicable in these cases.

Wavelet-based Finite Element Discretization

The wavelet-based finite element implementation of equation (3) is obtained in three steps: first, the original space domain $\mathbb{R}^d$ has to be *localized* to a bounded domain $\square := [-R, R]^d$, $R > 0$. Second, the test space V needs to be *discretized* by an increasing sequence of finite dimensional subspaces $V^L \subset V$, $L \in \mathbb{N}$. Third, a *time-stepping* scheme has to be applied to discretize in time.

Localization

For the localization, we find that in finance truncation of the original x -domain $\mathbb{R}^d$ to $\square$ corresponds to approximating the solution u of equation (2) by the price u_R of a corresponding barrier option on $[e^{-R}, e^R]^d$. In case the underlying stochastic process X admits semiheavy tails, the solution of the localized problem u_R converges pointwise exponentially to the solution u of equation (3). There exist constants $\alpha, \beta > 0$ such that

$$|u(t, x) - u_R(t, x)| \lesssim e^{-\alpha R + \beta \|x\|_\infty} \quad (5)$$

It therefore indeed suffices to replace the original price space domain $\mathbb{R}^d$ by $\square = [-R, R]^d$ with sufficiently large $R > 0$. For details we refer to [24].

Space Discretization

In wavelet methods, the space discretization is based on the concepts of classical finite element methods (see **Finite Element Methods**). To this end, for any level index $L \in \mathbb{N}$, let $V^L \subset V$ be a subspace of dimension $N := \dim V^L = \mathcal{O}(h^{-d})$ generated by a tensor product finite element basis $\Phi_L := \{\phi_{j,L} : j \in \Delta_L\}$ with some suitable index set Δ_L corresponding to a mesh of width $h = 2^{-L}$. There holds $V^L \subset V^{L+1}$

for all $L \in \mathbb{N}$. For further details on classical finite element approximations see, for example [10].

Denote by $\underline{U}_L$ the coefficient vector of u with respect to Φ_L . Then equation (3) is equivalent to: find $\underline{U}_L(t) \in \mathbb{R}^N$ such that $\underline{U}_L(0) = \underline{U}_0$ and

$$\mathbf{M}\underline{U}'_L(t) + \mathbf{A}\underline{U}_L(t) = 0, \quad t \in (0, T) \quad (6)$$

where $\mathbf{M}$ and $\mathbf{A}$ are the so-called mass and stiffness matrices.

Straightforward application of standard finite element schemes to calculate the stiffness matrix $\mathbf{A} = (\mathcal{E}(\phi_{i,L}, \phi_{j,L}))_{i,j \in \Delta_L}$ arising from general market models fails due to two reasons. For high-dimensional models, we have the “curse of dimension”: the number of degrees of freedom on a tensor product finite element mesh of width h in dimension d grows like $\mathcal{O}(h^{-d})$ as $h \rightarrow 0$. For jump models, the nonlocality of the underlying operator implies that the standard finite element stiffness matrix $\mathbf{A}$ consists of $\mathcal{O}(h^{-2d})$ nonzero entries, which is not practicable even in one dimension with small mesh widths.

Wavelets can overcome these issues while still being easy to compute. There are three main advantages.

- Break the curse of dimension using sparse tensor products (see **Sparse Grids**) $\Rightarrow$ dimension-independent complexity (up to log-factors).
- Multiscale compression of jump measure of $X \Rightarrow$ complexity of jump models can asymptotically be reduced to Black–Scholes complexity.
- Efficient preconditioning.

In one dimension, biorthogonal *complement* or *wavelet bases* $\Psi_L = \{\psi_{j,L} : j \in \nabla_L\}$, $\nabla_L := \Delta_{L+1} \setminus \Delta_L$ are constructed from the single-scale bases Φ_L ; for details see [7, 9, 22]. Denoting by W^L the span of Ψ_L , the spaces V^{L+1} admit a splitting

$$V^{L+1} = W^L \oplus V^L, \quad L > 0 \quad (7)$$

Each wavelet space W^L can be thought of as describing the increment of information when refining the finite element approximation from V^L to V^{L+1} . Furthermore, equation (7) implies that for any $L > 0$ the finite element space V^L can be written as a direct *multilevel* sum of the wavelet spaces W^ℓ , $\ell < L$. Thus, any $u_L \in V^L$ has the representation

$$u_L = \sum_{\ell=0}^{L-1} \sum_{j \in \nabla_\ell} d_{j,\ell} \psi_{j,\ell} \quad (8)$$

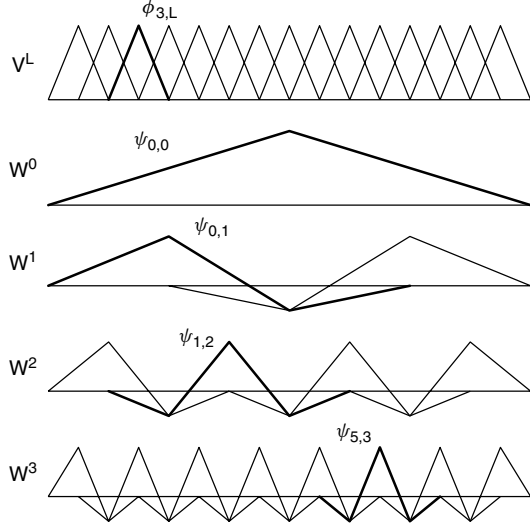


Figure 1 Schematic of single-scale space V^L and its decomposition into multiscale wavelet spaces W^ℓ

with suitable coefficients $d_{j,\ell} \in \mathbb{R}$. Figure 1 illustrates the decomposition of the finite element space V^L , $L = 4$, spanned by continuous, piecewise linear (nodal) basis functions $\phi_{i,L}$ into its increment spaces W^ℓ , $\ell = 0, \dots, 3$, spanned by wavelets $\psi_{j,\ell}$. In the multidimensional setting, we obtain multivariate wavelet basis functions by using tensor products. The finite element spaces V^L can then be characterized by

$$V^L = \text{span} \{ \psi_{j_1, \ell_1}(x_1) \cdots \psi_{j_d, \ell_d}(x_d) : \ell_1, \dots, \ell_d \leq L, j_i \in \nabla_{\ell_i} \} \quad (9)$$

Since these multivariate wavelet bases comprise of products of one-dimensional wavelets, they form *hierarchical bases* as in [12]. Thus, the spaces V^L can be replaced by *sparse tensor product* spaces

$$\widehat{V}^L = \text{span} \{ \psi_{j_1, \ell_1}(x_1) \cdots \psi_{j_d, \ell_d}(x_d) : \ell_1 + \dots + \ell_d \leq L, j_i \in \nabla_{\ell_i} \} \quad (10)$$

In [3, 26] it is shown that, under certain smoothness assumptions on the solution u of equation (3), the sparse tensor product spaces preserve the approximation properties of the full tensor product spaces, while there holds $\widehat{N} := \dim \widehat{V}^L = \mathcal{O}(h^{-1} |\log h|^{d-1}) \ll N$. Herewith, the complexity of the finite element

stiffness matrix can be reduced to $\mathcal{O}(h^{-2} |\log h|^{2(d-1)})$ instead of the original $\mathcal{O}(h^{-2d})$ nonzero entries.

Furthermore, wavelet basis functions give rise to certain *cancellation properties* and *norm equivalences* as illustrated in, for example [2, 7]. One therefore obtains sharp estimates for the entries of the corresponding stiffness matrix, cf. [20, 23]. Herewith *a priori* compression schemes can be defined that further reduce the complexity of the stiffness matrix. The compression exploits the fact that the position of large entries in the stiffness matrix arising from a model with jumps resembles the structure of a Black–Scholes stiffness matrix. The remaining entries can *a priori* be proved to be negligible. Therefore, the compression scheme (asymptotically) reduces the complexity of a model with jumps to that of the Black–Scholes model. To give a brief illustration of this idea, in Figure 2, the density pattern of the stiffness matrix of the Black–Scholes model (a) and a Lévy model of tempered-stable-type (b) in one dimension is given.

Combining the compression scheme with the sparse tensor product spaces results in a computational complexity of $\mathcal{O}(h^{-1} |\log h|^{2(d-1)})$ instead of the original $\mathcal{O}(h^{-2d})$. It is proved that these wavelet schemes preserve stability and convergence of the classical finite element schemes, cf. [23].

Time Stepping

To finally obtain a fully discrete solution of equation (3), there are various methods to approximate the solution of the ordinary differential equation (6). These time-stepping methods are not specific to wavelet-based finite element methods. We therefore mention only a few examples here: one obtains algebraic convergence $\mathcal{O}((\Delta t)^\alpha)$ in the time step size Δt for all the above market models using the implicit Euler scheme, then $\alpha = 1$ or the Crank–Nicholson scheme, then $\alpha = 2$; see, for example [1]. Furthermore, exponential convergence rates can be obtained by discretizing equation (6) by a *hp*-discontinuous Galerkin (dG) method as in [14, 25].

American Contracts

The price u of an American-type contract is given by an optimal stopping, free boundary problem:

$$u(t, x) = \sup_{t \leq \tau \leq T} \mathbb{E}[g(X_\tau) \mid X_t = x] \quad (11)$$

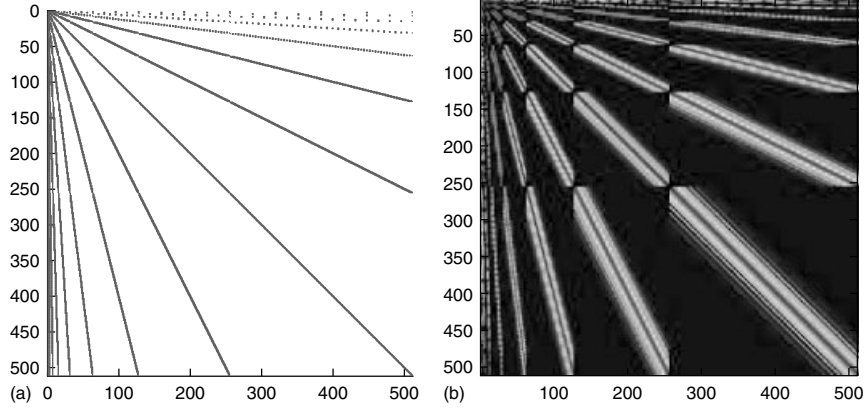


Figure 2 Density pattern for the stiffness matrix $\mathbf{A}$ for $L = 8$ refinement steps. Matrix for Black–Scholes model (a) and for tempered stable process (b)

with τ running over stopping times. The solution of equation (11) then solves a partial integro-differential *inequality* (see [19]). Similar to the variational formulation for European contracts, the wavelet-based finite element implementation solves the corresponding variational inequality: Find $u(t, \cdot) \in K_g := \{v \in V \mid v \geq g \text{ a.e.}\}$ such that

$$\left\langle \frac{\partial u}{\partial t}, v - u \right\rangle + \mathcal{E}(u, v - u) \leq 0 \quad \text{for all } v \in K_g \quad (12)$$

Choosing finite dimensional subspaces $V^L \subset V$, as above, to discretize in space and applying a time-stepping scheme leads to a sequence of matrix linear complementary problems. For large size $N = \dim V^L$, standard solution methods such as projected successive overrelaxation [8] are not suitable, since their rate of convergence depends on N . The wavelet-based solution algorithm suggested in [19] relies on a fixed point iteration where in each iteration step a V -projection P_{K_g} onto the convex cone K_g has to be realized. Owing to norm equivalences of the wavelet basis, the outer fixed point iteration converges at a rate independent of the dimension N of the space V^L . The projection P_{K_g} is based on a wavelet generalization of the classical Cryer algorithm [8].

Sensitivities and Greeks

As a PDE-based pricing method, wavelet methods are well suited for the fast and accurate calculation

of sensitivities of market models with respect to model parameters. Classical examples are variations of option prices with respect to the spot price or with respect to time to maturity, the so-called “Greeks” of the model.

There are two classes of sensitivities: the sensitivity of the solution u to variation of a model parameter, like the Greek Vega ($\partial_\sigma u$), and the sensitivity of the solution u to a variation of the state space such as the Greek Delta ($\partial_x u$).

Suppose that the infinitesimal generator $\mathcal{A}$ of X depends on a parameter η . We write $u(\eta_0)$ for a fixed parameter η_0 to emphasize the dependence of u on η_0 . Then, the derivative of u with respect to η

$$\tilde{u}(\delta\eta) := \lim_{s \rightarrow 0^+} \frac{1}{s} (u(\eta_0 + s\delta\eta) - u(\eta_0)) \quad (13)$$

is the solution of the PIDE (see [15])

$$\begin{aligned} \partial_t \tilde{u}(\delta\eta) + \mathcal{A}(\eta_0) \tilde{u}(\delta\eta) &= -D_\eta \mathcal{A} u(\eta_0) \\ \tilde{u}(\delta\eta)(0, \cdot) &= 0 \quad \text{in } \mathbb{R}^d \end{aligned} \quad (14)$$

where $D_\eta \mathcal{A}$ is the derivative of $\mathcal{A}$ with respect to η . Therefore, the derivative of u with respect to η can be obtained as a solution of the same pricing PIDE with a right-hand side depending on u , cf. [15].

For sensitivities with respect to a variation of the state space, a finite difference like differentiation procedure is presented in [15], which allows to obtain the sensitivities from the finite element forward price without additional forward solver.

Numerical Examples

In this section, we illustrate the use of wavelet-based finite element methods by a number of numerical examples.

Multidimensional Lévy Models

We consider a two-dimensional basket with payoff $g = (\frac{1}{2}(S_1 + S_2) - K)^+$, strike $K = 1$, and maturity $T = 1$. The price process is a pure jump process, which is given by the Clayton copula [16] and tempered stable margins [5]. The parameters are $C = (1, 1)$, $G = (10, 10)$, $M = (10, 10)$, and $Y = (0.5, 1.5)$. The copula depends on a parameter $\theta > 0$ parameterizing the dependence among jumps.

Different values for θ are used to analyze the different dependence structures. In Figure 3, we compare the model with a Black–Scholes model having the same marginal variances and the same correlation.

American-type Options in Lévy Models

We consider the finite horizon multiple stopping time problem as in [4] arising from swing options with maturity $T = 1$ and interest rate $r = 0.05$. The holder has $p = 5$ exercise rights and there is a refracting time $\delta = 0.1$ after each exercise. For computational details, see [27]. The computed prices and the exercise boundary are shown in Figure 4 where we use a tempered stable/CGMY model [5] and computed a swing put option with up

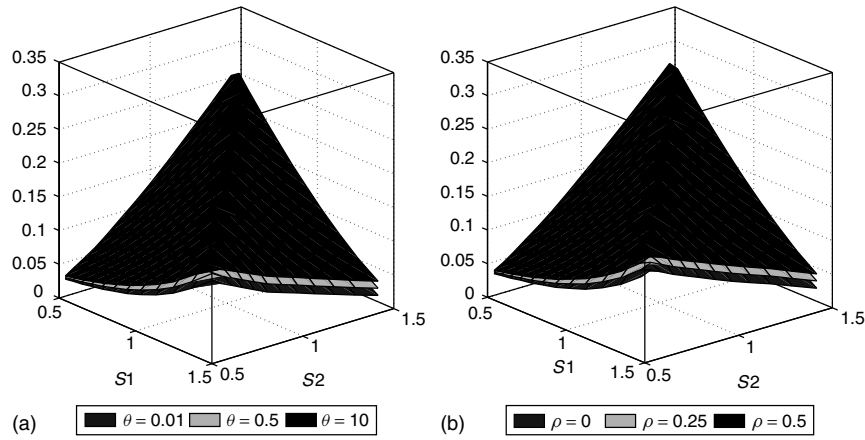


Figure 3 Time value for basket options using a Lévy model (a) and Black–Scholes model (b)

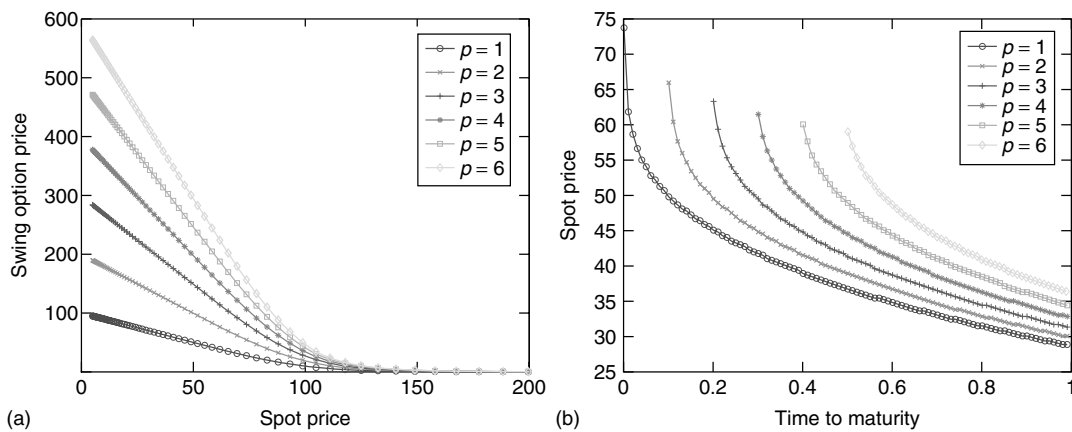


Figure 4 Swing option price (a) and exercise boundary (b) of a put option with up to six exercise rights

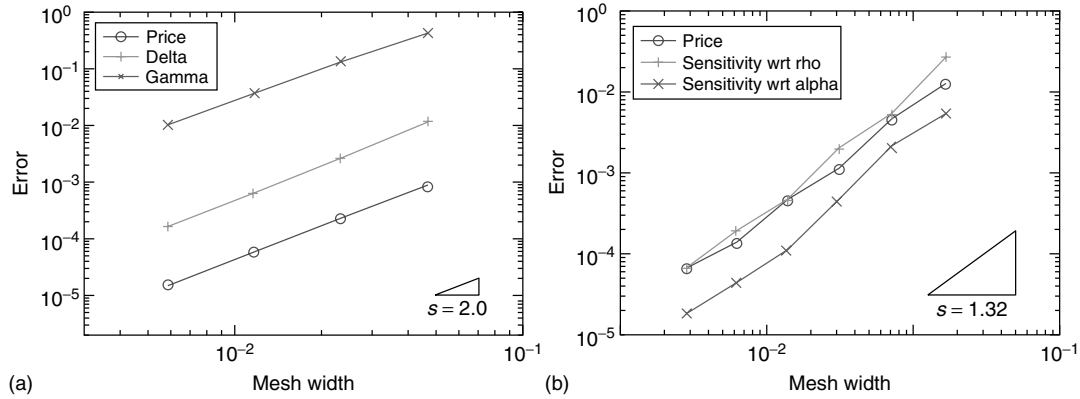


Figure 5 Convergence rates of sensitivities for a European call in the variance Gamma (a) and Heston (b) model

to five exercise rights and strike $K = 100$. The model parameters are $C = 1$, $G = 10$, $M = 10$, and $Y = 0.5$. In contrast to the result in the Black–Scholes model, the exercise boundary values in a Lévy model never reach the option’s strike price, which is well known for American options [17, 19].

Sensitivities

We compute various sensitivities for the variance Gamma [18] and the Heston model [13] where the price is known in closed form such that we are able to compute the errors between the exact price/sensitivities and their finite element approximations.

We consider a European call with strike $K = 1$ and maturity $T = 0.5$. For the variance Gamma model, we calculate the Greeks Delta, $\Delta = \partial u / \partial S$, and Gamma, $\Gamma = \partial^2 u / \partial S^2$, and use the parameters $\sigma = 0.4$, $\nu = 0.04$, and $\vartheta = -0.2$. For the Heston model, we calculate the sensitivities $\tilde{u}(\delta\rho)$ and $\tilde{u}(\delta\alpha)$ with respect to correlation ρ of the Brownian motions that drive the underlying and the volatility and the rate of mean reversion α . The model parameters are $\lambda = 0$, $\sigma = 0.5$, $m = 0.06$, $\rho = -0.5$, and $\alpha = 2.5$.

The convergence rates are shown in Figure 5. All sensitivities convergence with the same rate as the price u itself [15].

References

- [1] Achdou, Y. & Pironneau, O. (2005). Computational methods for option pricing, *Frontiers in Applied Mathematics*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Vol. 30.
- [2] Bramble, J.H., Cohen, A. & Dahmen, W. (2003). Multiscale problems and methods in numerical simulations, *Lecture Notes in Mathematics*, Springer-Verlag, Berlin, Vol. 1825.
- [3] Bungartz, H.-J. & Griebel, M. (1999). A note on the complexity of solving Poisson’s equation for spaces of bounded mixed derivative, *Journal of Complexity* **15**, 167–199.
- [4] Carmona, R. & Touzi, N. (2006). Optimal multiple stopping and valuation of swing options, *Mathematical Finance*, **18**(2), 239–268.
- [5] Carr, P., Geman, H., Madan, D.B. & Yor, M. (2002). The fine structure of assets returns: an empirical investigation, *Journal of Business* **75**(2), 305–332.
- [6] Carr, P., Geman, H., Madan, D.B. & Yor, M. (2007). Self-decomposability and option pricing, *Mathematical Finance* **17**(1), 31–57.
- [7] Cohen, A. (2003). *Numerical Analysis of Wavelet Methods*, Elsevier, Amsterdam.
- [8] Cryer, C.W. (1971). The solution of a quadratic programming problem using systematic overrelaxation, *SIAM Journal of Control* **9**(3), 385–392.
- [9] Dahmen, W., Kunoth, A. & Urban, K. (1999). Biorthogonal spline wavelets on the interval—stability and moment conditions, *Applied and Computational Harmonic Analysis* **6**, 259–302.
- [10] Ern, A. & Guermond, J.-L. (2004). *Theory and Practice of Finite Elements*, Springer Verlag, New York.
- [11] Farkas, W., Reich, N. & Schwab, C. (2007). Anisotropic stable Lévy copula processes—analytical and numerical aspects, *Mathematical Models and Methods in Applied Sciences* **17**, 1405–1443.
- [12] Griebel, M. & Oswald, P. (1995). Tensor product type subspace splittings and multilevel iterative methods for anisotropic problems, *Advances in Computational Mathematics* **4**, 171–206.
- [13] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility, with applications to bond and

- currency options, *The Review of Financial Studies* **6**, 327–343.
- [14] Hilber, N., Matache, A.-M. & Schwab, C. (2005). Sparse wavelet methods for option pricing under stochastic volatility, *The Journal of Computational Finance* **8**(4), 1–42.
- [15] Hilber, N., Schwab, C. & Winter, C. (2008). Variational Sensitivity Analysis of Parametric Markovian Market Models, in *Advances in Mathematics in Finance*, L. Stettner, ed, Banach Center Publications Vol. **83**, 85–106.
- [16] Kallsen, J. & Tankov, P. (2006). Characterization of dependence of multidimensional Lévy processes using Lévy copulas, *Journal of Multivariate Analysis* **97**, 1551–1572.
- [17] Levendorskiĭ, S.Z. (2004). Early exercise boundary and option prices in Lévy driven models, *Quantitative Finance* **4**(5), 525–547.
- [18] Madan, D.B., Carr, P. & Chang, E. (1998). The variance gamma process and option pricing, *European Finance Review* **2**, 79–105.
- [19] Matache, A.-M., Nitsche, P.-A. & Schwab, C. (2005). Wavelet Galerkin pricing of American options on Lévy driven assets, *Quantitative Finance* **5**(4), 403–424.
- [20] Matache, A.-M., Petersdorff, T. & Schwab, C. (2004). Fast deterministic pricing of options on Lévy driven assets, *M2AN Mathematical Modelling and Numerical Analysis* **38**(1), 37–71.
- [21] Matache, A.-M., Schwab, C. & Wihler, T.P. (2006). Linear complexity solution of parabolic integro-differential equations, *Numerical Mathematics* **104**(1), 69–102.
- [22] Nguyen, H. & Stevenson, R. (2003). Finite elements on manifolds, *IMA Journal of Numerical Analysis* **23**, 149–173.
- [23] Reich, N. (2008). *Wavelet Compression of Anisotropic Integrodifferential Operators on Sparse Tensor Product Spaces*, PhD Thesis, ETH, Zurich.
- [24] Reich, N. Schwab, C. & Winter, C. (2008). *Anisotropic multivariate Lévy processes and their Kolmogorov equation*, research report no. 2008-03, *Seminar for Applied Mathematics*, ETH, Zurich.
- [25] Schötzau, D. & Schwab, C. (2001). *hp*-discontinuous Galerkin time-stepping for parabolic problems, *Comptes Rendus de l'Académie des Sciences Series I Mathematics*, **333**(12).
- [26] von Petersdorff, T. & Schwab, C. (2004). Numerical solution of parabolic equations in high dimensions, *M2AN Mathematical Modelling and Numerical Analysis* **38**(1), 93–127.
- [27] Wilhelm, M. & Winter, C. (2008). Finite element valuation of swing options, *Journal of Computational Finance* **11**(3), 107–132.

NORBERT HILBER, NILS REICH & CHRISTOPH
WINTER

Integral Equation Methods for Free Boundaries

Free boundary problems (FBPs) are ubiquitous in modern mathematical finance. They arise as early exercise boundaries for American style options (arguably the first example in finance, introduced by McKean [15] in 1965), as default barriers in structural (value-of-firm) models of credit default [7], as the optimal strategies for refinancing mortgages [13, 21], exercising employee stock options [14] (*see Employee Stock Options*) and callable convertible bonds [20], and so on. There are many methods for treating the FBPs that arise as mathematical models of these finance problems, including variational inequalities [11], viscosity solutions [8], and the classical partial differential equations (PDE) approach [10, 16]. In this article, we focus on an integral equation (IE) approach that is particularly suited for the types of FBPs that arise in finance.

In the section Free Boundary Problems as Integral Equations, we sketch the method in the context of the American put option, perhaps the most widely known and best understood FBP in finance. After deriving an IE problem mathematically equivalent to the original Black, Scholes, and Merton (BSM) PDE FBP for the American put, we outline how the IE problem can be used to prove existence and uniqueness for the original problem and to derive analytical and numerical estimates for the location of the early exercise boundary [4, 5]. In the section Application of the IE Method to Other FBPs, we sketch how this IE approach can be carried over to other FBPs in finance [7, 21] with the goal of indicating that this is a unified approach to a diverse collection of problems.

Free Boundary Problems as Integral Equations

In this section, we outline the IE approach in the context of an American put option on a geometric Brownian motion underlier (*see American Options*). The BSM risk-neutral pricing theory says that the

option value, $p(S, t)$, satisfies the FBP

$$p_t + \frac{\sigma^2 S^2}{2} p_{SS} + rSp_S - rp = 0, \quad S_f(t) < S < \infty, 0 < t < T \quad (1)$$

$$p(S, t) = K - S \text{ on } S = S_f(t), \quad 0 < t < T \quad (2)$$

$$p_S(S, t) = -1 \text{ on } S = S_f(t), \quad 0 < t < T \quad (3)$$

$$p(S, t) \rightarrow 0 \text{ as } S \rightarrow \infty \quad (4)$$

$$p(S, T) = \max(K - S, 0), \quad K = S_f(T) < S < \infty \quad (5)$$

where r, σ, K, T have the conventional meanings and $S_f(t)$ is the location of the early exercise, free boundary to be determined along with $p(S, t)$.

Letting $\tau = \frac{\sigma^2}{2}(T - t)$ (the scaled time to expiry) and $x = \ln(S/K)$, then the scaled option price

$$P_{\text{new}} = \begin{cases} 1 - S/K & S < S_f(t) \\ p/K & S > S_f(t) \end{cases} \quad (6)$$

satisfies the transformed problem (dropping the subscript) in $-\infty < x < \infty, 0 < \tau < \sigma^2 T/2$

$$p_\tau - \{p_{xx} + (k - 1)p_x - kp\} = kH(x_f(\tau) - x) \quad (7)$$

$$p(x, 0) = \max(1 - e^x, 0) \quad (8)$$

where $k = 2r/\sigma^2$, H is the Heaviside function, $x_f(\tau) = \ln(S_f/S)$, and the coefficient k appears on the right-hand side (rhs) of equation (7) because the intrinsic payoff, $p_0(x) = 1 - e^x$ satisfies

$$p_{0\tau} - \{p_{0xx} + (k - 1)p_{0x} - kp_0\} = k \quad (9)$$

The solution to problems (7) and (8) can be written in terms of the free boundary, $x_f(\tau)$, using the fundamental solution of the pdo on the left-hand side (lhs) of equation (7),

$$\Gamma(x, \tau) = \frac{e^{-k\tau}}{2\sqrt{\pi\tau}} e^{-(x+(k-1)\tau)^2/4\tau} \quad (10)$$

in the form

$$p(x, \tau) = \int_{-\infty}^0 (1 - e^y) \Gamma(x - y, \tau) dy + k \int_0^\tau \int_{-\infty}^{x_f(u)} \Gamma(x - y, \tau - u) dy du \quad (11)$$

The first term is the price of the European style put while the second is the premium for the American optionality. Integral representations of this sort have been discussed in the finance literature for some time (see, e.g., [3]).

For the representation (11) to be useful, one must first determine the unknown location of the boundary, $x_f(\tau)$, which appears in the second integral on the rhs of equation (11). The usual approach in the free boundary literature proceeds by starting with equation (11) and evaluating the lhs at $x = x_f(\tau)$ using one or other of the conditions

$$p(x_f(\tau), \tau) = 1 - e^{x_f(\tau)} \quad (12)$$

$$p_x(x_f(\tau), \tau) = -e^{x_f(\tau)} \quad (13)$$

which are the transformed versions of the smooth pasting conditions (2) and (3). Instead, we use a trick here, based on financial considerations, to notice that $p_\tau(x_f(\tau), \tau) = 0$. Thus, from equation (11),

$$p_\tau(x, \tau) = \Gamma(x, \tau) + k \int_0^\tau \Gamma(x - x_f(u), \tau - u) \times \dot{x}_f(u) du \quad (14)$$

which, upon evaluation on the early exercise boundary, provides the following nonlinear IE for $x_f(\tau)$:

$$\Gamma(x_f(\tau), \tau) = -k \int_0^\tau \Gamma(x_f(\tau) - x_f(u), \tau - u) \times \dot{x}_f(u) du \quad (15)$$

This equation has proved, in our experience, to be much more effective in the mathematical analysis of this problem than the versions obtainable from equations (12) and (13). In addition, we have obtained still other versions, also derivable from the representation

(11), whose particular forms have proven useful in various situations [5]:

$$\int_0^\tau \{\Gamma_x(x_f(\tau), u) + k\Gamma(x_f(\tau), u)\} du = k \int_0^\tau \Gamma(x_f(\tau) - x_f(u), \tau - u) du \quad (16)$$

$$\Gamma(x_f(\tau), \tau) = \frac{k}{2} + k \int_0^\tau \{\Gamma(x_f(\tau) - x_f(u), \tau - u) - \Gamma_x(x_f(\tau) - x_f(u), \tau - u)\} du \quad (17)$$

$$\dot{x}_f(\tau) = \frac{-2\Gamma_x(x_f(\tau), \tau)}{k} + 2 \int_0^\tau \Gamma_x(x_f(\tau) - x_f(u), \tau - u) \times \dot{x}_f(u) du \quad (18)$$

Using the representation (11) to compute p_x , p_{xx} and $p_{x\tau}$, the above IEs for $x_f(\tau)$ follow (after some rearrangement of terms) by evaluation on the boundary.

The underlying theoretical rationale for using this IE approach to treat the original FBPs (1)–(5) or (7), (8), is summarized in the following result.

Theorem 1 (Theorem 3.2 and Sections 4, 5, 6 of [5]). *Suppose that $x_f \in C^1((0, \infty)) \cap C^0([0, \infty))$ and $\alpha(\tau) = x_f(\tau)^2/4\tau$. Assume that as $\tau \searrow 0$, $\alpha(\tau) = [-1 + o(1)]\ln\sqrt{\tau}$, and $\tau\dot{\alpha}(\tau) = o(1)$. Then x_f , together with p defined by equation (11), solves the (equivalent) FBPs (1–5) or (7, 8), if and only if x_f satisfies any of the equivalent integro-differential equations (IODEs) (15–18). Finally, equation (18) has a unique solution with the properties listed above.*

The equivalence of the IODEs (15–18) is established in Lemma 3.1 of [6] and the required estimates on α are rigorously derived from equation (15). The proof that equation (18) has a solution with the required properties is a highly technical analysis [6] based on Schauder's fixed point theorem. Peskir [18] used the IE obtained from equation (11) with equation (12), along with local time—space arguments, to prove uniqueness in the class of continuous boundaries.

Analytical and numerical estimates for the location of the early exercise boundary, that might be useful to practitioners, can also be obtained from the IODEs (15–18). For example, if we make the change of variables $\eta = (x_f(\tau) - x_f(u))/2\sqrt{\tau - u}$ in equation (15), the rhs for small τ (near expiry) behaves like

$$-k \int_0^{\alpha(\tau)} \left[1 - \frac{x_f(\tau) - x_f(u)}{2\dot{x}_f(u)(\tau - u)} \right]^{-1} \frac{e^{-\eta^2}}{\sqrt{\pi}} d\eta \quad (19)$$

which tends to k because $\alpha(\tau) \rightarrow -\infty$ (from the above theorem) and $[\cdot \cdot \cdot]^{-1} \rightarrow 1/2$ uniformly in u because of the convexity of x_f (proved separately in [6] using the method of Friedman and Jensen [12]); an independent proof of convexity was obtained by Ekstrom [9]. Thus, from equation (15) with small τ ,

$$\begin{aligned} \Gamma(x_f(\tau), \tau) &= \frac{e^{-k\tau}}{2\sqrt{\pi\tau}} e^{-(x_f(\tau) + (k-1)\tau)^2/4\tau} \cong \\ &= \frac{e^{-x_f(\tau)^2/4\tau}}{2\sqrt{\pi}} \cong k \end{aligned} \quad (20)$$

which leads to

$$x_f(\tau) \approx 2\sqrt{\tau} \sqrt{-\ln(4\pi k^2 \tau)^{1/2}} \text{ as } \tau \rightarrow 0 \quad (21)$$

This implies the first rigorous estimate for the near-expiry behavior of the early exercise boundary obtained by Barles *et al.* [2]

$$S_f(t) \simeq k \left[1 - \sigma \sqrt{-(T-t)\ln(T-t)} \right], t \sim T \quad (22)$$

In addition, it provides the first estimate for $\alpha(\tau)$ in the above existence theorem. Specifically,

$$\alpha(\tau) = x_f(\tau)^2 4\tau \approx -\ln(4\pi k^2 \tau)^{1/2} = -\frac{\xi}{2} \quad (23)$$

where $\xi = \ln(4\pi k^2 \tau)$.

More precise analytical and numerical estimates can be obtained for $\alpha(\tau)$ (equivalently $x_f(\tau)$ and $S_f(t)$), valid for intermediate and large times as well [5], by using Mathematica to iterate equation (23) through equation (15). Perhaps even more importantly, a very fast, accurate numerical scheme can be obtained from the IODE (18), which can be written in the equivalent form

$$\dot{x}_f(\tau) = \frac{x_f(\tau)}{2k\tau} \Gamma(x_f(\tau), \tau) [1 + m(\tau)] \quad (24)$$

where

$$\begin{aligned} m(\tau) &= k \int_0^\tau \left[\frac{2\tau}{x_f(\tau)} \left(\frac{x_f(\tau) - x_f(u)}{\tau - u} \right) - 1 \right] \\ &\quad \times \frac{\Gamma(x_f(\tau) - x_f(u), \tau - u)}{\Gamma(x_f(\tau), \tau)} \dot{x}_f(u) du \end{aligned} \quad (25)$$

that is to be solved with initial data $x_f(0) = 0$. Solving this iteratively with $m(\tau) = m_0(\tau) \equiv 0$ initially provides the fastest and most accurate approximation among all our estimates [5].

The above IE formalism can be carried over to the American put with a continuous dividend rate, d . The analog of equations (24) and (25) was used to show numerically that for a small interval of dividend rates, $r < d < (1 + \epsilon)r$, the early exercise boundary loses convexity near maturity (see Figure 1). This agrees qualitatively with the folklore on this problem generally attributed to M. Broadie.

Application of the IE Method to Other FBPs

In the previous section, we described how to formulate the American put FBP in terms of IODEs for the early exercise boundary and how to use this

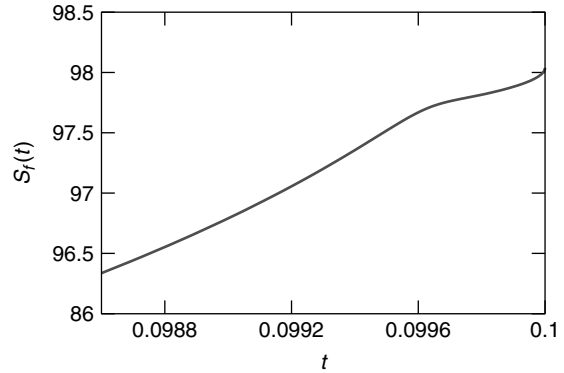


Figure 1 Graph of the early exercise boundary, $S_f(t)$, near expiry for a dividend paying stock with $r = 0.05$, $d = 0.51$, $K = 100$, $T = 0.1$. A numerical scheme based on the analog of equation (26) was used with $x_f(0) = \ln(r/d)$ and the first time step computed from the well-known $\sqrt{T-t}$ behavior near expiry

formulation to establish theoretical results (existence, uniqueness) as well as analytical and numerical estimates for the original problem. In this section, we indicate the wider applicability of the method by briefly discussing several other problems arising from finance.

Jump-diffusion Processes

These IE methods can be extended to jump-diffusion models (see **Lévy Processes**; **Poisson Process**). Specifically, letting $x = \ln(S/K)$, we now assume that the transformed asset price follows the process

$$X(t) = (\mu - \sigma^2/2)t + \sigma W(t) + N(t) \quad (26)$$

where $N(t)$ is a Poisson process with rate λt and has jumps of size $\pm\epsilon$ with equal probability. In this case, the transformed PDEs analogous to equations (7) and (8) are

$$\mathcal{L}p = \mathcal{L}(1 - e^x)H(x_f(\tau) - x) \quad (27)$$

$$p(x, 0) = \max(1 - e^x, 0) \quad (28)$$

where $\mathcal{L}$ is the nonlocal pdo

$$\begin{aligned} \mathcal{L}p = & p_\tau - \{p_{xx} + (k - 1)p_x - kp\} \\ & + \lambda\{p(x + \epsilon, \tau) - 2p + p(x - \epsilon, \tau)\} \end{aligned} \quad (29)$$

This problem is amenable by the methods outlined above because the fundamental solution can be explicitly calculated. Specifically,

$$\begin{aligned} \Gamma(x, \tau) = & \sum_{u=0}^{\infty} \frac{(\lambda\tau)^u}{2^u u!} e^{-\lambda\tau} \\ & \times \left(\sum_{j=0}^n \binom{n}{j} \Gamma_{BS}((2j - u)\epsilon + x, \tau) \right) \end{aligned} \quad (30)$$

where Γ_{BS} is the BSM fundamental solution in equation (10). Proceeding as in the section Free Boundary

Problems as Integral Equations, one obtains the analog to equation (15) in the form

$$\begin{aligned} \Gamma(x_f(\tau), \tau) = & - \int_0^\tau [k + \lambda\{2 - e^\epsilon - e^{-\epsilon}\}e^{x_f(u)} \\ & \times \Gamma(x_f(\tau) - x_f(u), \tau - u) \\ & \times \dot{x}_f(u)] du + \lambda \int_0^\epsilon (1 - e^{y-\epsilon}) \\ & \times \Gamma(x_f(\tau) - y, \tau) dy \end{aligned} \quad (31)$$

from which we obtain the near-expiry estimate for $\alpha(\tau) = x_f(\tau)^2/4\tau$ (see analog (23) with no jumps)

$$\sqrt{\tau}e^\alpha \approx 1/\sqrt{4\pi\tilde{k}^2} \text{ as } \tau \rightarrow 0 \quad (32)$$

where $\tilde{k} = k + \lambda(1 - e^{-\epsilon})$, agreeing with the result of Pham [19] using other methods.

Interest Rate Processes

These IE methods can also be used to study American style contracts on other underliers. For example, a mortgage prepayment option provides the holder with the right to prepay the outstanding balance of a fixed-rate mortgage

$$M(t) = \frac{m}{c}(1 - e^{-c(T-t)}) \quad (33)$$

where T is the maturity, c is the (continuous) fixed mortgage rate, and m is the (continuous) rate of payment of the mortgage (i.e., $m dt$ is the premium paid in any time interval dt). Clearly, the value of the prepayment option depends on $M(t)$ and also on the rate of return, $r(t)$, that the mortgage holder (borrower) can obtain by investing $M(t)$. If this short-term rate is assumed to follow the Vasicek model (see **Term Structure Models**)

$$dr = (\eta - \theta r) dt + \sigma dW \quad (34)$$

in a risk-neutral world, then the value of the prepayment option, $V(r, t)$, satisfies [13, 21]

$$\begin{aligned} V_t + \frac{\sigma^2}{2} V_{rr} + (\eta - \theta r) V_r + m - rV \\ = 0, R(t) < r < \infty, 0 < t < T \end{aligned} \quad (35)$$

$$V(r, t) = M(t), \quad r = R(t), \quad 0 < t < T \quad (36)$$

$$V_r(r, t) = 0, \quad r = R(t), \quad 0 < t < T \quad (37)$$

$$V(r, t) \rightarrow 0, \quad \text{as } r \rightarrow \infty, \quad 0 < t < T \quad (38)$$

$$V(r, T) = 0, \quad c = R(0) < r < \infty \quad (39)$$

The optimal strategy for the mortgage holder is to exercise the option to pay off the mortgage the first time that the rate r falls below $R(t)$ at time t . Existence and uniqueness for this FBP was proved using variational methods [13].

Because the fundamental solution for the Vasicek “bond pricing equation”—equation (35)—can be explicitly calculated, its form suggests a sequence of changes of dependent and independent variables (not relevant for this summary) that reduces the FBPs (35–39) to the following analog of equations (7) and (8) in $-\infty < x < \infty, s > 1$

$$u_s - \frac{1}{4}u_{xx} = f(x, s)H(x - x_f(s)) \quad (40)$$

$$u(x, 1) = 0 \quad (41)$$

where $f(x, s)$ is a specific function resulting from the transformations and $x_f(s)$ is the transformed free boundary with $u(x_f(s), s) = 0 = u_x(x_f(s), s)$.

In this form, the procedure outlined in the previous section can be followed to obtain

$$u(x, s) = \int_1^s \left[\int_{x_f(u)}^{\infty} \Gamma(x - y, s - u) f(y, u) dy \right] du \quad (42)$$

where Γ is the fundamental solution of the heat operator $\partial_s - \frac{1}{4}\partial_{xx}$, and the free boundary can be obtained by solving the IE

$$\int_1^s \left[\int_{x_f(u)}^{\infty} \Gamma(x_f(s) - y, s - u) f(y, u) dy \right] du = 0 \quad (43)$$

In [21], Dejun Xie used the integral representations (42) and (43) to obtain near-expiry estimates for the critical rate as well as to obtain a numerical scheme to determine $x_f(s)$ globally. Specifically, if $Q(x, s)$ denotes the integral on the rhs of equation

(42), he showed that the Newton–Raphson iterative scheme to solve equation (43), $Q(x_f(s), s) = 0$, can be written as

$$x_f(s)^{\text{new}} = x_f(s)^{\text{old}} + \frac{Q(x_f(s)^{\text{old}}, s)}{2f(x_f(s)^{\text{old}}, s)} \quad (44)$$

where, in the denominator, $Q_x(x_f(s), s)$ is approximated by

$$\begin{aligned} & \frac{1}{2} \{ Q_x(x_f(s)+, s) + Q_x(x_f(s)-, s) \} \\ &= \frac{1}{2} u_{xx}(x_f(s), s) = -2f(x_f(s), s) \end{aligned} \quad (45)$$

Default Barrier Models

As a final example, we outline how these methods can be used to obtain an IE formulation for the inverse first-crossing problem in a value-of-firm (structural) model for credit default (*see Structural Default Risk Models*). Suppose the default index of a company, $X(t)$, is a Brownian motion with drift following

$$dX(t) = a dt + \sigma dW(t), \quad X(0) = x_0 \quad (46)$$

(equivalently, the log of such an index that originally satisfied a geometric Brownian motion). Default of the firm is said to occur the first time τ that $X(t)$ falls below a preassigned value, $b(t)$. The survival pdf, $u(x, t)$ defined by

$$u(x, t) dx = Pr[x < x(t) < x + dx | t < \tau]$$

is known to satisfy the following problem for the forward Kolmogorov equation:

$$u_t = \frac{\sigma^2}{2} u_{xx} - a u_x, \quad b(t) < x < \infty, \quad 0 < t < T \quad (47)$$

$$u(x, t) = 0, \quad x = b(t), \quad 0 < t < T \quad (48)$$

$$u(x, t) \rightarrow 0 \quad x \rightarrow \infty, \quad 0 < t < T \quad (49)$$

$$u(x, 0) = \delta(x - x_0), \quad b(0) < x < \infty \quad (50)$$

and the resulting survival probability is given, in terms of the solution $u(x, t)$, by

$$Pr(\tau > t) = P(t) = \int_{b(t)}^{\infty} u(x, t) dx \quad (51)$$

Motivated by the work of Avellaneda and Zhu [1], Lan Cheng, studied the inverse first-crossing problem [7]: given the survival probability $P(t)$ for $0 < t < T$, find the time-dependent absorbing boundary $b(t)$ in equation (48), including $b(0)$, such that equations (47)–(51) are satisfied. The more usual extra Neumann boundary condition appearing in FBP's can be obtained by differentiating equation (51):

$$\begin{aligned} P'(t) &= -u(b(t), t)b'(t) + \int_{b(t)}^{\infty} u_x(x, t) dx \\ &= \frac{-\sigma^2}{2} u_x(b(t), t) \end{aligned} \quad (52)$$

using the PDE (47) and the boundary conditions. With $-P'(t) = (1 - P(t))' = Q'(t) = q(t)$ denoting the default pdf, the extra boundary condition becomes

$$u_x(x, t) = \frac{2}{\sigma^2} q(t), x = b(t), 0 < t < T \quad (53)$$

Following the outline in the section Free Boundary Problems as Integral Equations, one can derive IEs for $b(t)$ in the form

$$\Gamma(b(t), t) = \int_0^t \Gamma(b(t) - b(s), t - s) q(s) ds \quad (54)$$

$$\begin{aligned} \frac{1}{2} q(t) &= \Gamma_x(b(t), t) \\ &\quad - \int_0^t \Gamma_x(b(t) - b(s), t - s) q(s) ds \end{aligned} \quad (55)$$

where Γ is the fundamental solution of the pdo in equation (47). A fast and accurate numerical scheme for solving

$$\begin{aligned} F(x, t) &= \Gamma(x, t) - \int_0^t \Gamma(x - b(s), t - s) q(s) ds = 0 \end{aligned} \quad (56)$$

for $x = b(t)$ (i.e., solving equation (54)) is the Newton—Raphson iteration

$$b(t)^{\text{new}} = b(t)^{\text{old}} - \frac{F(b(t)^{\text{old}}, t)}{q(t)/2} \quad (57)$$

where, in computing F_x in the denominator, we have used

$$\begin{aligned} \frac{1}{2} q(t) &= F_x(b(t), t) \simeq F_x(b(t)^{\text{old}}, t) \\ &= \simeq \Gamma_x(b(t)^{\text{old}}, t) - \int_0^t \Gamma_x(b(t)^{\text{old}} \\ &\quad - b(s)^{\text{old}}, t - s) q(s) ds \end{aligned} \quad (58)$$

Finally, we mention that an IEs formulation of the first passage problem for Brownian motion was given by Peskir [17] but the inverse problem described here was not treated. In [7], the proof of existence and uniqueness used viscosity solution methods. A proof using integral equations is still an open question for this problem as well as the two others listed in this section.

Acknowledgments

The author acknowledges support by NSF award DMS 0707953.

References

- [1] Avellaneda, M. & Zhu, J. (2001). Modeling the distance-to-default of a firm, *Risk* **14**, 125–129.
- [2] Barles, G., Burdeau, J., Romano, M. & Samsoen, N. (1995). Critical stock price near expiration, *Mathematical Finance* **5**, 77–95.
- [3] Carr, P., Jarrow, J. & Myneni, R. (1992). Alternative characterizations of American put option, *Mathematical Finance* **2**, 87–105.
- [4] Chen, X. & Chadam, J. (2003). Analytical and numerical approximations for the early exercise boundary for American put options, *Continuous, Discrete and Impulsive Systems, Series A: Mathematical Analysis* **10**, 649–660.
- [5] Chen, X. & Chadam, J. (2006). A mathematical analysis for the optimal exercise boundary of American put options, *SIAM Journal of Mathematical Analysis* **38**, 1613–1641.
- [6] Chen, X., Chadam, J., Jiang, L. & Zheng, W. (2008). Convexity of the exercise boundary of the American put

- option on a zero dividend asset, *Mathematical Finance* **18**, 185–197.
- [7] Cheng, L., Chen, X., Chadam, J. & Saunders, D. (2006). Analysis of an inverse first passage problem from risk management, *SIAM Journal of Mathematical Analysis* **38**, 845–873.
- [8] Crandall, M., Iskii, H. & Lions, P.L. (1992). User's guide to viscosity solutions of second order partial differential equations, *Bulletin of AMS* **27**, 1–67.
- [9] Ekstrom, E. (2004). Convexity of the optimal stopping time for the American put option, *Journal of Mathematical Analysis and Applications* **299**, 147–156.
- [10] Friedman, A. (1964). *Partial Differential Equations of Parabolic Type*, Prentice Hall.
- [11] Friedman, A. (1983). *Variational Principles and Free Boundary Problems* Lems, John Wiley & Sons.
- [12] Friedman, A. & Jensen, R. (1978). Convexity of the free boundary in the Stefan problem and in the dam problem, *Archive for Rational Mechanics and Analysis* **67**, 1–24.
- [13] Jiang, L., Bian, B. & Yi, F. (2005). A parabolic variational inequality arising from the valuation of fixed rate mortgages, *European Journal of Applied Mathematics* **16**, 361–383.
- [14] Leung, T. & Sircar, R. (2009). Accounting for risk aversion, vesting, job termination risk and multiple exercises in valuation of employee stock options, *Mathematical Finance* **19**, 99–128.
- [15] McKean, H.P. Jr. (1965). A free boundary problem for the heat equation arising from a problem of mathematical economics, *Industrial Management Review* **6**, 32–39.
- [16] Ockendon, J., Howison, S., Lacey, A. & Movchan, A. (2003). *Applied Partial Differential Equations*, Oxford University Press.
- [17] Peskir, G. (2002). On integral equations arising in the first-passage problem for Brownian motion, *The Journal of Integral Equations and Applications* **14**, 397–423.
- [18] Peskir, G. (2005). On the American option problem, *Mathematical Finance* **15**, 169–181.
- [19] Pham, H. (1997). Optimal stopping, free boundary and American option in a jump-diffusion model, *Applied Mathematics Optimization* **35**, 145–164.
- [20] Sirbu, M. & Shreve, S. (2006). A two-person game for pricing convertible bonds, *SIAM Journal of Control and Optimization* **45**, 1508–1639.
- [21] Xie, D., Chen, X. & Chadam, J. (2007). Optimal prepayment of mortgages, *European Journal of Applied Mathematics* **18**, 363–388.

Related Articles

American Options; Finite Difference Methods for Early Exercise Options; Structural Default Risk Models; Term Structure Models.

JOHN CHADAM

Tikhonov Regularization

An important issue in quantitative finance is *model calibration*. The calibration problem is the *inverse* of the pricing problem. Instead of computing prices in a model with given values for its parameters, one wishes to compute the values of the model parameters that are consistent with observed prices (up to the bid–ask spread).

Many examples of such inverse problems are ill-posed. Recall that a problem is *well posed* (as defined by Hadamard) if its solution exists, is unique, and depends continuously on its input data. Thus there are three reasons for which a problem might be ill posed:

- It admits no solution.
- It admits more than one solution.
- The solution/solutions to the inverse problem does/do not depend on the input data in a continuous way.

In the case of calibration problems in finance, except for trivial situations, there exists typically *no instance* of a given class of models that is *exactly* consistent with a full calibration data set, including a number of option prices, a zero-coupons curve, an expected dividend yield curve on the underlying, and so on. However, there are often *various* instances of a given class of models that fit the data *within the bid–ask spread*. In such cases, if one perturbs the data (e.g., if the observed prices change by some small amount between today and tomorrow), it is quite typical that a numerically determined best fit solution of the calibration problem switches from one “basin of attraction” to the other, and thus the numerically determined solution is *not stable* either.

To get a well-posed problem, we need to introduce some *regularization*. The most widely known and applicable regularization method is the *Tikhonov* (–*Phillips*) regularization method [9, 14, 16].

Tikhonov Regularization of Nonlinear Inverse Problems

We consider a Hilbert space $\mathcal{H}$, a closed convex nonvoid subset $\mathcal{A}$ of $\mathcal{H}$, a direct operator (“pricing functional”)

$$\mathcal{H} \supseteq \mathcal{A} \ni a \xrightarrow{\Pi} \Pi(a) \in \mathbb{R}^d \quad (1)$$

(so a corresponds to the set of model parameters), noisy data (“observed prices”) π^δ , and a *prior* $a_0 \in \mathcal{H}$ (a priori guess for a). The Tikhonov regularization method for *inverting* Π at π^δ , or estimating the model parameter a given the observation π^δ , consists in

- reformulating the inverse problem as the following *nonlinear least squares problem*:

$$\min_{a \in \mathcal{A}} \|\Pi(a) - \pi^\delta\|^2 \quad (2)$$

- to ensure *existence* of a solution;
- selecting the solutions of the previous nonlinear least squares problem that minimize $\|a - a_0\|^2$ over the set of all solutions; and
- introducing a trade-off between accuracy and regularity, parameterized by a level of regularization $\alpha > 0$, to ensure *stability*.

More precisely, we introduce the following *cost criterion*:

$$J_\alpha^\delta(a) \equiv \|\Pi(a) - \pi^\delta\|^2 + \alpha \|a - a_0\|^2 \quad (3)$$

Given α , δ , and a further parameter η , where η represents an error tolerance on the minimization, we define a *regularized solution to the inverse problem* for Π at π^δ as any model parameter $a_\alpha^{\delta, \eta} \in \mathcal{A}$ such that

$$J_\alpha^\delta(a_\alpha^{\delta, \eta}) \leq J_\alpha^\delta(a) + \eta, \quad a \in \mathcal{A} \quad (4)$$

Under suitable assumptions, one can show that the regularized inverse problem is well posed, as follows. We first postulate that the direct operator Π satisfies the following regularity assumption.

Assumption 1 (Compactness) $\Pi(a_n)$ converges to $\Pi(a)$ in $\mathbb{R}^d$ if a_n weakly converges to a in $\mathcal{H}$.

We then have the following *stability* result.

Theorem 1 (Stability) Let $\pi^{\delta_n} \rightarrow \pi^\delta$, $\eta_n \rightarrow 0$ when $n \rightarrow \infty$. Then any sequence of regularized solutions $a_n^{\delta_n, \eta_n}$ admits a subsequence that converges toward a regularized solution $a_\alpha^{\delta, \eta=0}$.

Assuming further that the data lie in the range of the model leads to *convergence* properties of

regularized solutions to (unregularized) solutions of the inverse problem as $midi \rightarrow 0$. Let us then make the following additional assumption on Π .

Assumption 2 (Range Property) $\pi \in \Pi(\mathcal{A})$.

Definition 1 By an a_0 -solution to the inverse problem for Π at π , we mean any $a \in \underset{\{\Pi(a)=\pi\}}{\text{Argmin}} \|a - a_0\|$. Note that the set of a_0 -solutions is non-empty, by Assumption 2.

Theorem 2 (Convergence; see, for instance, Theorem 2.3 of Engl et al [10]) Let the perturbed parameters $\alpha_n, \delta_n, \eta_n$ and the perturbed data $\pi_n \in \mathbb{R}^d$ satisfy

$$\begin{aligned} (n \in \mathbb{N}) \quad & \|\pi - \pi_n\| \leq \delta_n \\ (n \rightarrow \infty) \quad & \alpha_n, \quad \delta_n^2/\alpha_n, \quad \eta_n/\alpha_n \longrightarrow 0 \end{aligned} \quad (5)$$

Then any sequence of regularized solutions $a_{\alpha_n, \eta_n}^{\delta_n}$ admits a subsequence that converges toward an a_0 -solution a of the inverse problem for Π at π . In particular, in the case when this problem admits a unique a_0 -solution a , $a_{\alpha_n, \eta_n}^{\delta_n}$ converges to a .

Remark 1 In the special case where the direct operator Π is linear, Tikhonov regularization thus appears as an approximating scheme for the pseudoinverse of Π .

Finally, assuming further regularity of Π , one can get *convergence rates* estimates, uniform over all data $\pi \in \Pi(\mathcal{A})$ sufficiently close and smooth with respect to the prior a_0 (so that the additional *source condition* 12 is satisfied). Let us thus make the following additional assumption on Π .

Assumption 3 (Twice Gateaux Differentiability)

There exist linear and bilinear forms $d\Pi(a)$ on $\mathcal{H}$ and $d^2\Pi(a)$ on $\mathcal{H}^2$ such that

$$\begin{aligned} \Pi(a + \varepsilon h) &= \Pi(a) + \varepsilon d\Pi(a) \cdot h \\ &+ \frac{\varepsilon^2}{2} d^2\Pi(a) \cdot (h, h) + o(\varepsilon^2); \\ a, a + h &\in \mathcal{A} \end{aligned} \quad (6)$$

$$\begin{aligned} \|d\Pi(a) \cdot h\| &\leq C \|h\|, \\ \|d^2\Pi(a) \cdot (h, h')\| &\leq C \|h\| \|h'\|; \\ a \in \mathcal{A}, \quad h, h' &\in \mathcal{H} \end{aligned} \quad (7)$$

where C is a constant independent of $a \in \mathcal{A}$.

In the following theorem, the operator

$$d\Pi(a)^* : \mathbb{R}^d \ni \lambda \mapsto d\Pi(a)^* \lambda \in \mathcal{H}^1 \quad (8)$$

denotes the adjoint of

$$d\Pi(a) : \mathcal{H}^1 \ni h \mapsto d\Pi(a) h \in \mathbb{R}^d \quad (9)$$

in the sense that (see [9])

$$\langle h, d\Pi(a)^* \lambda \rangle_{\mathcal{H}^1} = \lambda' d\Pi(a) \cdot h; \quad (h, \lambda) \in \mathcal{H}^1 \times \mathbb{R}^d \quad (10)$$

Theorem 3 (Convergence Rates; see, for instance, Theorem 10.4 of Engl et al [9]) Assume

$$\begin{aligned} (n \in \mathbb{N}) \quad & \|\pi - \pi_n\| \leq \delta_n, \\ (n \rightarrow \infty) \quad & \alpha_n \longrightarrow 0, \quad \alpha_n \sim \delta_n, \quad \eta_n = O(\delta_n^2) \end{aligned} \quad (11)$$

Then $\|a_{\alpha_n, \eta_n}^{\delta_n} - a\| = O(\sqrt{\delta_n})$, for any a_0 -solution a of the inverse problem for Π at π such that

$$a - a_0 = d\Pi(a)^* \lambda \quad (12)$$

for some λ sufficiently small in $\mathbb{R}^d$ (in particular, there exists at most one such a_0 -solution a).

Remark 2 An interesting feature of Tikhonov regularization is that the data set π does not need to belong to the range of the direct operator for applicability of the method—even if Assumption 2 is the simplest assumption for the previous results regarding convergence and convergence rates (in fact, a minimal assumption for such results is the existence of a least squares solution to the inverse problem; see Proposition 3.2 of Binder et al [2]).

An important issue, in practice, is the choice of the *regularization parameter* α that determines the trade-off between accuracy and regularity in the method. To set α , the two main approaches are

- *a priori* methods, in which the choice of α only depends on δ , the level of noise on the data (such as the size of the bid–ask spread, in the case of market prices data in finance);
- more general *a posteriori* methods, in which α may depend on the data in a less specific way.

In applications to calibration problems in finance, the most commonly used method for choosing α is the *a posteriori* method based on the so-called *discrepancy principle*, which consists in choosing the greatest level of α for which the “distance” $\|\Pi(a_{\alpha}^{\delta,\eta}) - \pi^{\delta}\|$ (for given δ, η) does not exceed the level of noise δ on the observations (as measured by the bid–ask spread).

Implementation

For implementation purposes, the minimization problem 3 is discretized, and thus it effectively becomes a *nonlinear minimization problem* on (some subset of) $\mathbb{R}^k$ (see, e.g., [13]), where k is the number of model parameters to be estimated.

In the case of a *strictly convex* cost criterion $J = J_{\alpha}^{\delta}$ in equation 3, and if, additionally, J is differentiable, one can prove the convergence to the (unique) minimum of various *gradient descent algorithms*. These consist, at each iteration, in making a step of a certain length (fixed-step descent vs. optimal step descent) in a direction defined by the gradient ∇J in the current step of the algorithm, in combination with, in some variants of the method (*conjugate gradient method, quasi-Newton algorithms, etc.*), the gradient(s) ∇J in the previous step(s).

In the *nonstrictly convex* case, (actually, in the context of calibration problems in finance, J is typically not even convex with respect to α), or if the cost criterion is only almost everywhere differentiable (as in the *American calibration problem*, see Remark 3 (i)), such algorithms can still be used, in which case, they typically converge to one among many *local minima* of J .

When there are no constraints (the case when $\mathcal{A} = \mathcal{H}$), the minimization problem is, in practice, much easier, and many implementations of the related gradient descent algorithms are available (e.g., in [15]). As for constrained problems, a state-of-the-art open-source implementation of the quasi-Newton method for minimizing a function in a box, the L-BFGS algorithm, is available on www.ece.northwestern.edu/~nosedal/lbfgsb.html.

When the gradient ∇J is neither computable in closed form nor computable numerically with the required accuracy, an alternative to gradient descent methods is to use the *nonlinear simplex method* (not to be confused with the simplex algorithm for solving linear programming problems, see [15]). As

opposed to gradient descent methods, the nonlinear simplex algorithm only uses the *values* (and not the *gradient*) of J , but the convergence of the algorithm is not proved in general, and there are known counterexamples in which it does not converge.

Application: Extracting Local Volatility

In the case of *parametric* models in finance, namely, models with a small number of *scalar* parameters, such as Heston’s stochastic volatility model or Merton’s jump-diffusion model (as opposed to models with *functional*, e.g., time-dependent, parameters), the choice of a suitable regularization term is generally not obvious. In this case, the calibration industry standard rather consists in solving the unregularized nonlinear least squares problem 2. So Tikhonov regularization is rather used for calibrating *nonparametric* financial models.

In this section, we consider the problem of inferring a *local volatility function* $\sigma(t, S)$ (see [7]) from observed option prices, namely, European vanilla calls and/or puts with various strikes and maturities on the underlying S . The local volatility function thus inferred may then be used to price exotic options and/or compute Greeks, consistently with the market (e.g., [5]).

The Ill-posed Local Volatility Calibration Problem

The local volatility calibration problem, however, is underdetermined (since the set of observed prices is finite whereas the nonparametric function σ has an infinite degrees of freedom) and ill posed. So a naive approach based on numerical differentiation using the so-called *Dupire’s formula* [7] gives a local volatility that is highly oscillatory (Figure 1), and thus unstable, for instance when performing a day-to-day calibration.

To address this issue, the first idea that comes to mind is to seek for σ within a parameterized family of functions. However, finding classes of functions with all the flexibility required for fitting implied volatility surfaces with several hundred implied volatility points and a variety of shapes turns out to be a very challenging task (unless a large family of splines is considered, see Coleman *et al.* [3], in which case, the ill-posedness of the problem shows up again).

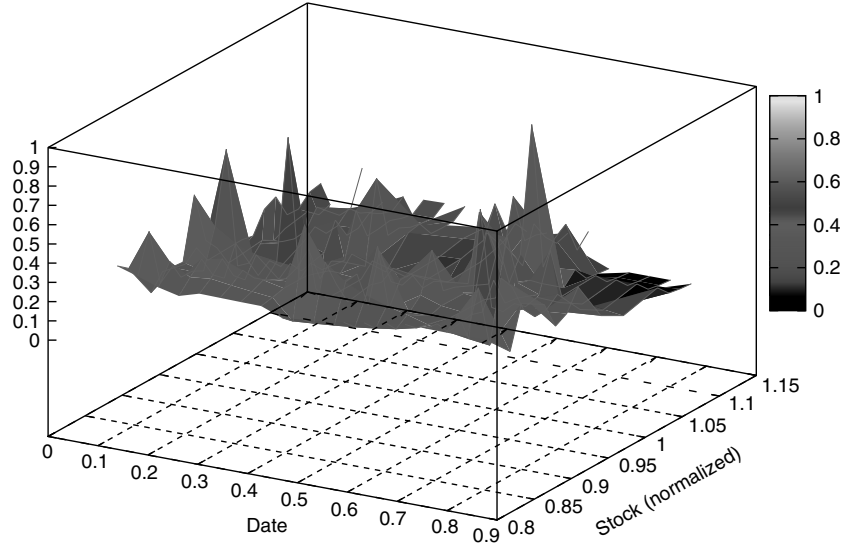


Figure 1 Local Variance $\sigma(t, S)^2$ obtained by application of Dupire's formula on the DAX index, May 2, 2001

The best way to proceed is to stay nonparametric, and to use regularization methods to stabilize the calibration procedure. Since we use a nonparametric local volatility, the model contains sufficient number of degrees of freedom to provide a perfect fit to virtually any market smile. Furthermore, the regularization method guarantees that the local volatility thus calibrated is nice and smooth.

Approach by Tikhonov Regularization

Among the various regularization methods at hand, the most popular one is the Tikhonov regularization method described in the section Tikhonov Regularization of Nonlinear Inverse Problems. One thus rewrites the local volatility calibration problem as the following nonlinear minimization problem:

$$\begin{aligned} \min_{\{\sigma \equiv \sigma(t, S); \underline{\sigma} \leq \sigma \leq \bar{\sigma}\}} J(\sigma) \\ = \|\Pi(\sigma) - \pi\|^2 + \alpha \|\sigma - \sigma_0\|_{\mathcal{H}^1}^2 \end{aligned} \quad (13)$$

where

- the bounds $\underline{\sigma}$ and $\bar{\sigma}$ are given positive constants specifying the abstract set $\mathcal{A}$ in the section Tikhonov Regularization of Nonlinear Inverse Problems ;

- π is the vector of market prices observed at the calibration time; $\Pi(\sigma)$ is the related vector of prices in the Dupire model with volatility function σ ;
- σ_0 is a suitable prior (*a priori* guess on σ), and for $u \equiv u(t, S)$,

$$\begin{aligned} \|u\|_{\mathcal{H}^1}^2 = \int_{t_0}^{\infty} \int_0^{\infty} [u(t, S)^2 + (\partial_t u(t, S))^2 \\ + (\partial_S u(t, S))^2] dt dS \end{aligned} \quad (14)$$

Problem 13 and a related gradient descent approach to solve it numerically (cf. the section Implementation) were introduced in [12]. Crépey [6] (see also [8]) further showed that the general conditions of the section Tikhonov Regularization of Nonlinear Inverse Problems are satisfied in this case. Stability and convergence of the method follow.

In [5] an efficient trinomial tree implementation of this approach was presented, based on an exact computation of the gradient of the (discretized) cost criterion J in equation 13. Figure 2 displays the local variance surface $\sigma(t, S)^2$ (to be compared with that of Figure 1), the corresponding implied volatility surface, and the accuracy of the calibration, obtained by running this algorithm on the DAX index European options data set of May 2, 2001 (consisting

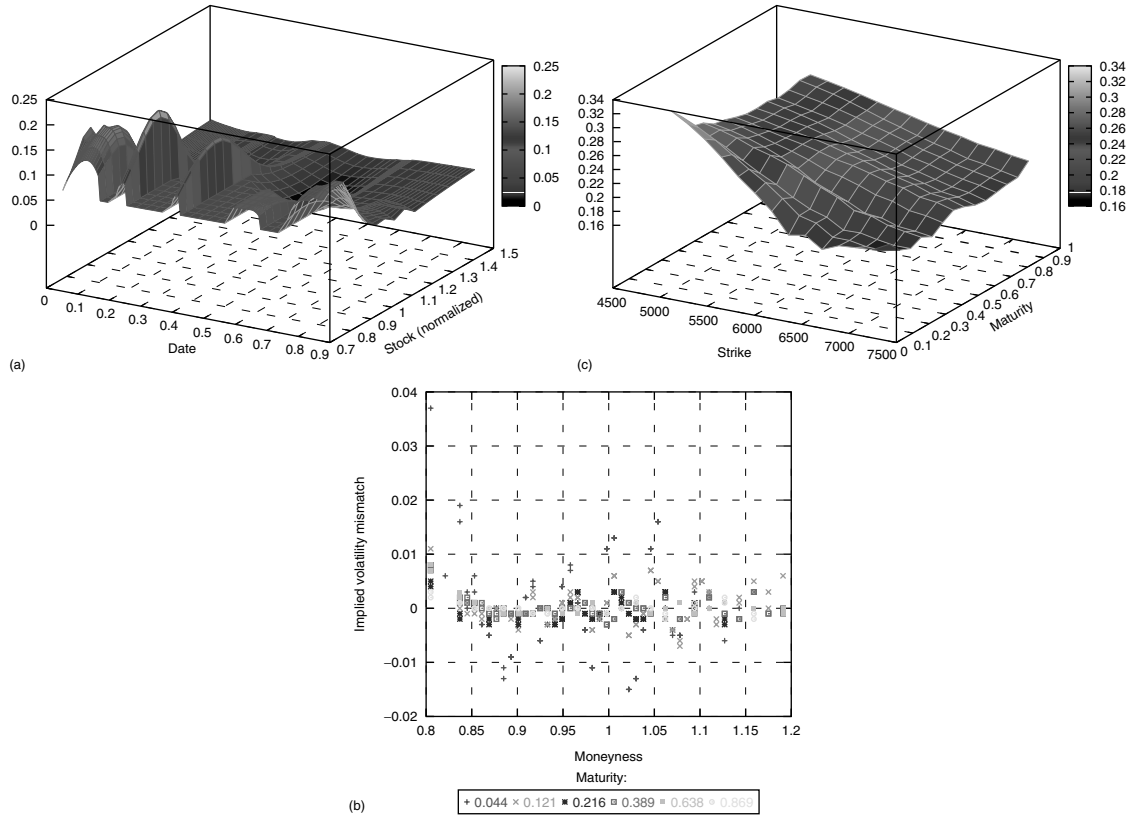


Figure 2 (a) Local variance, (b) implied volatility, and (c) calibration accuracy obtained by application of the Tikhonov regularization method on the DAX index, May 2, 2001

of about 300 European vanilla option prices distributed throughout six maturities with moneyness $K/S_0 \in [0.8, 1.2]$. At the initiation of the algorithm, the norm of the gradient of the cost criterion J in equation 13 was equal to $5.73\text{E} - 02$, and upon convergence after 65 iterations of the gradient descent algorithm, a local minimum of the cost criterion was found, with related value of the norm of the gradient of the cost criterion equal to $6.83\text{E} - 07$. In the accuracy graph, implied volatility mismatch refers to the difference between the Black–Scholes implied volatility corresponding to the market price of an option and its price in the calibrated local volatility model, for each option in the calibration data set.

Such calibration procedures are typically computationally intensive; however, it is possible to make them faster by resorting to *parallel computing* (see Table 1 and [5]).

Table 1 Calibration CPU times on a cluster of $nproc$ 1.3-GHz processors connected on a fast Myrinet network, using a calibration tree with n time steps (thus $n^2/2$ nodes in the tree)

$n \times nproc$	1	3	6
54	25s	9s	10s
101	4m30s	1m57s	1m36s

Remark 3 (1) This approach by Tikhonov regularization can be extended to the problem of calibrating a local volatility function using *American* observed option prices as input data [5], or to the problem of calibrating a *Lévy model with local jump measure* (see [4] and [11]).

(2) An alternative approach for this problem is to use *entropic regularization*, rewriting the local volatility calibration problem as a related *stochastic control problem* [1].

References

- [1] Avellaneda, M., Friedman, C., Holmes, R. & Samperi, D. (1997). Calibrating volatility surfaces via relative-entropy minimization, *Applied Mathematical Finance* **41**, 37–64.
- [2] Binder, A., Engl, H.W., Groetsch, C.W., Neubauer, A. & Scherzer, O. (1994). Weakly closed nonlinear operators and parameter identification in parabolic equations by Tikhonov regularization, *Applicable Analysis* **55**, 13–25.
- [3] Coleman, T., Li, Y. & Verma, A. (1999). Reconstructing the unknown volatility function, *Journal of Computational Finance* **2**(3), 77–102.
- [4] Cont, R. & Rouis, M. (2006). *Estimating Exponential Lévy Models from Option Prices via Tikhonov Regularization*, Working Paper.
- [5] Crépey, S. (2003). Calibration of the local volatility in a trinomial tree using Tikhonov regularization, *Inverse Problems* **19**, 91–127.
- [6] Crépey, S. (2003). Calibration of the local volatility in a generalized Black–Scholes model using Tikhonov regularization, *SIAM Journal on Mathematical Analysis* **34**(5), 1183–1206.
- [7] Dupire, B. (1994). Pricing with a smile, *Risk* **7**, 18–20.
- [8] Egger, H. & Engl, H.W. (2005). Tikhonov regularization applied to the inverse problem of option pricing: convergence analysis and rates, *Inverse Problems* **21**, 1027–1045.
- [9] Engl, H.W., Hanke, M. & Neubauer, A. (1996). *Regularization of Inverse Problems*, Kluwer, Dordrecht.
- [10] Engl, H.W., Kunisch, K. & Neubauer, A. (1989). Convergence rates for Tikhonov regularisation of nonlinear ill-posed problems, *Inverse Problems* **5**(4), 523–540.
- [11] Kindermann, S., Mayer, P., Albrecher, H. & Engl, H.W. (2008). Identification of the local speed function in a Lévy model for option pricing, *Journal of Integral Equations and Applications* **20**(2), 161–200.
- [12] Lagnado, R. & Osher, S. (1997). A technique for calibrating derivative security pricing models: numerical solution of an inverse problem, *Journal of Computational Finance* **1**(1), 13–25.
- [13] Nocedal, J. & Wright, S.J. (2006). *Numerical Optimization*, 2nd Edition, Springer.
- [14] Phillips, D. (1962). A technique for the numerical solution of certain integral equations of the first kind, *Journal of the ACM* **9**, 84–97.
- [15] Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1992). *Numerical Recipes in C: The Art of Scientific Computing*, 2nd Edition, Cambridge University Press.
- [16] Tikhonov, A. (1963). Solution of incorrectly formulated problems and the regularization method, *Soviet Mathematics Doklady* **4**, 1035–1038, English translation of *Doklady Akademii Nauk SSSR* **151**, 501–504, 1963.

Related Articles

Conjugate Gradient Methods; Dupire Equation; Local Volatility Model; Model Calibration; Tree Methods.

STÉPHANE CRÉPEY

Tree Methods

Tree methods are among the most popular numerical methods to price financial derivatives. Mathematically speaking, they are easy to understand and do not require severe implementation skills to obtain algorithms to price financial derivatives. Tree methods basically consist in approximating the diffusion process modeling the underlying asset price by a discrete random walk.

In fact, the price of a European option of maturity T can always be written as an expectation either of the form

$$\mathbb{E}(e^{-rT} \psi(S_T))$$

in the case of vanilla options or of the form

$$\mathbb{E}(e^{-rT} \psi(S_t, 0 \leq t \leq T))$$

in the case of exotic options, where $(S_t, t \geq 0)$ is a stochastic process describing the evolution of the stock price, ψ is the payoff function, and r is the instantaneous interest rate. The basic idea of tree methods is to approximate the stochastic process $(S_t, t \geq 0)$ by a discrete Markov chain $(\tilde{S}_n^N, n \geq 0)$, such that

$$\begin{aligned} \mathbb{E}(e^{-rT} \psi(S_t, 0 \leq t \leq T)) \\ \approx \mathbb{E}(e^{-rT} \psi(\tilde{S}_n^N, 0 \leq n \leq N)) \end{aligned} \quad (1)$$

for N large enough ($\approx$ is used to remind the reader that the equality is only guaranteed for $N = \infty$). To ensure the quality of the approximation, we are interested in a particular notion of convergence called *convergence in distribution* (weak convergence) of discrete Markov chains to continuous stochastic processes. It is interesting to note that tree methods can also be regarded as a particular case of explicit finite difference algorithms.

Tree methods provide natural algorithms to price both European and American options when the risky asset is modeled by a geometric Brownian motion (see [27], for an introduction on how to use tree methods in financial problems). When considering more complex models—such as models with jumps or stochastic volatility models—the use of tree methods

is much more difficult; analytic approaches like finite difference (see **Partial Differential Equations**; **Partial Integro-differential Equations (PIDEs)**) or finite element (see **Finite Element Methods**) methods are usually preferred, and Monte Carlo methods are also widely used for complex models. Binomial (see **Binomial Tree**) and trinomial trees may also be constructed to approximate the stochastic differential equation governing the short rate [21, 26] or the intensity of default [32], permitting hereby to obtain the price of, respectively, interest rate derivatives or credit derivatives. Implied binomial trees, which enable us to construct trees consistent with the market prices of plain vanilla options, are a generalization of the standard tree methods used to price more exotic options (see [11, 14]).

For the sake of simplicity, consider a market model where the evolution of the risky asset is driven by the Black–Scholes stochastic differential equation

$$dS_t = S_t(r \, dt + \sigma \, dW_t), \quad S_0 = s_0 \quad (2)$$

in which $(W_t)_{0 \leq t \leq T}$ is a standard Brownian motion (under the so-called risk neutral probability measure) and the positive constant σ is the volatility of the risky asset.

The seminal work of Cox–Ross–Rubinstein (CRR) [10] initiated the use of tree methods and many variants have been introduced to improve the quality of the approximation when pricing plain vanilla or exotic options.

Plain Vanilla Options

The multiplicative binomial CRR model [10] is interesting on its own as a basic discrete-time model for the underlying asset of a financial derivative, since it converges to a log-normal diffusion process under appropriate conditions. One of its most attractive features is the ease of implementation to price plain vanilla options by backward induction.

Let N denote the number of steps of the tree and $\Delta T = T/N$ the corresponding time step. The log-normal diffusion process $(S_{n\Delta T})_{0 \leq n \leq N}$ is approximated by the CRR binomial process $(\tilde{S}_n^N = s_0 \prod_{j=1}^n Y_j)_{0 \leq n \leq N}$ where the random variables $Y_1, \dots, Y_N$ are independent and identically distributed with values in $\{d, u\}$ (u is called the *up factor* and d the *down factor*) with $p_u = \mathbb{P}(Y_n = u)$

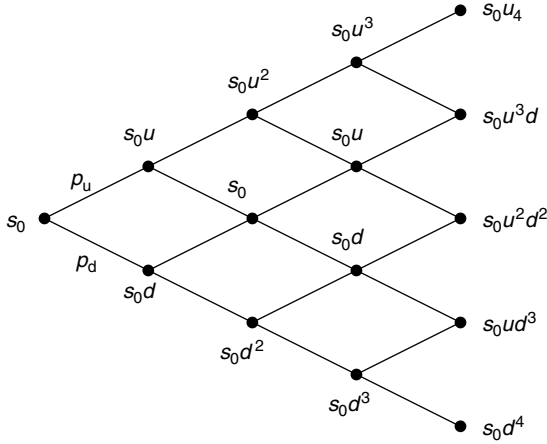


Figure 1 CRR tree

and $p_d = \mathbb{P}(Y_n = d)$. The dynamics of the binomial tree (see Figure 1) is given by the following Markov chain:

$$\bar{S}_{n+1}^N = \begin{cases} \bar{S}_n^N u & \text{with probability } p_u \\ \bar{S}_n^N d & \text{with probability } p_d \end{cases} \quad (3)$$

Kushner and Dupuis [23] proved that the local consistency conditions given by equation (4)—that is, the matching at the first order in ΔT of the first and second moments of the logarithmic increments of the approximating chain with those of the continuous-time limit—grant the convergence in distribution:

$$\begin{cases} \mathbb{E} \log \frac{\bar{S}_{n+1}^N}{\bar{S}_n^N} = \mathbb{E} \log \frac{S_{(n+1)\Delta T}}{S_{n\Delta T}} + o(\Delta T) \\ \mathbb{E} \log^2 \frac{\bar{S}_{n+1}^N}{\bar{S}_n^N} = \mathbb{E} \log^2 \frac{S_{(n+1)\Delta T}}{S_{n\Delta T}} + o(\Delta T) \end{cases} \quad (4)$$

This first-order matching condition rewrites

$$\begin{cases} p_u \log u + (1 - p_u) \log d = (r - \frac{\sigma^2}{2})\Delta T \\ p_u \log^2 u + (1 - p_u) \log^2 d = \sigma^2 \Delta T \end{cases} \quad (5)$$

The usual CRR tree corresponds to the choice $u = 1/d = e^{\sigma\sqrt{\Delta T}}$, which leads to $p_u = e^{r\Delta T} - e^{-\sigma\sqrt{\Delta T}} / (e^{\sigma\sqrt{\Delta T}} - e^{-\sigma\sqrt{\Delta T}}) = 1/2 + r - \sigma^2/2/2\sigma\sqrt{\Delta T} + \mathcal{O}(\Delta T^{3/2})$. When ΔT is small enough (i.e., for N large), the above value of p_u belongs to $[0,1]$. For this choice of u , d , and p_u , the difference between both sides of each equality in equation (5) is of order $(\Delta T)^2$. This is sufficient to ensure the convergence to the Black–Scholes model when N tends to infinity.

As $(\bar{S}_n)_n$ defined by the CRR model is Markovian, the price at time $n \in \{0, \dots, N\}$ of an American put option (see **American Options**) in the CRR model with maturity T and strike K can be written as $v(n, \bar{S}_n)$ where the function $v(n, x)$ can be computed by the following backward dynamic programming equations:

$$\begin{cases} v_N(N, x) = (K - x)^+ \\ v_N(n, x) = \max(\psi(x), \\ e^{-r\Delta T} [p_u v_N(n+1, xu) \\ v_N(n+1, xd)]) \end{cases} \quad (6)$$

where $\psi(x) = (K - x)^+$. Note that the algorithm requires the comparison between the intrinsic value and the continuation value. When considering European options, $\psi \equiv 0$.

The initial price of a put option in the Black–Scholes model can be approximated by $v_N(0, s_0)$. The initial delta, which is the quantity of risky asset in the replicating portfolio on the first time step in the CRR model, is approximated by $v_N(1, s_0u) - v_N(1, s_0d)/s_0(u - d)$. Note that to obtain the approximated price and delta, one only needs to compute $(v_N(n, s_0u^k d^{n-k}), 0 \leq k \leq n)$ by backward induction on n from N to 0. Figure 2 gives an example of backward computation of the price of an American put option using $N = 4$ time steps. The complexity of the algorithm is of order N^2 , more precisely the function v_N has to be evaluated at $(N + 1)(N + 2)/2$ nodes.

For the computation of the delta, Pelsser and Vorst [28] suggested to enlarge the original tree by adding two new initial nodes generated by an extended two-period back tree (dashed lines in Figure 2). To achieve the convergence in distribution, many other choices for u , d , p_u , and p_d can be made, leading to as many other Markov chains. We may choose other two-point schemes, such as a random walk approximation of the Brownian motion, three-point schemes (trinomial trees), or more general p -point schemes. The random walk approximation of the Brownian motion $(Z_{n+1} = Z_n + U_{n+1})$ with $(U_i)_i$ independent and identically distributed with $\mathbb{P}(U_i = 1) = \mathbb{P}(U_i = -1) = 1/2$ can be used as long as S_T is given by

$$S_T = s_0 e^{(r - \sigma^2/2)T + \sigma W_T} \quad (7)$$

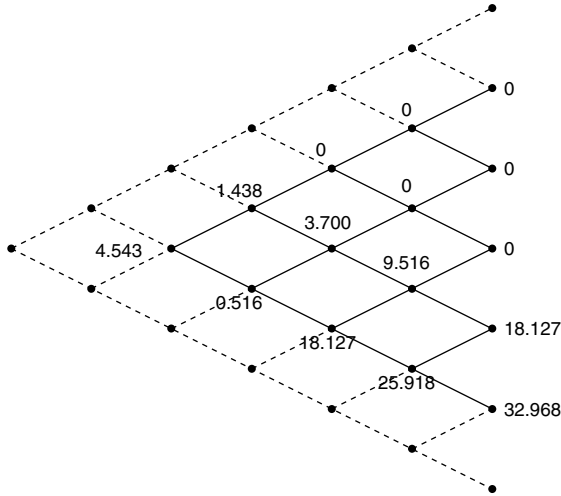


Figure 2 Backward induction for a CRR tree with $N = 4$ for an American put option with parameters $s_0 = K = 100$, $r = 0.1$, $\sigma = 0.2$, $T = 1$

This leads to $p_u = 1/2$ and

$$\begin{aligned} u &= e^{(r-\sigma^2/2)\Delta T + \sigma\sqrt{\Delta T}} \\ d &= e^{(r-\sigma^2/2)\Delta T - \sigma\sqrt{\Delta T}} \end{aligned} \quad (8)$$

The most popular trinomial tree has been introduced by Kamrad and Ritchken [22] who have chosen to approximate $(S_{n\Delta T})_{0\leq n\leq N}$ by a symmetric three-point Markov chain $(\tilde{S}_n)_{0\leq n\leq N}$

$$\bar{s}_{n+1} = \begin{cases} \bar{s}_n u & \text{with probability } p_u \\ \bar{s}_n & \text{with probability } p_m \\ \bar{s}_n d & \text{with probability } p_d \end{cases} \quad (9)$$

The convergence is ensured as soon as the first two moment matching condition on $\log(S_{\Delta T}/s_0)$ is satisfied. With $u = e^{\lambda\sigma\sqrt{\Delta T}}$ and $d = 1/u$, this condition leads to

$$\begin{aligned} p_u &= \frac{1}{2\lambda^2} + \frac{(r - \sigma^2/2)\sqrt{\Delta T}}{2\lambda\sigma}, & p_m &= 1 - \frac{1}{\lambda^2} \\ p_d &= \frac{1}{2\lambda^2} - \frac{(r - \sigma^2/2)\sqrt{\Delta T}}{2\lambda\sigma} \end{aligned} \quad (10)$$

The parameter λ —the stretch parameter—appears as a free parameter of the geometry of the tree, which can be tuned to improve the convergence. The value $\lambda \approx 1.22474$ corresponding to $p_m = 1/3$

is reported to be a good choice for at-the-money plain vanilla options. Note that choosing $u = 1/d$ is essential to avoid a complexity explosion. In this case, the complexity is still of order N^2 , but this time $(N + 1)(N + 3)$ evaluations of the function v_N are required. The value $\lambda = 1$ corresponds to the CRR tree.

Note that complexity is intimately related to the quality of the approximation. Therefore, one should always try to balance the additional computational cost with the improvement of the convergence it brings out.

Convergence Issues

Over the last years, significant advances have been made in understanding the convergence behavior of tree methods for option pricing and hedging (see [13, 24, 25, 33]). As noted in [12, 15], there are two sources of error in tree methods for call (**Call Options**) or put option: the first one (*the distribution error*) ensues from the approximation of a continuous distribution by a discrete one, whereas the second one (*the nonlinearity error*) stems from the interplay between the strike and the grid nodes at the final time step. Because of the nonlinearity error, the convergence is slow except for at-the-money options. Let P_N^{CRR} and P^{BS} denote the initial price of the European put option with maturity T and strike K , respectively, in the CRR tree (with N steps) and in the Black–Scholes model. Using the call–put parity relationship in both the models and the results given for the call option in [13], one finds

$$P_N^{\text{CRR}} = P^{\text{BS}} - \frac{Ke^{-rT}}{N} e^{-d_2^2/2} \sqrt{\frac{2}{\pi}} \\ \times [\kappa_N(\kappa_N - 1)\sigma\sqrt{T} + D_1] + \mathcal{O}\left(\frac{1}{N^{3/2}}\right) \quad (11)$$

where $d_2 = (\log(s_0/K) + (r - (\sigma^2/2)T)/\sigma\sqrt{T})$, κ_N denotes the fractional part of $\log(K/s_0)/2\sigma\sqrt{N/T} - N/2$ and D_1 is a constant. For at-the-money options (i.e., $K = s_0$), $\kappa_N = 0$ for N even and $\kappa = 1/2$ for N odd; hence, the difference $(P_N^{\text{CRR}} - P^{\text{BS}})_N$ is an alternating sequence. Figure^a 3 shows that for N even P_N^{CRR} gives an upper estimate of P^{BS} , whereas for N odd P_N^{CRR} gives a lower estimate. The monotonicity of $(P_{2N+1})_N$ and $(P_{2N})_N$ for at-the-money

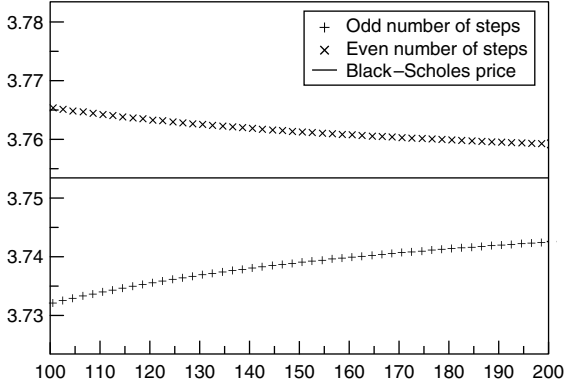


Figure 3 Convergence for an at-the-money put option with parameters $s_0 = K = 100$, $r = 0.1$, $\sigma = 0.2$, $T = 1$

options enables us to use a Richardson extrapolation (i.e., consider $2P_{4N}^{\text{CRR}} - P_{2N}^{\text{CRR}}$ or $2P_{4N+1}^{\text{CRR}} - P_{2N+1}^{\text{CRR}}$) to make the terms of order $1/N$ disappear. For not at-the-money options, Figure 4 shows that the sequences $(P_{2N+1}^{\text{CRR}})_N$ and $(P_{2N}^{\text{CRR}})_N$ are not monotonic and present an oscillating behavior. In this context, a naive Richardson extrapolation performs badly.

Several tree methods [6, 15, 18] try to deal with the nonlinearity error at maturity reproducing in some sense an at-the-money situation. The binomial Black-Scholes (BBS) method introduced by Broadie and Detemple [6] replaces at each node of the last but one time step before maturity, the continuation value with the Black-Scholes European one [3]. A two-point Richardson extrapolation aiming

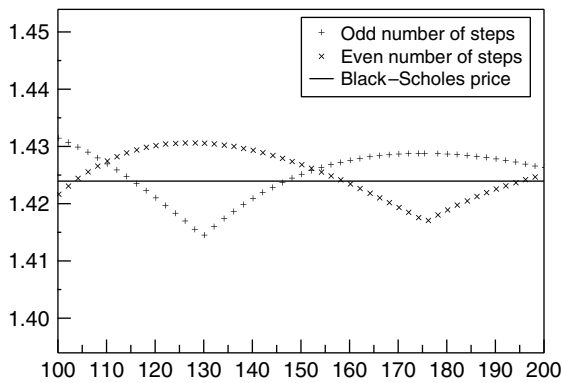


Figure 4 Convergence for a not at-the-money put option with parameters $s_0 = 100$, $K = 90$, $r = 0.1$, $\sigma = 0.2$, $T = 1$

at improving the convergence leads to the binomial Black-Scholes-Richardson (BBSR) method.

Adaptive mesh model (AMM) is a trinomial-based method introduced by Figlewski and Gao [15]. By taking into account that the nonlinearity error at maturity only affects the node nearest to the strike, AMM resorts to thickening the trinomial grid only around the strike and only at maturity time.

The binomial interpolated (with Richardson extrapolation) method (BI(R)) introduced by Gaudenzi and Pressacco [18] tries to recover the regularity of the sequences giving the CRR price of the European at-the-money options. The logic of the BI approach is to create a set of computational options, each one with a computational strike lying exactly on a final node of the tree. The value of the option with the contractual strike is then obtained by interpolating the values of the computational options.

Let us finally remark that similar techniques have been developed in numerical analysis for PDEs associated to option pricing problems (see [19]). In the case of nonsmooth initial conditions, to get good convergence of finite element and finite difference methods, there should always be a node at the strike and the payoff may have to be smoothed (see [30] for a classical reference).

In the American option case, a new source of error arises compared to the European case: the loss of opportunity to early exercise in any interval between two discrete times of the tree.

Exotic Options

The classical CRR approach may be troublesome when applied to barrier options (see **Barrier Options**) because the convergence is very slow in comparison with plain vanilla options. The reason is quite obvious: let L be the barrier and n_L denote the index such that

$$s_0 d^{n_L} \geq L > s_0 d^{n_L+1}$$

Then, the binomial algorithm yields the same result for any value of the barrier between $s_0 d^{n_L}$ and $s_0 d^{n_L+1}$, while the limiting value changes for every barrier L .

Several different approaches have been proposed to overcome this problem.

Boyle and Lau [5] choose the number of time steps in order to have a layer of nodes of the tree as close

as possible to the barrier. Ritchken [31] noted that the trinomial method is more suitable than the binomial one. The main idea is to choose the stretch parameter λ such that the barrier is exactly hit. Later, Cheuk and Vorst [9] presented a modification of the trinomial tree (based on a change of the geometry of the tree), which enables to set a layer of the nodes exactly on the barrier for any choice of the number of time steps. Numerical interpolation techniques have also been provided by Derman *et al.* [12].

In the case of Asian options (*see Asian Options; Lattice Methods for Path-dependent Options*) with arithmetic average, the CRR method is not efficient since the number of possible averages increases exponentially with the number of the steps of the tree. For this reason, Hull and White [20] and in a similar way Barraquand and Pudet [2] proposed more feasible approaches. The main idea of their procedure is to restrict the possible arithmetic averages to a set of representative values. These values are selected in order to cover all the possible values of the averages reachable at each node of the tree. The price is then computed by a backward induction procedure, whereas the prices associated to the averages outside of representative value set are obtained by some suitable interpolation methods.

These techniques drastically reduce the computation time compared to the pure binomial tree; however, they present some drawbacks (convergence and numerical accuracy) as observed by Forsyth *et al.* [16]. Chalasani *et al.* [7, 8] proposed a completely different approach to obtain precise upper and lower bounds on the pure binomial price of Asian options. This algorithm significantly increases the precision of the estimates but induces a different problem: the implementation requires a lot of memory compared to the previous methods.

In the case of lookback options (*see Lookback Options*), the complexity of the pure binomial algorithm is of order $O(N^3)$ and the methods proposed in [2, 20] do not improve the efficiency. Babbs [1] gave a very efficient and accurate solution to the problem for American floating strike lookback options by using a procedure of complexity of order $O(N^2)$. He proposed a change of “numeraire” approach, which cannot be applied in the fixed strike case.

Gaudenzi *et al.* [17] introduced the singular point method to price American path-dependent options. The main idea is to give a continuous representation

of the option price function as a piecewise linear convex function of the path-dependent variable. These functions are characterized only by a set of points named *singular points*. Such functions can be evaluated by backward induction in a straightforward manner. Hence, this method provides an alternative and more efficient approach to evaluate the pure binomial prices associated with the path-dependent options. Moreover, because the piecewise linear function representing the price is convex, it is easy to obtain upper and lower bounds of the price.

For the rainbow options, extensions of the binomial approach for pricing American options on two or more stocks have been made by Boyle *et al.* [4] and Kamrad and Ritchken [22]. In higher dimensional problems (say, dimension greater than three), the straightforward application of tree methods fails because of the so-called curse of dimension: the computational cost and the memory requirement increase exponentially with the dimension of the problem.

End Notes

^aThe graphics have been generated using PREMIA software [29].

References

- [1] Babbs, S. (2000). Binomial valuation of lookback options, *Journal of Economic Dynamics and Control* **24**, 1499–1525.
- [2] Barraquand, J. & Pudet, T. (1996). Pricing of American path-dependent contingent claims, *Mathematical Finance* **6**, 17–51.
- [3] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.
- [4] Boyle, P.P., Evnine, J. & Gibbs, S. (1989). Numerical evaluation of multivariate contingent claims, *Review of Financial Studies* **2**, 241–250.
- [5] Boyle, P.P. & Lau, S.H. (1994). Bumping up against the barrier with the binomial method, *Journal of Derivatives* **1**, 6–14.
- [6] Broadie, M. & Detemple, J. (1996). American option valuation: new bounds, approximations and a comparison of existing methods, *The Review of Financial Studies* **9–4**, 1221–1250.
- [7] Chalasani, P., Jha, S., Egriboyun, F. & Varikooty, A. (1999). A refined binomial lattice for pricing American Asian options, *Review of Derivatives Research* **3**, 85–105.

-
- [8] Chalasani, P., Jha, S. & Varikooty, A. (1999). Accurate approximations for European Asian Options, *Journal of Computational Finance* **1**, 11–29.
 - [9] Cheuk, T.H.F. & Vorst, T.C.F. (1996). Complex barrier options, *Journal of Derivatives* **4**, 8–22.
 - [10] Cox, J., Ross, S.A. & Rubinstein, M. (1979). Option pricing: a simplified approach, *Journal of Financial Economics* **7**, 229–264.
 - [11] Derman, E. & Kani, I. (1994). Riding on a smile, *Risk Magazine* **7**, 32–39.
 - [12] Derman, E., Kani, I., Bardhan, D. & Ergener, E.B. (1995). Enhanced numerical methods for options with barriers, *Finance Analyst Journal* **51–6**, 65–74.
 - [13] Diener, F. & Diener, M. (2004). Asymptotics of the price oscillations of a European call option in a tree model, *Mathematical Finance* **14**(2), 271–293.
 - [14] Dupire, B. (1994). Pricing on a smile, *Risk Magazine* **7**, 18–20.
 - [15] Figlewski, S. & Gao, B. (1999). The adaptive mesh model: a new approach to efficient option pricing, *Journal of Financial Economics* **53**, 313–351.
 - [16] Forsyth, P.A., Vetzal, K.R. & Zvan, R. (2002). Convergence of numerical methods for valuing path-dependent options using interpolation, *Review of Derivatives Research* **5**, 273–314.
 - [17] Gaudenzi, M., Lepellere, M.A. & Zanette, A. (2007). *The Singular Points Binomial Method For Pricing American Path-Dependent Options*, Working Paper, Finance Department, University of Udine, Vol. 1, pp. 1–17.
 - [18] Gaudenzi, M. & Pressacco, F. (2003). An efficient binomial method for pricing American put options, *Decisions in Economics and Finance* **4–1**, 1–17.
 - [19] Giles, M.B. & Carter, R. (2006). Convergence analysis of Crank-Nicolson and Rannacher time-marching, *Journal of Computational Finance* **4–9**, 89–112.
 - [20] Hull, J. & White, A. (1993). Efficient procedures for valuing European and American path-dependent options, *Journal of Derivatives* **1**, 21–31.
 - [21] Hull, J. & White, A. (1993). Numerical procedures for implementing term structure models I: single factor models, *Journal of derivatives* **2**, 7–16.
 - [22] Kamrad, B. & Ritchken, P. (1991). Multinomial approximating models for options with k state variables, *Management Science* **37**, 1640–1652.
 - [23] Kushner, H. & Dupuis, P.G. (1992). *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer Verlag.
 - [24] Lamberton, D. (1998). Error estimates for the binomial approximation of American put options, *Annals of Applied Probability* **8**(1), 206–233.
 - [25] Lamberton, D. (2002). Brownian optimal stopping and random walks, *Applied Mathematics and Optimization* **45**, 283–324.
 - [26] Li, A., Ritchken, P. & Sankarasubramanian, L. (1995). Lattice methods for pricing American interest rate claims, *The Journal of Finance* **2**, 719–737.
 - [27] Martini, C. (1999). *Introduction to Tree Methods for Financial Derivatives*. PREMIA software documentation. <http://www.premia.fr>.
 - [28] Pelsser, A. & Vorst, T. (1994). The binomial model and the Greeks, *Journal of Derivatives* **1–3**(3), 45–49.
 - [29] PREMIA. *An Option Pricer Project*. MathFi, INRIA-ENPC. <http://www.premia.fr> (accessed 2008).
 - [30] Rannacher, A. (1984). Finite element solution of diffusion problems with irregular data, *Numerische Mathematik* **43–2**, 309–327.
 - [31] Ritchken, P. (1995). On pricing barrier options, *Journal of Derivatives*, **3**, 19–28.
 - [32] Schonbucher, P.J. (2002). A tree implementation of a credit spread model for credit derivatives, *Journal of Computational Finance* **6–2**, 1–38.
 - [33] Walsh, J.B. (2003). The rate of convergence of the binomial tree scheme, *Finance and Stochastics* **7**, 337–361.

JÉRÔME LELONG & ANTONINO ZANETTE

Quadrature Methods

Beyond the Black–Scholes equations [6] for European style options and certain other cases, generally with no early exercise, “numerical” techniques involving repetitive calculations are required in order to extract approximate solutions, which can be refined by further calculation to make them comparable in accuracy with analytic solutions. The numerical techniques are then classified as trees [18], various finite-difference methods following from the most basic explicit method [8], Monte Carlo simulation [7], and—introduced only recently—QUAD [4], each of which has been subject of modification and refinement.

Quadrature *via* the QUAD method was first presented as a flexible, robust option pricing tool of wide applicability *via* cross-disciplinary work by the research group of Duck (mathematics) and Newton (finance), then at Manchester University [4]. Just as the mathematics of tree, finite-difference and Monte Carlo approaches were known and used in other areas, such as engineering and natural sciences, long before their introduction into finance; basic quadrature goes back centuries and is, in simple conversational terms, the calculation of an area under a graph *via* an approximation. Pictorially, this can be the splitting of an area into a series of shapes, such as rectangles, that approximate the total area and summing their individual areas. Taking smaller shapes produces more accurate results, converging on the correct one. For doing this, the well-known methods are the trapezium rule, Simpson’s method, and Gaussian quadrature, and there are others as well. Each has differing properties and is more or less easy to program, but of particular interest is the rate of convergence to a correct solution as the number of calculations is increased in finer and finer approximations.

A key concept in financial quadrature—sometimes not appreciated—is that the mathematical quadrature component is merely a computational engine to be chosen appropriately to fit into the wider calculations of the particular options problem [4, 5]. Thus, even the very simple trapezium rule can be adequate when elements in wider calculation are less refined. Similarly, Gaussian quadrature only provides useful extra speed over Simpson’s rule in pricing problems where unusually heavy calculational

demands on the “engine” are requisite. Details of quadrature schemes can be found in [1].

It was well known for many years that the majority of options could be written as either a finite or an infinite set of nested (multiple) integrals. Geske and Johnson [20] valued the American put option using a set of multivariate integrals, and Huang *et al.* [25] and Sullivan [33] valued them using univariate integration at discrete observation points followed by Richardson extrapolation. The Geske and Johnson method, however, is computationally very time consuming when more than four nested (i.e., multiple) integrals have to be evaluated. Huang *et al.* used univariate integrals to achieve the same effect, but not more than three observation times were used due to computational complexity. Sullivan [33] also used univariate integrals using a Gaussian quadrature scheme. The first flexible, robust, and widely applicable option pricing tool in quadrature came with QUAD [4]. Its range has been extended to cover the most difficult problems in combination: path dependency, early exercise, and multiple underlyings [5]. In a parallel work, Broadie and Yamamoto [11, 12] have adapted the fast Gauss transform (FGT) technique of Greengard and Strain [21], which, like QUAD, omits time steps between exercise dates, thus vastly reducing the computational load.

As with trees and finite-difference methods, QUAD works backward in time from the maturity of the option, in contrast to Monte Carlo that calculates forward. In the past, this tended to make the backward-working techniques the methods of choice when there could be early (American-style) exercise but favored Monte Carlo both for path-dependent options and for those with several underlyings, since although Monte Carlo starts by requiring a higher computational load than the other techniques (from averaging a large number of simulations), unlike the other techniques it does not suffer from the “curse of dimensionality”, which raises their loads exponentially rather than linearly as dimensions are added. This lack of immediate ability to handle American-style options with Monte Carlo and a desire to tackle problems with both forward and backward elements lead to improvements in all the techniques, with the result that they now compete as alternatives across the spectrum of options types. Since the work of Longstaff and Schwartz [28], in particular, Monte Carlo has been able to handle early exercise (*see Bermudan Options*). In mathematics,

where “high dimension” can mean that a large number of dimensions are of interest, Monte Carlo may become the only technique of concern; however, in finance five dimensions can be considered large. Moreover, quadrature-based methods are so fast to converge that they may only be overtaken in speed by Monte Carlo in dimensions above five (clearly depending on which two quadrature and Monte Carlo techniques are compared; see below for more on this).

QUAD is a particularly effective tool because of the way it deals with the two fundamental causes of error in the numerical methods for finance: distribution error and nonlinearity error. The distribution error arises from approximating a continuous distribution with a discrete one and in other methods can be reduced by adding more steps to a lattice or grid to improve the approximation. This, however, adds to the computational load, whereas with QUAD no calculations are needed at time steps between exercise dates which, combined with the well-known high accuracy of quadrature methods in approximating integrals of functions, results in a huge advantage (measured in orders of magnitude). An extreme example is a plain European option which, if valued numerically not analytically, would require just one set of QUAD calculations in moving from maturity to valuation, with no intermediate steps.

Nonlinearity error is dealt with by the flexibility in placing calculational nodes so as not to miss non-smooth changes (the simplest example being the kink in payoff for a plain European option). Nonlinearity error is observed when valuing options using lattice methods (such as binomial and trinomial trees) or grid methods (notably finite-difference techniques). The familiar saw tooth shape observed in plots of the error with varying time steps when pricing options with binomial trees is directly the result of a nonlinearity in the option price for certain values of the underlying [9, 14, 15, 19, 27, 31, 36]. Extrapolation procedures are possible only when nonlinearity error has been removed. For products such as European vanilla and continuous barrier options, the lattice or the grid can be adapted to remove this nonlinearity error, the main advantage of which is to provide smooth, monotonically converging results that can be improved significantly by standard extrapolation procedures. For options containing more and more exotic features, it becomes increasingly difficult, if not impossible, to adjust the lattice or the grid to

remove new sources of nonlinearity error: for example, options in which the source of nonlinearity moves with time as in barrier or American-style options.

The QUAD method improves on lattice and grid methods in two important ways: superior convergence and, most important, far greater flexibility. The convergence of the method is dependent on which underlying quadrature scheme is used; with Simpson’s rule the convergence is with $1/N^4$, where N is the number of steps, that is, $N = 1/h$, with h being the quadrature mesh size of steps (contrast $1/N^{1/2}$ for the Monte Carlo method). This, combined with its flexibility, makes the improvement offered significant for many types of options. Using the QUAD method with one underlying and either one or two discontinuities at any time, the grid can be maneuvered so as to remove all nonlinearity error, leaving only distribution error and providing monotonic, smooth, fast convergence. These results can then be extrapolated, providing remarkably accurate results rapidly.

One significant problem with lattice and grid methods is what is dubbed the curse of dimensionality; as more underlying assets are introduced, computation time increases exponentially. The usual alternative is to solve by means of Monte Carlo methods, for which the computational effort increases only linearly as more underlying assets are added. For a single underlying asset, Monte Carlo methods are inferior to lattice and grid methods because of slow convergence, but as more underlying assets are considered, Monte Carlo methods can take over as the preferred technique; however, as we shall see, this is not the case with QUAD, which retains its considerable speed/accuracy advantage across the gamut of practical options problems.

To quantify the trade-off between dimensions and convergence, Broadie and Glasserman [10] express convergence as a function of computational effort or work. Under this measure, for an n -underlying asset option pricing problem, lattice or explicit finite-difference methods (*see Finite Difference Methods for Barrier Options*) have convergence rates of $O(\text{work}^{-1/(n+1)})$, the Crank–Nicolson finite-difference method (*see Crank–Nicolson Scheme*) has $O(\text{work}^{-2/(n+1)})$, whereas Monte Carlo methods (*see Monte Carlo Simulation*), because of the linear increase in effort, exhibit $O(\text{work}^{-1/2})$ convergence. Where time stepping is required (e.g., for Bermudan options), under this metric, the QUAD method when using

unextrapolated Simpson's rule has $O(\text{work}^{-4/(n+1)})$ convergence, increasing to $O(\text{work}^{-8/(n+1)})$ when extrapolation is employed. Andricopoulos *et al.* [5] have compared QUAD in a deliberately unfavorable circumstance with [28] Monte Carlo on multiple underlyings for American options (where QUAD is naturally slowed by more exercise times, Bermudan-style, followed by extrapolation) and found that QUAD still has an advantage with five underlyings, thus covering most cases in practical finance.

Extending the Method

The fast convergence of QUAD is partly due to the careful placing of nodes in the calculations. For example, the payoff conditions at an option's maturity show a break in the direction, and nodes are correctly placed so as not to calculate across the break, which would lead to propagation of errors. This does mean that some appreciation of the "geometry" of solution is necessary in each problem in order to attain the highest accuracy. This is, perhaps, not a great difficulty (and careful placement of nodes for improved accuracy is a general feature of all the lattice and grid methods); nevertheless, it is useful that the degradation of results through programming roughly, without care for node placement, can still give adequate results in a quickly programmed exploratory setup [13]. For convenience, a hybrid QUAD finite-difference approach was used by Law [26] for the case of "arithmetic Asian tail" options—a reminder that methods can be mixed.

The foundation work for QUAD was presented in the Black–Scholes framework but this method applies whenever the conditional probability density function is known. This restricts the immediate use of the QUAD technique to the Black–Scholes setup, to Merton's jump-diffusion model (*see* **Jump-diffusion Models**), and to certain interest-rate models such as those of Vasicek [34] and Cox *et al.* [17] (*see* **Term Structure Models**). Extending the technique to Merton's process is straightforward, but the interest-rate models are more subtle and Heap [23] has successfully extended the coverage to interest-rate derivatives including these mean-reverting underlying processes.

A notable advance was made by O'Sullivan [30], who used the observation that many useful processes that do not have a well-known density function

do nonetheless have a well-understood characteristic function. The density function, as the inverse Fourier transform of the characteristic function, can be computed using fast Fourier transform and the output may then be inserted in the QUAD scheme to price derivatives. We refer to this method as *FFT-QUAD*. O'Sullivan's method applies in particular to exponential Levy processes. This made FFT-QUAD an important advance, but it does suffer from drawbacks [22]. First, it requires two integrations even for a derivative on a single underlying process. This brings the complexity of the algorithm to at least $O(N^2)$, where N is the number of grid points used in the numerical integrations; by comparison, the original QUAD had a much better complexity of just $O(N)$ for vanilla options. Second, FFT-QUAD cannot be used to price heavily path-dependent options in stochastic volatility frameworks, since it does not keep track of the evolution of the volatility process in moving from one observation point to the next.

O'Sullivan's FFT-QUAD was considerably improved by the CONV technique of Lord *et al.* [29] (*see also* **Fourier Methods in Options Pricing**). We refer to this method as *CONV-QUAD* [32]. This excellent method uses the observation that the basic pricing integral may usually be regarded as the convolution (strictly speaking, the cross correlation) of the payoff and the density function. The beauty of this insight is that the two integrals of FFT-QUAD may then be replaced by two fast Fourier transforms. This brings the complexity of the algorithm down to $O(N \log(N))$ and, for example, for Bermudan options (on M observation points), the complexity remains at $O(MN \log(N))$, which beats even QUAD's $O(MN^2)$. The CONV method applies in particular to exponential Levy processes (*see* **Exponential Lévy Models**) and hence in particular to CGMY processes (*see* **Tempered Stable Process**). Owing to its nearly linear speed, it is clearly the method of choice for a great many processes. CONV-QUAD, however, suffers from the same difficulty of application to stochastic volatility processes: the volatility process is ignored again. Moreover, there is no possibility of treating, say, Heston's processes [24] (*see also* **Heston Model**) by a two-dimensional CONV-QUAD; the volatility component is simply not sufficiently well behaved.

In summary thus far, we have, in various forms of QUAD, a remarkably accurate (fast-converging) technique that covers complex combinations of option

characteristics in multiple dimensions. It outperforms other methods in most circumstances, making it the method of choice in situations when this power is needed and is a convenient alternative to trees, finite-difference grids, and Monte Carlo among others. There remains, however, some further development underway in order to complete the universality of QUAD: full coverage of underlying processes, particularly stochastic volatility and Heston's framework [24]. This appears possible *via* the basic QUAD scheme, reducing complexity *via* CONV-QUAD techniques when possible. When pricing, say, Bermudan options, complexity will grow linearly, not exponentially, with the number of observation points, that is, with B observation points, the overall complexity will be $O(BM^2N \log(N))$. This complexity is worse than what CONV can produce in other cases, but it is necessary to keep track of an extra process, and some extra penalty is to be expected as CONV-QUAD techniques cannot be used on the volatility component [22].

QUAD Mathematics

We now proceed to a demonstration of the basic technique for QUAD. The original description of the QUAD method can be found in detail in [3] and in overview in [5]. Starting with the well-known [6] partial differential equation for an option with an underlying asset following geometric Brownian motion (*see Black–Scholes Formula*):

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r - D_c)S \frac{\partial V}{\partial S} - rV = 0 \quad (1)$$

where V is the price of the derivative product and is a function of S , the value of the underlying asset and time, t . The risk-free interest rate is r , the volatility of the underlying asset is σ , and the exercise price is X . D_c is a continuous dividend yield. A convenient log transform may be made:

$$x = \log(S) \quad (2)$$

Then suppose that y is the corresponding value of the transform of the underlying at time $t + \Delta t$, where Δt is a time step. It is important to note that Δt is not restricted to small time periods; for example, were QUAD has to be applied to a plain European call option (this is not required since we have the

Black–Scholes analytic solution), then the complete time to expiry would be taken in a single time step, Δt . At expiry, the final condition (payoff) becomes $\max(e^y - X, 0)$. The solution for the value of the option at time t on an underlying asset S is then

$$V(x, t) = A(x) \int_{-\infty}^{\infty} B(x, y) V(y, t + \Delta t) dy \quad (3)$$

where

$$A(x) = \frac{1}{\sqrt{2\sigma^2\pi\Delta t}} e^{-(kx/2) - (\sigma^2 k^2 \Delta t/8) - r\Delta t} \quad (4)$$

$$B(x, y) = e^{-((x-y)^2/(2\sigma^2\Delta t)) + (ky/2)} \quad (5)$$

$$k = \frac{2(r - D)}{\sigma^2} - 1 \quad (6)$$

for a plain vanilla call, for example, the integrand becomes

$$f(x, y) = B(x, y) \times \max(e^y - 1, 0) \quad (7)$$

This integral is the key. Next, any of the many quadrature methods can be employed as a valuation “engine” for what is a European option with known payoff. The integration covers a single time step, Δt . As already noted, this is not of any special use for a plain European option; the (considerable) advantage comes from treating more complex and interesting options problems as equivalent to series of European options. Regions where there is no boundary condition to deal with can be jumped in a single step. Contrast this with other techniques, such as finite difference or trees, in which many intermediate calculations are employed (though there are techniques for thinning the grids/trees). Once more complex option features are incorporated, the choice of quadrature “engine” becomes more important. The speed of convergence comes from a combination of the calculations required for the quadrature “engine” and the calculations arising from particular option features (boundary conditions). A faster engine is needed only in some cases in order to attain the required speed and accuracy. Andricopoulos [3] originally considered the trapezium rule, Simpson's method, and Gaussian quadrature. From the exceedingly simple technique of the trapezium rule, familiar in schoolrooms, through to Gaussian Quadrature, these three are progressively more accurate (fast converging). At the extreme end of fast “engines”, the superior convergence of the Gauss–Legendre method of Sullivan [33] used within

QUAD has been subject to detailed examination by Chung *et al.* [16].

Note that the range of integration is infinite, but this is not a practical problem; the range can be truncated, provided the integrand outside the truncated range is suitably small. Highly accurate calculations are then possible to any level required. A vanilla call option has a discontinuity in the first derivative at expiry, where $S = X$ or $y = \log(X)$. Integration is needed from $\log(X)$ and above, with a frequency of quadrature points and a cutoff to suit the desired accuracy. Extension to options observed discretely M times ($T_1; \dots; T_M$ with the current time T_0 and expiry T_M) is made possible by splitting the option into M separate options, V_m . Each option, V_m , runs from T_{m-1} to T_m . The values at expiry (time T_M) are simply the option payoff. The values of the option at T_{m-1} can be found for all values of y required by the quadrature routine at T_m . American exercise is readily handled by a combination of extra observation times and extrapolation. Any additional operations, such as the imposition of a barrier, can be performed on these values and then these can be used to value the option at time T_{m-2} . The process is continued until the option value is found at the present time, T_0 .

Regarding truncation, for the logarithm of an underlying asset, $\log(S)$, the probability of moving more than five standard deviations in any time period is negligible and truncation to this range typically affects, at worst, only the fifth significant figure of the option price. The probability of moving more than 10 standard deviations is so small that truncating to this range is comparable to machine accuracy on a computer (note that this is important in attaining machine accuracy rather than specifically the choice of quadrature technique supposed in [16]). The choice of range, ξ (the number of standard deviations of the range of integration), is left to the practitioner, but it should be somewhere between these bounds, 5 and 10. The value 7.5 is generally good because it is accurate enough, yet it does not affect extrapolation and does not render the computational effort prohibitively high. For an option valued at time T_m , the range in y , $[y^{\min}; y^{\max}]$ used is

$$y^{\min} = x - \xi \sqrt{T_m - T_{m-1}} \quad (8)$$

and

$$y^{\max} = x + \xi \sqrt{T_m - T_{m-1}} \quad (9)$$

If the QUAD grid is constructed to coincide with the discontinuities such as a strike price, barrier, or exercise boundary, then convergence is perfectly smooth and suitable for improvement through extrapolation because only distribution error remains. It is useful because it is so smooth that extrapolated results can often be further extrapolated themselves. This extrapolation is applicable in all cases, including those with early exercise (but see [5], for adaptive quadrature). Extrapolation is straightforward, *via* a simple Richardson-type procedure. If results converge at a known rate $(\delta y)^d$, then consider two calculations undertaken with step sizes, δy_1 and δy_2 , producing option values V_1 and V_2 , respectively. An improved value V_{ext} is given as

$$V_{\text{ext}} = \frac{(\delta y_1)^d V_2 - (\delta y_2)^d V_1}{(\delta y_1)^d - (\delta y_2)^d} \quad (10)$$

Assuming that the discontinuities are correctly located, the extrapolated trapezium rule results converge as $(\delta y)^4$, which is $1/N^4$. For Simpson's rule, the extrapolated results converge as $(\delta y)^8$, which is $1/N^8$, a remarkable rate of convergence compared with other methods. For comparison, a trinomial tree converges merely with N^{-1} or in some cases only with $N^{-1/2}$. Even the more sophisticated finite-difference methods at best converge at the rate of $(\Delta S^2, \Delta t^2)$, where ΔS and Δt are the step sizes in the S and the t directions, respectively.

Simpson's rule remains one of the most accurate and popular methods for approximating integrals. For a function of y , $f(y)$, plotted against t , y divides the desired range, $[a_1, a_2]$ into n intervals of a fixed length δy such that $n\delta y = a_2 - a_1$. Then approximate the area under the curve by summing the area of the individual regions. This yields the following expression:

$$\begin{aligned} \int_{a_1}^{a_2} f(y) dy \approx & \frac{\delta y}{6} \left\{ f(a_1) + 4f\left(a_1 + \frac{1}{2}\delta y\right) \right. \\ & + 2f(a_1 + \delta y) + 4f\left(a_1 + \frac{3}{2}\delta y\right) \\ & + 2f(a_1 + 2\delta y) + \dots + 2f(a_2 - \delta y) \\ & \left. + 4f\left(a_2 - \frac{1}{2}\delta y\right) + f(a_2) \right\} \quad (11) \end{aligned}$$

It is easily shown that, for smooth functions, the error term associated with this method decreases

at a rate of order $(\delta y)^4$ and so a doubling of the number of steps reduces the error by a factor of 16. Here, it is useful to reiterate the analogy of quadrature techniques as merely “engines” to be fitted into the bodywork of QUAD. When pricing options for which the overall valuation technique puts other, more stringent, limitations on convergence, very basic quadrature may be appropriate. For example, the trapezium rule is an even simpler quadrature method, slower to converge than Simpson’s rule, at a rate of $(\delta y)^2$, but in rare cases its use can save computational time, as in the case of lookback options priced in three dimensions, where the use of more accurate quadrature schemes would be superfluous. QUAD benchmarks may be found in [3, 35].

Speed and Accuracy

For once-only calculations under no research or business time constraint, options valuations are now usually accessible *via* more than one technique; however, there remain some difficult or high-volume repetitive requirements for fast and accurate answers and for these QUAD competes across all option types. For a once-only calculation, it does not probably matter at all whether the chosen technique delivers acceptable accuracy in 10 ms or 10 min for a moderately simple problem, but this difference becomes compounded with the hardest problems and with large numbers of repeat calculations. Thus, speed and convergence results quoting such differences in research papers are of particular interest; a doubling of computer speed, say, since publication, is of less interest than the relative benchmarking. The original QUAD results were computed on a Pentium III 550 MHz system, using the NAG Fortran 95 v4.0 compiler with optimization [4], later on a 2.4 GHz Pentium computer, approximately four times faster [5]. Even given the current faster computers and projected speed increases for the next several years, the practical convergence/accuracy gains for many options problems remain exceptional, generally measurable in several orders of magnitude. Pricing to penny accuracy with a trinomial tree (mathematically similar to an explicit finite difference) takes hours, if not days, for some options (e.g., a lookback barrier) but they are priced correct to four decimal places in seconds with QUAD on the outdated Pentium III. In the valuation of a three-underlying American call option, QUAD

achieves penny accuracy in just over a minute, which is a level of accuracy not achieved by the explicit finite-difference method in over 85 min. For European options on two and three underlyings, QUAD exhibits a huge improvement upon Monte Carlo and finite-difference schemes. For two underlyings, the accuracy after 0.01 s is an order of magnitude better than that with Monte Carlo (the next best method) after 21 s. For three underlyings, QUAD is more accurate after 0.04 s than Monte Carlo is after 35 s. By five underlyings, the results start to become comparable, although QUAD still has the edge in accuracy. The new and popular technique introduced by Figlewski’s group at NYU [2, 19], the adaptive mesh model (AMM), bears comparison. This improves lattice and grid methods by setting up more concentrated areas of lattice in areas of greater importance for errors, such as at payoff and on barriers. Another way of thinking of this is as a heavy pruning dense lattices everywhere except in the areas of the greatest importance. Levels of AMM calculation (not explained here) are expressed as AMM1, AMM2, and so on. As an example, for moving barrier options (on a Pentium III), AMM8 with 80 time steps between observations gives root-mean-squared errors of 0.00034 and takes an average time of 8.65 s, and extrapolated QUAD achieves better accuracy in 0.28 s.

Acknowledgments

I thank Dr Hannu Härkönen for his mathematical insights into FFT-QUAD and CONV-QUAD.

References

- [1] Abramowitz, M. & Stegun, I.A. (1965). *Handbook of Mathematical Functions*, Dover, New York.
- [2] Ahn, D.-G., Figlewski, S. & Gao, B. (1999). Pricing discrete barrier options with an adaptive mesh model, *Journal of Derivatives* **6**, 33–44.
- [3] Andricopoulos, A.D. (2003). *Option Valuation Using Quadrature and other Numerical Methods*. PhD thesis, Manchester University, UK.
- [4] Andricopoulos, A.D., Widdicks, M., Duck, P.W. & Newton, D.P. (2003). Universal option valuation using quadrature methods, *Journal of Financial Economics* **67**, 447–471. See also Corrigendum, (2004). *Journal of Financial Economics*, **73**, 603.
- [5] Andricopoulos, A.D., Widdicks, M., Newton, D.P. & Duck, P.W. (2007). Extending quadrature methods to value multi-asset and complex path dependent options, *Journal of Financial Economics* **83**, 471–499.

- [6] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- [7] Boyle, P.P. (1977). Options: a Monte Carlo approach, *Journal of Financial Economics* **4**, 323–338.
- [8] Brennan, M.J. & Schwartz, E.S. (1977). Convertible bonds: valuation and optimal strategies for call and conversion, *Journal of Finance* **32**, 1699–1715.
- [9] Broadie, M. & Detemple, J. (1996). American option valuation: new bounds, approximations, and a comparison of existing methods, *Review of Financial Studies* **9**, 1211–1250.
- [10] Broadie, M. & Glasserman, P. (2004). A stochastic mesh method for pricing high-dimensional American options, *Journal of Computational Finance* **7**, 35–72.
- [11] Broadie, M. & Yamamoto, Y. (2003). Application of the fast Gauss transform to option pricing, *Management Science* **49**, 1071–1088.
- [12] Broadie, M. & Yamamoto, Y. (2005). A double-exponential fast Gauss transform algorithm for pricing discrete path-dependent options, *Operations Research* **53**, 764–779.
- [13] Broni-Mensah, E.K., Duck, P.W. & Newton, D.P. (2008). *A Simple and Generic Methodology to Suppress Non-linearity Error in Multi-Dimensional Option Pricing*. working paper.
- [14] Cheuk, T.H.F. & Vorst, T.C.F. (1996). Complex barrier options, *Journal of Derivatives* **4**, 8–22.
- [15] Cheuk, T.H.F. & Vorst, T.C.F. (1997). Currency look-back options and observation frequency: a binomial approach, *Journal of International Money and Finance* **16**, 173–187.
- [16] Chung, S.L., Ko, K. & Shackleton, M.B. (2005). *Toward Option Values of Near Machine Precision using Gaussian Quadrature*. working paper.
- [17] Cox, J.C., Ingersoll, J.E. & Ross, S.A. (1985). A theory of the term structure of interest rates, *Econometrica* **53**, 385–408.
- [18] Cox, J.C., Ross, S.A. & Rubinstein, M. (1979). Option pricing: a simplified approach, *Journal of Financial Economics* **7**, 229–264.
- [19] Figlewski, S. & Gao, B. (1999). The adaptive mesh model: a new approach to efficient option pricing, *Journal of Financial Economics* **53**, 313–351.
- [20] Geske, R. & Johnson, H.E. (1984). The American put valued analytically, *Journal of Finance* **39**, 1511–1524.
- [21] Greengard, L. & Strain, J. (1991). The fast Gauss transform, *SIAM Journal on Scientific and Statistical Computing* **12**, 79–94.
- [22] Härkönen, H.J. & Newton, D.P. (2009). *Completing the Universality of Option Valuation using Quadrature Methods*. working paper.
- [23] Heap, J. (2008). *Enhanced Techniques For Complex Interest Rate Derivatives*. Ph.D. thesis, Manchester University, UK.
- [24] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *The Review of Financial Studies* **6**, 327–343.
- [25] Huang, J., Subrahmanyam, M. & Yu, G. (1996). Pricing and hedging American options: a recursive integration approach, *Review of Financial Studies* **9**, 277–300.
- [26] Law, S.H. (2009). *On the Modelling, Design and Valuation of Commodity Derivatives*. Submitted Ph.D. thesis, Manchester University, UK.
- [27] Leisen, D.P.J. & Reimer, M. (1996). Binomial models for option valuation: examining and improving convergence, *Applied Mathematical Finance* **3**, 319–346.
- [28] Longstaff, F.A. & Schwartz, E.S. (2001). Valuing American options by simulation: a simple least-squares approach, *Review of Financial Studies* **14**, 113–147.
- [29] Lord, R., Fang, F., Bervoets, F. & Oosterlee, K. (2007). *A Fast and Accurate FFT-Based Method for Pricing Early-Exercise Options Under Levy Processes*, <http://ssrn.com/abstract=966046>.
- [30] O’Sullivan, C. (2005). *Path Dependent Option Pricing under Levy Processes*, <http://ssrn.com/abstract=673424>.
- [31] Ritchken, P. (1995). On pricing barrier options, *Journal of Derivatives* **3**, 19–28.
- [32] Staunton, M. (2007). Convolution for Lévy, *Wilmott Magazine* September, 62–63.
- [33] Sullivan, M.A. (2000). Valuing American put options using Gaussian quadrature, *Review of Financial Studies* **13**, 75–94.
- [34] Vasicek, O. (1977). An equilibrium characterization of the term structure, *Journal of Financial Economics* **5**, 177–188.
- [35] Widdicks, M. (2002). *Examination Extension and Creation of Methods for Pricing Options with Early Exercise Features*. PhD thesis, Manchester University, UK.
- [36] Widdicks, M., Andricopoulos, A.D., Newton, D.P. & Duck, P.W. (2002). On the enhanced convergence of lattice methods for option pricing, *Journal of Futures Markets* **22**, 315–338.

Related Articles

Bermudan Options; Finite Difference Methods for Barrier Options; Finite Difference Methods for Early Exercise Options; Fourier Methods in Options Pricing; Fourier Transform; Lattice Methods for Path-dependent Options; Sparse Grids.

DAVID P. NEWTON

Partial Integro-differential Equations (PIDEs)

Partial integro-differential equations (PIDEs) appear in finance in the context of option pricing in discontinuous models. They generalize the Black–Scholes partial differential equation (PDE) when the continuous Black–Scholes dynamics for the underlying price is replaced by a Markov process with jumps.

The jump models proposed for option pricing are mostly based on Lévy processes (see **Exponential Lévy Models**). The jumps of a Lévy process are characterized by a positive measure $\nu(dx)$ on $\mathbb{R} \setminus \{0\}$, called *Lévy measure*, which satisfies the following integrability condition:

$$\int_{\mathbb{R} \setminus \{0\}} \min(1, x^2) \nu(dx) < \infty \quad (1)$$

In particular, it may be singular at the origin.

Consider an asset whose risk-neutral dynamics are described by an exponential Lévy model $S_t = S_0 \exp X_t$, where X_t is a Lévy process with coefficient σ and Lévy measure ν (the drift is determined by the martingale condition on $e^{-rt} S_t$). The value $V_t = V(t, S_t)$ at time $t \leq T$ of a European option on this underlying with maturity T and payoff $H(S_T)$ is then given by the solution of the following PIDE:

$$\begin{aligned} V_t + \frac{1}{2} \sigma^2 S^2 V_{SS} + r S V_S - r V \\ + \int [V(t, S e^y) - V(t, S) \\ - S(e^y - 1) V_S(t, S)] \nu(dy) = 0 \end{aligned} \quad (2)$$

The subscripts indicate partial derivatives and r is the continuously compounded risk-free interest rate. We can see that this equation contains the same terms as the Black–Scholes PDE and an integral term with respect to the Lévy measure ν . In particular, if ν is equal to 0, the integral term disappears and equation (2) reduces to the Black–Scholes equation. This is compatible with the fact that an exponential Lévy model without jumps is simply the Black–Scholes model.

Similar to the diffusion case, the PIDE formulation can be used for other types of options, such as barrier or American options, leading to

boundary-value problems or variational inequalities with the same integro-differential operator. The PIDE approach can also be generalized to more complicated cases: for example, stochastic volatility with jumps, nonhomogeneous jump measure $\nu_{t,x}(dx)$, multiasset derivatives (see **Jump Processes; Exponential Lévy Models**). Finally, one can obtain a *forward PIDE*—in maturity and strike variables—for call options, analogous to the Dupire equation (see [6, 14]) in one-factor Markovian models with jumps.

We first present the specific features of the integro-differential equations compared with the PDEs. In particular, we highlight the difficulties that arise for the numerical solution of these equations. We see that the use of standard techniques, such as finite differences or finite elements, is not straightforward. We then survey the existing approaches to adapt numerical methods developed for PDEs to the solution of partial integro-differential equations.

General Remarks

To solve the PIDE (2), it is convenient to make the change of variables $x = \log(S/S_0)$ and $\tau = T - t$. Denote $v(\tau, x) = e^{r(T-t)} V(t, S) = e^{r\tau} V(T - \tau, S_0 e^x)$. This leads to the following equation:

$$\begin{aligned} v_\tau = \frac{1}{2} \sigma^2 v_{xx} + \left(r - \frac{1}{2} \sigma^2 \right) v_x \\ + \int [v(\tau, x + y) - v(\tau, x) \\ - (e^y - 1) v_x(\tau, x)] \nu(dy) \end{aligned} \quad (3)$$

The payoff function provides a terminal condition for (2): $V(T, S) = H(S)$ and an initial condition for equation (3):

$$v(0, x) = H(S_0 e^x) \equiv h(x), \quad x \in \mathbb{R} \quad (4)$$

The main sources of difficulties with the PIDEs of type (2) or (3) are the following:

- nonlocal character of the operator;
- possible lack of regularity of the solution, especially in the pure jump case ($\sigma = 0$); and
- possible singularity of the Lévy measure at 0.

Let us get a closer look at these issues. Observe that the operator in equation (3) is nonlocal in space: that is, to evaluate it at a given point (τ, x_*) , we need

the values $v(\tau, x)$ for all $x \in \mathbb{R}$. This has multiple implications. First of all, if we consider equation (3) on a bounded domain in x , which is always the case if we solve it numerically, the boundary conditions must be specified not only *at* the boundary but also *beyond* the boundary:

$$v(\tau, x) = g(\tau, x) \quad \forall x \notin (x_{\min}, x_{\max}) \quad (5)$$

where g is a given function. Second, a finite difference or finite element discretization of equation (3) gives rise to a dense matrix, in contrast with the PDE case where the discretized operator is a sparse (usually tridiagonal) matrix easy to invert. The direct solution of a nonsparse linear system of N equations requires, in general, $O(N^3)$ operations, which makes the direct application of these methods inefficient. Finally, the integral operator propagates possible irregularities of the solution. For instance, the price of a barrier option is usually discontinuous at the barrier. While in the PDE case this discontinuity influences only the neighborhood of the boundary, a PIDE will propagate it inside the domain. It is worth mentioning that the regularity of the solution is not only a theoretical issue but also has a direct impact on the efficiency (stability and order of convergence) of numerical methods.

If the volatility σ is strictly positive (jump diffusion case), the integro-differential operator has the same regularizing property as the differential Black–Scholes operator: the solution is smooth for all $t < T$ even for discontinuous payoffs. The same is true in the pure jump case (without diffusion component) if there are “sufficiently many” jumps. More precisely, we have the following result [15, 16]:

Proposition 1 *Let H be a measurable function with a finite set of discontinuities. Suppose that $h(x) \equiv H(S_0 e^x)$ has at most polynomial growth: $\exists p \geq 0, |h(x)| \leq C(1 + |x|^p)$. Let X be a Lévy process with characteristic triplet (σ, ν, γ) satisfying $\int_{|y|>1} |y|^{2p} \nu(dy) < \infty$ and*

$$\begin{aligned} &\sigma > 0 \quad \text{or} \quad \exists \beta \in (0, 2), \\ &\liminf_{\varepsilon \downarrow 0} \frac{1}{\varepsilon^{2-\beta}} \int_{-\varepsilon}^{\varepsilon} |y|^2 \nu(dy) > 0 \end{aligned} \quad (6)$$

Then the forward European option price $v(\tau, x) = E[h(x + X_\tau)]$ belongs to the class $C^\infty((0, T] \times \mathbb{R})$ with $|\partial^{n+m} v / \partial x^n \partial t^m(x)| \leq C(1 + |x|^p)$, for all

$n, m \geq 0$. Moreover, $v(\tau, x)$ is a solution of the PIDE (3) with the initial condition (4).

The main idea of the proof is to apply the Itô formula to the discounted option price, which is a martingale, identify the drift term, and put it to 0: this gives the corresponding PIDE (see also **Partial Differential Equations**). The condition (6) ensures the regularity of the option price, so that we can apply the Itô formula. While the condition (6) is satisfied by most of the exponential Lévy models in the literature, there are some exceptions such as the popular Variance Gamma model (see **Variance-gamma Model**). When equation (6) is not satisfied, the irregularities in the payoff are not smoothened immediately and the option price may not be sufficiently regular to apply the Itô formula (see the examples in [16]). In the case of barrier options, the option price may be irregular even if $\sigma > 0$. This is due to the fact that the irregularity at the barrier is propagated inside the domain by the integral operator.

The possible lack of regularity of option prices led to consider them as different kinds of weak solutions of the PIDEs. Existence and uniqueness of solutions of integro-differential equations in Sobolev spaces has been studied in [10] and more recently in [25, 26, 33]. This approach provides the necessary framework for the solution of PIDEs by finite element methods. Another type of weak solution is the viscosity solution (see **Monotone Schemes**). Viscosity solutions of PIDEs are studied in [4, 5, 7, 9, 23, 27–29]. This approach is naturally linked to the probabilistic interpretation of the solution [16]. Moreover, it can be shown that any monotone scheme converges to the viscosity solution of the equation (see **Monotone Schemes**). This property is exploited in [11, 17].

We stress that Proposition 1 gives the PIDE satisfied by a *European* option price in the relatively simple *exponential Lévy models*. Although a rigorous proof of similar results for more complicated jump models or option types is often not available, the formal derivation of the pricing PIDEs is very easy and is regularly used for numerical valuation of option prices in models with jumps. However, we should be aware that, in many cases, the PIDE approach remains to be justified.

Finally, let us comment on the singularity of the Lévy measure. In pure jump models, such as Variance Gamma or CGMY (see **Tempered Stable Process**), the jumps are of infinite activity in order to

compensate the absence of diffusion component. This implies that the jump measure ν is not integrable at 0, which brings an additional difficulty for the discretization of equation (3). For finite element discretization, this is not really a problem because we do not need to discretize ν separately but only to evaluate the integral on some basis functions. To solve the PIDE using finite differences, we can approximate small jumps of the Lévy process by a Brownian motion with appropriately chosen volatility. This approximation is based on the results obtained in [8], and its impact on the option price is estimated in [17]. In the case of jumps of finite variation, as in the Variance Gamma model, it is also possible to replace small jumps by a drift rather than a diffusion term, as done in [24].

Numerical Solution

To illustrate the issues discussed above, we describe a simple finite difference scheme for equation (3) proposed in [17]. We consider here the case of nonzero diffusion and finite jump measure. As noted above, the general case may be reduced to this one by approximating small jumps.

European Vanilla and Barrier Options

To solve numerically the integro-differential problem (3)–(4), we first need to localize it to a bounded computational domain in x and truncate the integration domain in the integral part.

Truncation of the integration domain. Since we cannot calculate numerically an integral on the infinite range $(-\infty, \infty)$, the domain is truncated to a bounded interval (B_l, B_r) . In terms of the Lévy process, this corresponds to removing large jumps. Usually, the tails of ν decrease exponentially, so the probability of large jumps is small and the impact on the solution of the truncation is exponentially small in the domain size (see [17] for pointwise estimates).

Localization. Similarly, for the computational purposes, the domain of definition of the equation has to be bounded. For barrier options, the barriers are the natural limits for this domain and the rebate is the natural boundary condition. In the absence of barriers, we have to choose artificial bounds $(x_{\min}, x_{\max})$ and impose artificial boundary conditions. Recall that “boundary” conditions in this case must extend the

solution beyond the bounds as well (cf. equation (5)). It can be shown that any bounded function $g(\tau, x)$ leads to an exponentially decreasing localization error when the size of the domain increases. However, from the numerical point of view, it is better to take into account the asymptotic behavior of the solution. For instance, a good choice for a put option is $g(\tau, x) = (K - S_0 e^{x+r\tau})^+$. Of course, in the case of one barrier option, we need this condition only on one side of the domain: the other is zero or given by the rebate.

Remark In [25, 26], the authors subtract the payoff from the option price and solve a PIDE with zero boundary conditions on the excess to payoff. This leads to a source term in the equation which is, in the case of a put option, a Dirac delta-function. This is easily handled in the finite element framework but makes problem for finite difference discretization.

Discretization. We consider now the localized problem on $(x_{\min}, x_{\max})$:

$$v_\tau = Lv \quad \text{on } (0, T] \times (x_{\min}, x_{\max}) \quad (7)$$

$$v(0, x) = h(x), \quad x \in (x_{\min}, x_{\max}) \quad (8)$$

$$v(\tau, x) = g(\tau, x), \quad x \notin (x_{\min}, x_{\max}) \quad (9)$$

where L is the following integro-differential operator:

$$Lv = \frac{1}{2}\sigma^2 v_{xx} - \left(\frac{1}{2}\sigma^2 - r\right)v_x + \int_{B_l}^{B_r} \nu(dy)v(\tau, x+y) - \lambda v - \alpha v_x \quad (10)$$

with $\lambda = \int_{B_l}^{B_r} \nu(dy)$, $\alpha = \int_{B_l}^{B_r} (e^y - 1)\nu(dy)$. We introduce a uniform grid on $[0, T] \times \mathbb{R}$:

$$\begin{aligned} \tau_n &= n\Delta t, \quad n = 0, \dots, M, \\ x_i &= x_{\min} + i\Delta x, \quad i \in \mathbb{Z} \end{aligned} \quad (11)$$

with $\Delta t = T/M$, $\Delta x = (x_{\max} - x_{\min})/N$. The values of v on this grid are denoted by $\{v_i^n\}$. The space derivatives of v are approximated by finite differences:

$$(v_{xx})_i \approx \frac{v_{i+1} - 2v_i + v_{i-1}}{(\Delta x)^2} \quad (12)$$

$$(v_x)_i \approx \frac{v_{i+1} - v_i}{\Delta x} \quad \text{or} \quad (v_x)_i \approx \frac{v_i - v_{i-1}}{\Delta x} \quad (13)$$

The choice of the approximation of the first-order derivative—forward or backward difference—depends on the parameters σ , r , and α .

To approximate the integral term, we use the trapezoidal rule with the same discretization step Δx . Choose integers K_l , K_r such that $[B_l, B_r]$ is contained in $[(K_l - 1/2)\Delta x, (K_r + 1/2)\Delta x]$. Then,

$$\int_{B_l}^{B_r} v(dy)v(\tau, x_i + y) \approx \sum_{j=K_l}^{K_r} v_j v_{i+j} \quad \text{where} \quad v_j = \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} v(dy) \quad (14)$$

Using equations (12)–(14), we obtain an approximation for $Lv \approx D_\Delta v + J_\Delta v$, where $D_\Delta v$ and $J_\Delta v$ are chosen as follows:

Explicit–implicit Scheme. Without loss of generality, suppose that $\sigma^2/2 - r < 0$. Then

$$(D_\Delta v)_i = \frac{\sigma^2}{2} \frac{v_{i+1} - 2v_i + v_{i-1}}{(\Delta x)^2} - \left(\frac{\sigma^2}{2} - r \right) \frac{v_{i+1} - v_i}{\Delta x}$$

If $\sigma^2/2 - r > 0$, to ensure the stability of the algorithm, we must change the discretization of v_x by choosing the backward difference instead of the forward one. Similarly, if $\alpha < 0$ we discretize J as follows:

$$(J_\Delta v)_i = \sum_{j=K_l}^{K_r} v_j v_{i+j} - \lambda v_i - \alpha \frac{v_{i+1} - v_i}{\Delta x} \quad (15)$$

Otherwise, we change the approximation of the first derivative. Finally, we replace the problem (7)–(9) with the following semiimplicit scheme:

Initialization:

$$v_i^0 = h(x_i) \quad \text{if } i \in \{0, \dots, N-1\} \quad (16)$$

$$v_i^0 = g(0, x_i) \quad \text{otherwise} \quad (17)$$

For $n = 0, \dots, M-1$:

$$\frac{v_i^{n+1} - v_i^n}{\Delta t} = (D_\Delta v^{n+1})_i + (J_\Delta v^n)_i \quad \text{if } i \in \{0, \dots, N-1\} \quad (18)$$

$$v_i^{n+1} = g((n+1)\Delta t, x_i) \quad \text{otherwise} \quad (19)$$

Here, the nonlocal operator J is treated explicitly to avoid the inversion of the fully populated matrix J_Δ , while the differential part D is treated implicitly. At each time step, we first evaluate vector $J_\Delta v^n$ where v^n is known from the previous iteration, and then solve the tridiagonal system (18) for $v^{n+1} = (v_0^{n+1}, \dots, v_{N-1}^{n+1})$.

Remark A direct computation of the term $J_\Delta v^n$ would require $O(N^2)$ operations and would be the most expensive step of the method. Fortunately, the particular form of the sum—discrete convolution of two vectors—allows to compute it efficiently and simultaneously for all i using fast Fourier transform (FFT) (see [21]). Note, however, that this is only possible in the case of a translation invariant jump measure. Another way to solve the problem of the dense matrix, valid also for nonhomogeneous jump models, is proposed in [26]. The authors use a finite element method with a special basis of wavelet functions (see also **Wavelet Galerkin Method**). In this basis, most of the entries in the matrix operator are very small, so that they can be replaced by zeros without affecting the solution much. The efficiency of the wavelet compression depends on the degree of singularity of the jump measure. Although it does not appear explicitly, this is also the case for the finite difference scheme, because of the trade-off that has to be done between the parameter ε of truncation of small jumps and discretization step Δx (see [17]). Finally, let us mention [31] where the integral is evaluated in linear complexity using a recursive formula in the particular case of the Kou model (see **Kou Model**).

Stability. The scheme (16)–(19) is stable if

$$\Delta t < \frac{\Delta x}{|\alpha| + \lambda \Delta x} \quad (20)$$

It is possible to make it unconditionally stable by moving the local terms in equation (15) in the implicit part. However, numerical experiments show that this leads to large errors for infinite activity Lévy measures.

Remark

- In [21, 26, 30–32], fully implicit or Crank–Nicolson finite difference schemes are used, which are unconditionally stable. To solve the resulting dense linear systems, the authors use

iterative methods that require only matrix-vector multiplication performed using FFT.

- Equation (2) can be solved directly in the original variables. We can also choose a nonuniform grid with, for example, more nodes near the strike and maturity [2, 31]. In this case, an interpolation at each time step is needed in order to apply FFT [21].
- Similar semiimplicit scheme in the finite activity case is used in [11]. In [6], the operator is also split into differential and integral parts, and then an alternating direction implicit (ADI) time stepping is used.
- The situation is more challenging for PIDEs in more than one dimension. The main idea here is to devise an operator-splitting scheme as above where each component only acts on one of the variables, thus reducing the dimension in each step of the computation [12, 13, 18].

American Options

Pricing American options in jump models leads to integro-differential variational inequalities [10, 27]. Equivalently, option price may be represented as solution of a linear complementarity problem (LCP) of the following form (see also **Finite Difference Methods for Early Exercise Options**):

$$V_\tau - \mathcal{L}V \geq 0 \quad (21)$$

$$V - V^* \geq 0 \quad (22)$$

$$(V_\tau - \mathcal{L}V)(V - V^*) = 0 \quad (23)$$

For example, in exponential Lévy models, $\mathcal{L}$ is the same integro-differential operator as in the PIDE (2) and V^* is the payoff received upon exercise. Pricing American options in Lévy driven models based on equations (21)–(23) is considered in [1–3, 19–22, 24, 25, 32, 33]. Numerical solution of the integro-differential problem (21)–(23) faces globally the same difficulties as that of PIDEs. The dense and nonsymmetric matrix of the operator makes unfeasible or inefficient standard methods for solving LCPs. The solutions proposed rely on similar ideas as in the European case: splitting the operator [1, 33], wavelet compression [25], using iterative methods with suitable preconditioning. In [19, 21, 32], the LCP is replaced by an equation with a nonlinear penalty term. We refer to the references cited above for the details on these methods.

References

- [1] Almendral, A. (2005). Numerical valuation of American options under the CGMY process, in *Exotic Options Pricing and Advanced Levy Models*, A. Kyprianou, W. Schoutens & P. Wilmott, eds, Wiley.
- [2] Almendral, A. & Oosterlee, C.W. (2006). Highly accurate evaluation of European and American options under the variance gamma process, *Journal of Computational Finance* **10**(1), 21–42.
- [3] Almendral, A. & Oosterlee, C.W. (2007). Accurate evaluation of European and American options under the CGMY Process, *SIAM Journal on Scientific Computing* **29**, 93–117.
- [4] Alvarez, O. & Tourin, A. (1996). Viscosity solutions of non-linear integro-differential equations, *Annales de l'Institut Henri Poincaré* **13**(3), 293–317.
- [5] Amadori, A.L. (2003). Nonlinear integro-differential evolution problems arising in option pricing: a viscosity solutions approach, *Journal of Differential Integral Equations* **16**(7), 787–811.
- [6] Andersen, L. & Andreasen, J. (2000). Jump-diffusion models: volatility smile fitting and numerical methods for pricing, *Review of Derivatives Research* **4**(3), 231–262.
- [7] Arisawa, M. (2006). A new definition of viscosity solutions for a class of second-order degenerate elliptic integro-differential equations, *Annales de l'Institut Henri Poincaré (C) Non Linear Analysis*, **23**(5), 695–711.
- [8] Asmussen, S. & Rosiński, J. (2001). Approximations of small jumps of Lévy processes with a view towards simulation, *Journal of Applied Probability* **38**(2), 482–493.
- [9] Barles, G., Buckdahn, R. & Pardoux, E. (1997). Backward stochastic differential equations and integral-partial differential equations, *Stochastics and Stochastic Reports* **60**, 57–83.
- [10] Bensoussan, A. & Lions, J.-L. (1982). *Contrôle Impulsionnel et Inéquations Quasi-Variationnelles*, Dunod, Paris.
- [11] Briani, M., Natalini, R. & Russo, G. (2007). Implicit-explicit numerical schemes for jump–diffusion processes, *Calcolo* **44**, 33–57.
- [12] Carr, P. & Itkin, A. (2007). *A Finite-Difference Approach to the Pricing of Barrier Options in Stochastic Skew Model*, Working Paper.
- [13] Clift, S. & Forsyth, P. (2008). Numerical solution of two asset jump diffusion models for option valuation, *Applied Numerical Mathematics*, **58**(6), 743–782.
- [14] Cont, R. & Tankov, P. (2008). *Financial Modelling with Jump Processes*, 2nd Edition, CRC Press.
- [15] Cont, R., Tankov, P. & Voltchkova, E. (2005). Hedging with options in models with jumps, in *Stochastic Analysis and Applications: The Abel Symposium 2005 in honor of Kiyosi Ito*, F.E. Benth, G. Di Nunno, T. Lindstrom, B. Oksendal & T. Zhang, eds, Springer, pp. 197–218.

- [16] Cont, R. & Voltchkova, E. (2005). Integro-differential equations for option prices in exponential Lévy models, *Finance and Stochastics* **9**, 299–325.
- [17] Cont, R. & Voltchkova, E. (2005). Finite difference methods for option pricing in jump-diffusion and exponential Lévy models, *SIAM Journal on Numerical Analysis* **43**(4), 1596–1626.
- [18] Farkas, W., Reich, N. & Schwab, C. (2006). *Anisotropic Stable Lévy Copula Processes—Analytical and Numerical Aspects*. Research Report No. 2006-08, Seminar for Applied Mathematics, ETH Zürich.
- [19] d’Halluin, Y., Forsyth, P. & Labahn, G. (2004). A penalty method for American options with jump diffusion processes, *Numerische Mathematik* **97**, 321–352.
- [20] d’Halluin, Y., Forsyth, P. & Labahn, G. (2005). A semi-Lagrangian approach for American Asian options under jump diffusion, *SIAM Journal on Scientific Computing* **27**, 315–345.
- [21] d’Halluin, Y., Forsyth, P. & Vetzal, K. (2005). Robust numerical methods for contingent claims under jump diffusion processes, *IMA Journal on Numerical Analysis* **25**, 87–112.
- [22] Hirta, A. & Madan, D.B. (2003). Pricing American options under variance gamma, *Journal of Computational Finance* **7**(2), 63–80.
- [23] Jakobsen, E.R. & Karlsen, K.H. (2006). A “maximum principle for semicontinuous functions” applicable to integro-partial differential equations, *Nonlinear Differential Equations Applications* **13**, 137–165.
- [24] Levendorskii, S., Kudryavtsev, O. & Zherder, V. (2005). The relative efficiency of numerical methods for pricing American options under Lévy processes, *Journal of Computational Finance* **9**(2), 69–97.
- [25] Matache, A.-M., Nitsche, P.-A. & Schwab, C. (2005). Wavelet Galerkin pricing of American options on Lévy-driven assets, *Quantitative Finance* **5**(4), 403–424.
- [26] Matache, A.-M., von Petersdorff, T. & Schwab, C. (2004). Fast deterministic pricing of options on Lévy driven assets, *Mathematical Modelling and Numerical Analysis* **38**(1), 37–71.
- [27] Pham, H. (1998). Optimal stopping of controlled jump-diffusion processes: a viscosity solution approach, *Journal of Mathematical Systems* **8**(1), 1–27.
- [28] Sayah, A. (1991). Equations d’Hamilton Jacobi du premier ordre avec termes integro-differentiels, *Communications in Partial Differential Equations* **16**, 1057–1093.
- [29] Soner, H.M. (1986). Optimal control of jump-Markov processes and viscosity solutions, *IMA Volumes in Mathematics and Applications*, Springer-Verlag, New York, Vol. 10, 501–511.
- [30] Tavella, D. & Randall, C. (2000). *Pricing Financial Instruments: The Finite Difference Method*, Wiley, New York.
- [31] Toivanen, J. (2008). Numerical valuation of European and American options under Kou’s jump-diffusion model, *SIAM Journal on Scientific Computing* **30**(4), 1949–1970.
- [32] Wang, I.R., Wan, J.W.L. & Forsyth, P. (2007). Robust numerical valuation of European and American options under the CGMY process, *Journal of Computational Finance* **10**, 31–69.
- [33] Zhang, X.L. (1997). Numerical analysis of American option pricing in a jump-diffusion model, *Mathematics of Operations Research* **22**(3), 668–690.

Related Articles

Backward Stochastic Differential Equations; Numerical Methods; Econometrics of Option Pricing; Exponential Lévy Models; Finite Difference Methods for Barrier Options; Finite Difference Methods for Early Exercise Options; Jump Processes; Lattice Methods for Path-dependent Options; Monotone Schemes; Option Pricing Theory: Historical Perspectives; Partial Differential Equations; Stochastic Volatility Models; Time-changed Lévy Process.

EKATERINA VOLTCHKOVA

Method of Lines

The complexity of a partial differential equation (PDE) grows with the number of independent variables. In applied mathematics, the *method of lines* (MoL) is frequently used to reduce the number of independent variables by one. For this reduction, one has to pay a price, which is twofold:

1. A *system* of several dependent variables arises.
2. The system is an approximation—that is, a discretization error is introduced.

In finance, the method of lines is applied to the Black–Scholes PDE, which involves two independent variables, namely, time t and price S of the underlying asset. The continuous domain is the *half strip* $0 \leq t \leq T$, $S > 0$, where T denotes the time to maturity. Introducing the backward running time $\tau := T - t$, the Black–Scholes equation is

$$-\frac{\partial V(S, \tau)}{\partial \tau} + \mathcal{L}_{\text{BS}}(V(S, \tau)) = 0 \quad \text{with} \quad (1)$$

$$\mathcal{L}_{\text{BS}}(V(S, \tau)) := \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + (r - \delta)S \frac{\partial V}{\partial S} - rV$$

Here, the dependent variable $V(S, \tau)$ denotes the value function of a vanilla American put option, r is the risk-free interest rate, δ denotes a constant dividend yield, and σ the volatility (*see Options: Basic Definitions*). According to the Black–Scholes model, r, σ, δ are taken as constants. For a put option of the American style with strike K , the payoff at time T is

$$(K - S)^+ := \max(K - S, 0) \quad (2)$$

The initially unknown early exercise curve S_f separates the half strip into two parts: stopping region $S \leq S_f$, where V equals the payoff, and the continuation region, where V solves the Black–Scholes equation:

$$V = (K - S)^+ \quad \text{for } S \leq S_f \quad (3)$$

$$V \text{ solves equation (1) for } S > S_f$$

The geometry of the domain is illustrated in Figure 1.

Since equation (1) depends on two independent variables, MoL leads to a system with only one independent variable—that is, to a system of ordinary differential equations (ODEs). For introducing “lines” in our context, we have two possibilities: either we set up lines parallel to the t -axis or we work with lines parallel to the S -axis. The former approach leads to a fully numerical method; it is described in Exercise 4.10 of [6]. Our focus is on the latter MoL approach, with lines parallel to the S -axis. This approach (which goes back to [3]) is attractive because the resulting ODEs can be solved analytically.

Semidiscretization

The method of lines replaces the half strip by a set of equidistant lines, each line defined by a constant value of τ . To this end, the interval $0 \leq \tau \leq T$ is discretized into n subintervals by

$$\tau_v := v\Delta\tau, \quad \Delta\tau := T/n, \quad v = 1, \dots, n-1 \quad (4)$$

(Figure 1). On this discrete set of lines, the partial derivative $\partial V / \partial \tau$ is approximated by the difference quotient

$$\frac{V(S, \tau) - V(S, \tau - \Delta\tau)}{\Delta\tau} \quad (5)$$

This gives a semidiscretized version of equation (1), namely, for each τ_v the ODE

$$w(S, \tau - \Delta\tau) - w(S, \tau) + \Delta\tau \mathcal{L}_{\text{BS}}(w(S, \tau)) = 0 \quad (6)$$

which holds for $S \geq S_f$. Here, we use the notation w rather than V to indicate that a discretization error is involved. As this semidiscretized version of equation (6) is applied for each of the parallel lines, $\tau = \tau_v$, $v = 1, \dots, n-1$, a coupled system of ODEs is derived.

Analytic Solution

Substituting equation (1) into equation (6) gives the equation to be solved for each line τ_v

$$\frac{1}{2}\Delta\tau \sigma^2 S^2 \frac{\partial^2 w}{\partial S^2} + \Delta\tau (r - \delta)S \frac{\partial w}{\partial S} - (1 + r\Delta\tau)w = -w(S, \tau_{v-1}) \quad (7)$$

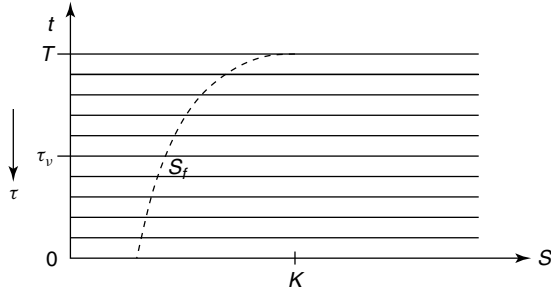


Figure 1 Method of lines applied to an American-style vanilla put

where the argument of w on the left-hand side is (S, τ_v) . This is a second-order ODE for $w(S, \tau_v)$, with boundary conditions for $S_f(\tau_v)$ and $S \rightarrow \infty$. For any line $\tau = \tau_v$, the function $w(S, \tau_{v-1})$ of the right-hand side is known from the previous line, starting from the known payoff for $\tau = 0$. The right-hand function $q(S) := -w(S, \tau_{v-1})$ is an inhomogeneous term of the ODE.

The analytic solution exploits that the equation (7) is linear in w and of the simple type of an inhomogeneous Euler equation

$$\begin{aligned} \alpha S^2 w'' + \beta S w' + \gamma w &= q(S) \\ \text{with } \alpha &= \frac{1}{2} \Delta \tau \sigma^2, \quad \beta = (r - \delta) \Delta \tau, \quad (8) \\ \gamma &= -(1 + r \Delta \tau) \end{aligned}$$

where the prime denotes differentiation with respect to S . The solution method for such a simple type of ODE is standard and found in any ODE text book. It is based on substituting S^λ into the homogeneous ODE ($q \equiv 0$). This yields a quadratic equation with

zeros

$$\begin{aligned} \lambda_{1,2} &:= \frac{1}{2} - \frac{r - \delta}{\sigma^2} \\ &\pm \sqrt{\left(\frac{1}{2} - \frac{r - \delta}{\sigma^2}\right)^2 + \frac{2(1 + r \Delta \tau)}{\sigma^2 \Delta \tau}} \quad (9) \end{aligned}$$

Solutions to the homogeneous ODE are obtained by linear combinations of the S^λ ,

$$aS^{\lambda_1} + bS^{\lambda_2} \quad (10)$$

for suitable constants a and b . A solution of the inhomogeneous equation is added. Note that this analytical solution avoids a truncation error in S -direction.

Matching Solution Parts

In our context, we need (at least) two such solutions for every line, because the inhomogeneous terms change. The early exercise curve S_f separates each of the parallel lines into two parts (Figure 2). As for the previous line τ_{v-1} , the separation point lies more “on the right” (recall that for a put the curve $S_f(\tau)$ is monotonically decreasing for growing τ), the inhomogeneous term $w(\cdot, \tau_{v-1})$ consists of (at least) two parts as well, but separated differently. Neglecting for a moment the previous history of lines $\tau_{v-2}, \tau_{v-3}, \dots$, the analytic solution of equation (7) for τ_v consists of three parts, defined on the three subintervals

$$\begin{aligned} \text{A: } & 0 < S \leq S_f(\tau_v) \\ \text{B: } & S_f(\tau_v) < S \leq S_f(\tau_{v-1}) \\ \text{C: } & S_f(\tau_{v-1}) < S \end{aligned} \quad (11)$$

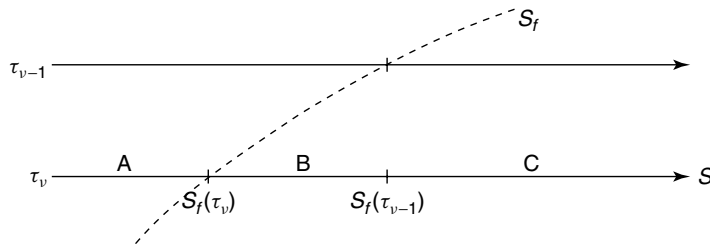


Figure 2 Detail of Figure 1, situation along line τ_v : A: solution is given by payoff; B: inhomogeneous term of differential equation given by payoff; and C: inhomogeneous term given by $-w(\cdot, \tau_{v-1})$

On the left-hand interval A, w equals the payoff; nothing needs to be calculated. For the middle interval B, the inhomogeneous term $-w(., \tau_{v-1})$ is given by the payoff, $q(S) = -(K - S)$. As we need solutions for both subintervals B and C, a second pair of constants is required for C. In this subinterval, the value of a put vanishes for $S \rightarrow \infty$, which contradicts $\lambda_1 > 0$; the root λ_1 drops out for the solution part in C. Hence, this solution is of the type cS^{λ_2} for a constant c . When we consider the dependence on previous lines, then we realize that there are recursively several B-type subintervals, and only one C-type interval for $S > S_f(\tau_0) = K$. To illustrate this, merge Figures 1 and 2. For $S_f(\tau_0) = K$ to hold, assume in addition $\delta \leq r$ for a put ($\delta \geq r$ for a call).

First Line

Let us discuss for the first line $v = 1$ how the solution is setup. For this exposition, we take specifically $\delta = 0$. For $v = 1$, we have $S_f(\tau_{v-1}) = K$ and $q = w(S, \tau_{v-1}) = 0$ for $S > K$. Hence, in subinterval C the inhomogeneous solution is 0. And $q = S - K$ for $S_f(\tau_1) < S < K$, hence

$$\frac{K}{1 + r \Delta \tau} - S \quad (12)$$

solves the inhomogeneous equation for subinterval B. This leads to the three parts of the solution along the first line $\tau = \tau_1$

$$\begin{aligned} w(S; \tau_1) &= K - S \quad \text{for A} \\ w(S; \tau_1) &= \frac{K}{1 + r \Delta \tau} - S + a^{(1)} S^{\lambda_1} \\ &\quad + b^{(1)} S^{\lambda_2} \quad \text{for B} \end{aligned} \quad (13)$$

$$w(S; \tau_1) = c^{(1)} S^{\lambda_2} \quad \text{for C}$$

The value of $S_f(\tau_1)$ is still undetermined as well as the three constants $a^{(1)}$, $b^{(1)}$ and $c^{(1)}$. To determine these four parameters, we require four equations.

The unknown separation point $S_f(\tau_v)$ is fixed by the high-contact conditions

$$V(S_f, \tau) = K - S_f, \quad \frac{\partial V(S_f, \tau)}{\partial S} = -1 \quad (14)$$

This is applied to the approximation w as well. Two remaining conditions are given by the requirement

that both w and dw/dS are continuous at the matching point $S_f(\tau_{v-1})$. This fixes all variables. For the first line, the four equations for the parameters are (with $E := S_f(\tau_1)$)

$$\frac{K}{1 + r \Delta \tau} - E + a^{(1)} E^{\lambda_1} + b^{(1)} E^{\lambda_2} = K - E \quad (15)$$

$$-1 + \lambda_1 a^{(1)} E^{\lambda_1-1} + \lambda_2 b^{(1)} E^{\lambda_2-1} = -1 \quad (16)$$

$$\frac{K}{1 + r \Delta \tau} - K + a^{(1)} K^{\lambda_1} + b^{(1)} K^{\lambda_2} = c^{(1)} K^{\lambda_2} \quad (17)$$

$$-1 + \lambda_1 a^{(1)} K^{\lambda_1-1} + \lambda_2 b^{(1)} K^{\lambda_2-1} = \lambda_2 c^{(1)} K^{\lambda_2-1} \quad (18)$$

which has made use of $S_f(\tau_0) = K$. The solution of this system is tedious, we skip the derivation. This explains the analytic solution of equation (13) along the first line.

General Case

The following lines ($v \geq 2$) lead to even more involved equations, because in subinterval C the inhomogeneous solution is nontrivial, and additional subintervals B are inserted for each new line. The general structure of the solutions of the ODE is the same in the subintervals B, but the coefficients differ. And, of course, the points $S_f(\tau_v)$ vary. The resulting analytic method of lines is quite involved. The final formulas from [3] for a put with $\delta \leq r$ are the following:

Notations

$$\gamma = \frac{1}{2} - \frac{r - \delta}{\sigma^2}$$

$$R = \frac{1}{1 + r \Delta \tau} \quad (\text{single-period discount factor})$$

$$D = \frac{1}{1 + \delta \Delta \tau}$$

$$\varepsilon = \sqrt{\gamma^2 + \frac{2}{R\sigma^2\Delta\tau}}$$

$$p = \frac{\varepsilon - \gamma}{2\varepsilon}, \quad q = 1 - p$$

$$\hat{p} = \frac{\varepsilon - \gamma + 1}{2\varepsilon}, \quad \hat{q} = 1 - \hat{p}$$

$$S_{f,v} := S_f(\tau_v)$$

$V^{(n)}(S)$ = approximation of the American put for
 $t = 0$

After n lines, the solutions consist of $n + 2$ pieces,

$$V^{(n)}(S) = \begin{cases} K - S & \text{for } S \leq S_{f,n} \\ v_v^{(n)}(S) + b_v^{(n)}(S) + A_v^{(n)}(S; 1) & \text{for } S_{f,v} < S \leq S_{f,v-1}, v = 1, \dots, n \\ p_0^{(n)}(S) + b_1^{(n)}(S) & \text{for } S > S_{f,0} \equiv K \end{cases} \quad (19)$$

This piecewise defined function represents $w(S, \tau_n)$ and corresponds to $V(S, 0)$. The coefficients are

$$p_0^{(n)}(S) = \left(\frac{S}{K}\right)^{\gamma-\varepsilon} \sum_{k=0}^{n-1} \frac{(2\varepsilon \ln(S/K))^k}{k!} \times \sum_{l=0}^{n-k-1} \binom{n-1+l}{n-1} \times [K R^n q^n p^{l+k} - K D^n \hat{q}^n \hat{p}^{l+k}] \quad (20)$$

$$v_v^{(n)}(S) = K R^{n-v+1} - S D^{n-v+1} \quad (21)$$

$$b_v^{(n)}(S) = \sum_{j=1}^{n-v+1} \left(\frac{S}{S_{f,n-j+1}}\right)^{\gamma-\varepsilon} \times \sum_{k=0}^{j-1} \frac{(2\varepsilon \ln(S/S_{f,n-j+1}))^k}{k!} \times \sum_{l=0}^{j-k-1} \binom{j-1+l}{j-1} \times [q^j p^{k+l} R^j K r - \hat{q}^j \hat{p}^{k+l} D^j S_{f,n-j+1} \delta] \Delta \tau \quad (22)$$

$$A_v^{(n)}(S; i) = \sum_{j=i}^{n-v+1} \left(\frac{S}{S_{f,n-j+1}}\right)^{\gamma+\varepsilon} \times \sum_{k=0}^{j-1} \frac{(2\varepsilon \ln(S_{f,n-j+1}/S))^k}{k!}$$

$$\times \sum_{l=0}^{j-k-1} \binom{j-1+l}{j-1} \times [p^j q^{k+l} R^j K r - \hat{p}^j \hat{q}^{k+l} D^j S_{f,n-j+1} \delta] \Delta \tau \quad (23)$$

The approximation of the optimal exercise prices $S_{f,m}$ are solutions of the equations

$$c_1^{(m)}(K) - A_1^{(m)}(K; 2) = \left(\frac{K}{S_{f,m}}\right)^{\gamma+\varepsilon} [p R K r - \hat{p} D S_{f,m} \delta] \Delta \tau \quad (24)$$

for $m = 1, \dots, n$, where

$$c_1^{(m)}(K) = \sum_{l=0}^{m-1} \binom{m-1+l}{m-1} \times [K D^m \hat{p}^m \hat{q}^l - K R^m p^m q^l] \quad (25)$$

This equation is solved iteratively with Newton's method. In case no dividend is paid ($\delta = 0$, $D = 1$), no iteration is needed, and the solution is

$$S_{f,m} = K \left(\frac{p R K r \Delta \tau}{c_1^{(m)}(K) - A_1^{(m)}(K; 2)} \right)^{1/(\gamma+\varepsilon)} \quad (26)$$

This value would serve as initial guess for the Newton iteration in case $\delta > 0$. The formulas of these triple sums are also collected in [2]. The corresponding formulas for a call are available too. Instead, the put-call symmetry relation from [4] can be used as well.

Extrapolation

For small values of n , the method does not provide high accuracy in its basic state. To enhance the quality of the approximation, Richardson extrapolation is applied. Let $\bar{V}_n$ denote the result of the above MoL with n lines, evaluating equation (19) for a given value of S . Assume that three approximations $\bar{V}_1$, $\bar{V}_2$, and $\bar{V}_3$ are calculated. Note that $\bar{V}_1$ means that only a single timestep of size $\Delta \tau = T$ is used. Then the extrapolated value

$$\bar{V}^{1:3} := \frac{1}{2}(9\bar{V}_3 - 8\bar{V}_2 + \bar{V}_1) \quad (27)$$

Table 1 Test results of the tuned $\bar{V}^{1:3}$ for $K = 100$

S	T	σ	r	δ	$V(S, 0)$
80	0.5	0.4	0.06	0.00	21.6257
100	0.5	0.4	0.02	0.00	10.7899
80	3.0	0.4	0.06	0.02	29.2323

gives an accurate result with order $(\Delta\tau)^3$. As shown in [1], the obtainable accuracy of the combined MoL/extrapolation approach compares well to the other methods. Reference [3] applies a fine tuning to the three-point formula $\bar{V}^{1:3}$ replacing the coefficient 8 in the above formula by $8(1 - 0.0002(5 - T)^+)$. Even with only two approximations, $\bar{V}_1, \bar{V}_2$, extrapolation enhances the accuracy significantly. The formula

$$\bar{V}^{1:2} := -\bar{V}^1 + 2\bar{V}^2 \quad (28)$$

is of the order $(\Delta\tau)^2$. The justification of Richardson extrapolation in this context does not yet appear fully explored; smoothness in time is assumed. This is not a disadvantage since equations (27) or (28) serve as analytic-approximation formulas.

Further Remarks

For comparison, we provide in Table 1, three test values of the tuned version of the three-line extrapolation

equation (27). To check the accuracy, an independent computation with a highly accurate version of a finite-difference approach was run. This has revealed relative errors of about 10^{-3} in the three examples reported in Table 1—that is, three digits are correct. For testing purposes, the MoL method with the tuned version of the formula $\bar{V}^{1:3}$ of equation (27) is installed in the option-calculator of the website www.compfin.de. Note that the above describes the analytic method of [3]; the MoL approach of [5] solves equation (7) numerically.

References

- [1] Broadie, M. & Detemple, J. (1996). American option valuation: new bounds, approximations, and a comparison of existing methods, *Review of Financial Studies* **9**, 1211–1250.
- [2] Carr, P. (1998). Randomization and the American put, *Review Financial Studies* **11**, 597–626.
- [3] Carr, P. & Faguet, D. (1995). *Fast Accurate Valuation of American Options*, Working Paper, Cornell University.
- [4] McDonald, R.L. & Schroder, M.D. (1998). A parity result for American options, *Journal of Computational Finance* **1**(3), 5–13.
- [5] Meyer, G.H. & van der Hoek, J. (1997). The valuation of American options with the method of lines, *Advances in Futures and Options Research* **9**, 265–285.
- [6] Seydel, R. (2006). *Tools for Computational Finance*, Springer, Berlin.

RÜDIGER U. SEYDEL

Fourier Methods in Options Pricing

Pricing European Options

There are various ways to price European options *via* Fourier inversion. Before we consider such methods, it is worth mentioning why Fourier option pricing methods can be useful. First recall that the risk-neutral valuation theorem states that the forward price of a European call option on a single asset S can be written as

$$C(S(t), K, T) = \mathbb{E}[(S(T) - K)^+] \quad (1)$$

where C denotes the value, T the maturity, and K the strike price of the call. The expectation is taken under the T -forward probability measure. As equation (1) is an expectation, it can be calculated *via* numerical integration, provided we know the density in closed form. For many models the density is either not known in closed form, or is quite cumbersome, whereas its characteristic function is often much easier to obtain. A good example in finance is the variance gamma model, introduced by Madan and Seneta [15]. Its density involves a Bessel function of the third kind, whereas its characteristic function only consists of elementary functions.

Heston [8] was among the first to utilize Fourier methods to price European options within his stochastic volatility model. Since Heston's seminal paper, the pricing of European options by means of Fourier inversion has become more and more commonplace. Heston's approach starts from the observation that equation (1) can be recast as

$$C(S(t), K, T) = F(t, T) \cdot \mathbb{S}(S(T) > K) - K \cdot \mathbb{P}(S(T) > K) \quad (2)$$

with $F(t, T)$ the forward price of the asset and $\mathbb{P}$ and $\mathbb{S}$ respectively the T -forward probability measure and the measure induced by taking the asset price as the numeraire. The cumulative probabilities in equation (2) can subsequently be found by Fourier inversion, an approach dating back to [6, 7, 11]. As this approach necessitates the evaluation of two Fourier inversions, and is inaccurate for out-of-the-money options, due to cancellation, we do not discuss

it here in further detail. Instead we focus on more recent approaches due to [3, 16], and [12].

Carr and Madan's approach was to consider the Fourier transform of the damped European call price with respect to the logarithm of the strike price:

$$\begin{aligned} \psi(v, \alpha) &\equiv \int_{-\infty}^{\infty} e^{ivk} e^{\alpha k} C(k) dk \\ &= \frac{\phi(v - i(\alpha + 1))}{-(v - i\alpha)(v - i(\alpha + 1))} \end{aligned} \quad (3)$$

and subsequently invert this to arrive at the desired call price:

$$\begin{aligned} C(S(t), K, T) &= \frac{e^{-\alpha k}}{2\pi} \int_{-\infty}^{\infty} e^{-ivk} \psi(v, \alpha) dv \\ &= \frac{e^{-\alpha k}}{\pi} \int_0^{\infty} \text{Re}(e^{-ivk} \psi(v, \alpha)) dv \end{aligned} \quad (4)$$

In equations (3) and (4), $\phi(u)$ is the characteristic function; $\phi(u) = \mathbb{E}[\exp(iu \ln S(T))]$. Sufficient conditions for equation (4) to exist are that the damping factor $\alpha > 1$, and that the $(\alpha + 1)$ st moment of the asset, $\phi(-(\alpha + 1)i)$, is finite. The first condition is required to make the damped call price an L^1 -integrable function, which is a sufficient condition for the existence of its Fourier transform.

Whereas Carr and Madan took the Fourier transform with respect to the strike price of the call option, Raible [16] and Lewis [12] used an approach that is slightly more general in that it does not require the existence of a strike in a payoff. Raible took the transform with respect to the log-forward price, Lewis used the log-spot price. Note that for all three methods, the Fourier transform of the option price can be decoupled into two parts, a payoff-dependent part, the payoff transform, and a model-dependent part, the characteristic function.

One of the restrictions on the damping factor α for a call price is that it must be larger than 1. However, as Lewis [12] and Lee [9] point out, this is not a real restriction if we recast equation (4) as a contour integral in the complex plane. Shifting the contour, equivalent to varying α in equation (4), and carefully applying Cauchy's residue theorem leads to the following option pricing equation:

$$\begin{aligned} C(S(t), K, T, \alpha) \\ = R(F(t, T), K, \alpha) \end{aligned}$$

$$+ \frac{e^{-\alpha k}}{\pi} \int_0^\infty \operatorname{Re}(e^{-ivk} \psi(v, \alpha)) dv \quad (5)$$

where the residue term $R(F, K, \alpha)$ equals 0 for $\alpha > 0$, $1/2F$ for $\alpha = 0$, F for $\alpha \in (-1, 0)$, $F - 1/2K$ for $\alpha = -1$, and $F - K$ for $\alpha < -1$. For values of $\alpha < -1$, for example, this means that the integral in equation (5) yields the value of a put, from which we obtain the price of a call *via* put–call parity.

As far as the numerical implementation of equation (5) goes, the appropriate numerical algorithm depends, as always, on the user’s demands. If one is interested in the option price at a great many strike prices, equation (5) can be discretized in such a way that it is amenable to use of the fast Fourier transform (FFT), as detailed in [3] and in the article **Fourier Transform**. If one is calibrating a model to a volatility surface, one often only needs to evaluate option prices at a handful of strikes, at which point a direct integration of equation (5) becomes computationally advantageous. Important points to consider in approximating the semi-infinite integral in equation (5) are the discretization error, the truncation error, and the choice of contour or damping factor α . Lee [9] extensively analyzes these choices when equation (5) is discretized using the discrete Fourier transform (DFT), and proposes a minimization algorithm to determine the parameters of the discretization.

Lord and Kahl [14] propose a different approach. If an appropriate transformation function is available that maps the semi-infinite interval into a finite one, the truncation error can be avoided altogether. This leaves the discretization error, which can be

controlled by using adaptive integration methods. Finally, the speed and accuracy of the integration algorithm can be controlled by choosing an appropriate value of α . A good proxy for the optimal value of α , the value that minimizes the approximation error given a fixed computational budget, is

$$\begin{aligned} \alpha^* &= \arg \min_{\alpha \in (\alpha_{\min}, \alpha_{\max})} |e^{-\alpha k} \psi(0, \alpha)| \\ &= \arg \min_{\alpha \in (\alpha_{\min}, \alpha_{\max})} -\alpha k + \frac{1}{2} \ln(\psi(0, \alpha)^2) \quad (6) \end{aligned}$$

where $(\alpha_{\min}, \alpha_{\max})$ is the allowed range of α , corresponding to $\phi(-(\alpha + 1)i) < \infty$. This choice of contour is closely linked to how the optimal contour is chosen in saddlepoint approximations. That α really can have a significant impact on the accuracy should become clear from the following example.

Example 1—Impact of α on the Numerical Approximation in Heston’s Model

As an example, we look at the impact of α in Heston’s stochastic volatility model (*see Heston Model*). The parameters we pick are $\kappa = \omega = 1$, $\rho = -0.7$, $\theta = v(0) = 0.1$, $F = 1$, $K = 1.5$ and the time to expiry is 0.1 years. Figure 1 shows the impact of α on the approximation error when two different ways are used to discretize equation (5).

If one plots the function that is minimized in equation (6), one obtains a very similar pattern, suggesting that α^* is indeed close to optimal.

Finally, Andersen and Andreasen [2] and Cont and Tankov [4] have suggested that the Black–Scholes

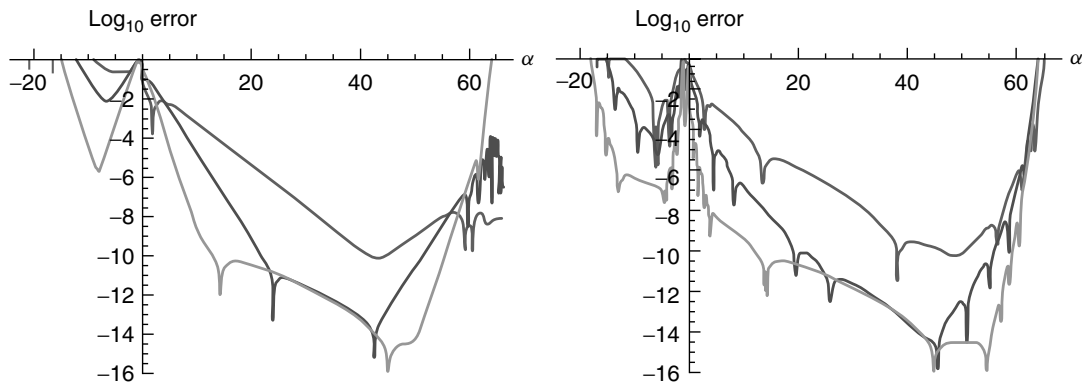


Figure 1 Impact of α using (a) Lee [9] DFT discretization, or (b) Gauss–Legendre quadratures. The various lines represent the number of abscissae used (8, 16, or 32)

option price could be used as a control variate in the evaluation of equation (5), that is, we could subtract the Black–Scholes integrand from the integrand, and subsequently add the Black–Scholes price back to the equation. While this could work for some models, the approach does require a good correspondence between both characteristic functions, and also requires an educated guess as to which Black–Scholes volatility should be used.

Pricing Bermudan and American Options

Now that we can price European options using Fourier methods, the next question is whether options with early exercise features can be priced in a similar framework. The answer is in the affirmative. The first paper to attempt this in the framework of Carr and Madan was by O’Sullivan [18], who extended the QUAD method of [1] (see **Quadrature Methods**) to allow for models where the density is not known in closed form, but has to be approximated *via* Fourier inversion. This method has complexity $O(MN^2)$, where M is the number of time steps and N is the number of discretization points used in a one-dimensional model.

Building upon a presentation by Reiner [17], Lord *et al.* [13] noticed that the key to extending Carr and Madan’s approach to early exercise options was to abandon the idea of working with an analytical Fourier transform of the option payoff and to numerically approximate it. If at time t_m we have an expression for the value of the option contract, then its continuation value at t_{m-1} can be obtained by calculating its convolution with the transition density. As we know the Fourier transform of a convolution is the product of the individual Fourier transforms, all we need to do is numerically calculate the Fourier transform of the continuation value. Having calculated the continuation value, we obtain the value at time t_{m-1} simply by comparison with the exercise value. The CONV method of Lord *et al.* [13] utilizes the FFT to approximate the convolutions. As such, the algorithm is $O(MN \log N)$. For Bermudan options, the algorithm is certainly competitive with the fastest partial integro-differential differential equation (PIDE) methods (see **Partial Integro-differential Equations (PIDEs)**); see the numerical comparison in [13]. The prices of American options can be obtained *via* Richardson extrapolation. It is

here that PIDE methods have an advantage. Another area where PIDE methods are advantageous is at the choice of gridpoints—as the CONV method employs the FFT, the grid for the logarithm of the asset price needs to be uniform. This makes it harder to place discontinuities on the grid, something which is much easier to achieve in, for example, the QUAD or PIDE methods. Extensions of the CONV method to multiple dimensions can be found in [10].

Finally, we mention a recent paper by Fang and Oosterlee [5], in which Bermudan options are efficiently priced *via* Fourier cosine series expansions. While this method is also $O(MN \log N)$ and has some similarities with the CONV method, a great advantage is that the exercise boundary is directly solved, as in the QUAD method. Hence, the cosine series coefficients can be calculated exactly, instead of being approximated, which is the case in the CONV method. Whereas the convergence of the CONV method is dictated by the chosen Newton–Cotes rule, the convergence of the COS method is dictated by the rate of decay of the characteristic function.

References

- [1] Andricopoulos, A.D., Widdicks, M., Duck, P.W. & Newton, D.P. (2003). Universal option valuation using quadrature, *Journal of Financial Economics* **67**(3), 447–471.
- [2] Andersen, L.B.G. & Andreasen, J. (2002). Volatile volatilities, *Risk* **15**(12), 163–168.
- [3] Carr, P. & Madan, D.B. (1999). Option valuation using the Fast Fourier Transform, *Journal of Computational Finance* **2**(4), 61–73.
- [4] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, Chapman and Hall.
- [5] Fang, F. & Oosterlee, C.W. (2008). *Pricing Early-exercise and Discrete Barrier Options by Fourier-cosine Series Expansions*, working paper, Delft University of Technology and CWI.
- [6] Gil-Pelaez, J. (1951). Note on the inversion theorem, *Biometrika* **37**, 481–482.
- [7] Gurland, J. (1948). Inversion formulae for the distribution of ratios, *Annals of Mathematical Statistics* **19**, 228–237.
- [8] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**(2), 327–343.
- [9] Lee, R.W. (2004). Option pricing by transform methods: extensions, unification and error control, *Journal of Computational Finance* **7**(3), 51–86.

- [10] Leentvaar, C.C.W. & Oosterlee, C.W. (2008). Multi-asset option pricing using a parallel Fourier-based technique, *Journal of Computational Finance* **12**(1), 1–26.
- [11] Lévy, P. (1925). *Calcul des Probabilités*, Gauthier-Villars, Paris.
- [12] Lewis, A. (2001). *A Simple Option Formula for General Jump-diffusion and other Exponential Lévy Processes*, working paper, OptionCity.net, <http://ssrn.com/abstract=282110>.
- [13] Lord, R., Fang, F., Bervoets, F. & Oosterlee, C.W. (2008). A fast and accurate FFT-based method for pricing early-exercise options under Lévy processes, *SIAM Journal on Scientific Computing* **30**(4), 1678–1705.
- [14] Lord, R. & Kahl, C. (2007). Optimal Fourier inversion in semi-analytical option pricing, *Journal of Computational Finance* **10**(4), 1–30.
- [15] Madan, D.B. & Seneta, E. (1990). The variance gamma (V.G.) model for share market returns, *Journal of Business* **63**(4), 511–524.
- [16] Raible, S. (2000). *Lévy Processes in Finance: Theory, Numerics and Empirical Facts*. PhD thesis, Institut für Mathematische Stochastik, Albert-Ludwigs-Universität, Freiburg.
- [17] Reiner, E. (2001). Convolution methods for path-dependent options, *Financial Mathematics Workshop*, Institute for Pure and Applied Mathematics, UCLA, January 2001, available at: http://www.ipam.ucla.edu/publications/fm2001/fm2001_4272.pdf.
- [18] O’Sullivan, C. (2005). *Path Dependent Option Pricing under Lévy Processes*, EFA 2005 Moscow meetings paper, available at: <http://ssrn.com/abstract=673424>.

Related Articles

Fourier Transform; Partial Integro-differential Equations (PIDEs); Quadrature Methods; Wavelet Galerkin Method.

ROGER LORD

Quantization Methods

The origin of optimal vector quantization goes back to the early 1950s as a way to discretize a (stationary) signal so that it could be transmitted for a given cost with the lowest possible degradation. The starting idea is to consider the best approximation in the mean quadratic sense—or, more generally, in an L^p -sense—of an $\mathbb{R}^d$ -valued random vector X by a random variable $q(X)$ taking at most N values (with respect to a given norm on $\mathbb{R}^d$, usually the canonical Euclidean norm).

More recently (in the late 1990s), it has been introduced as an efficient tool in numerical probability—first, for numerical integration in medium dimensions [15, 16], and soon as a method for the computation of conditional expectations. The main motivation was the pricing and hedging of multias-set American-style options [2, 4] and more generally to devise some realistic numerical schemes for the reflected backward stochastic differential equations (SDEs) (see **Backward Stochastic Differential Equations** and [1, 3]). Presently, this ability to compute conditional expectations has led to tackling other nonlinear problems like stochastic control (portfolio management [18], pricing of swing options [5, 6]), nonlinear filtering with some applications to stochastic volatility models [20], and some classes of stochastic partial differential equations (PDEs) like stochastic Zakai and McKean–Vlasov equations [8]. In all these problems, quantization is used to produce a space discretization of the underlying (Markov) dynamics at the time discretization instants (see also [19]).

Optimal Vector Quantization: A Short Background

Assume $\mathbb{R}^d$ is equipped with the Euclidean norm $|\cdot|$. Let $X : (\Omega, \mathcal{A}, \mathbb{P}) \rightarrow \mathbb{R}^d$ be a random vector. For a given set $\Gamma = \{x_1, \dots, x_N\} \subset \mathbb{R}^d$, $N \geq 1$, any (Borel) projection $\widehat{X}^\Gamma$ of X on Γ following the *nearest neighbor rule* provides an optimal solution

$$|X - \widehat{X}^\Gamma| = d(X, \Gamma) = \min_{1 \leq i \leq N} |X - x_i| \quad (1)$$

The projection is essentially unique if all hyperplanes have 0-mass for the distribution of X . If $X \in L^2(\mathbb{P})$,

the induced mean quadratic error $\|X - \widehat{X}^\Gamma\|_2$ reaches a minimum as Γ runs over all subsets of $\mathbb{R}^d$ of size at most N . Any such minimizer $\Gamma^{N,*}$ is called an *optimal quadratic N -quantizer* of X and $\widehat{X}^{\Gamma^{N,*}}$ is called an *optimal quadratic N -quantization* of X . Using the property that conditional expectation given a σ -field $\mathcal{B}$ is the best $\mathcal{B}$ -measurable quadratic approximation, one derives the result that an optimal quantizer satisfies the so-called *stationary property*:

$$\mathbb{E}(X|\widehat{X}) = \widehat{X} \quad \text{with} \quad \widehat{X} := \widehat{X}^{\Gamma^{N,*}} \quad (2)$$

It is easy to show that the minimal mean quantization error $\|X - \widehat{X}^{\Gamma^{N,*}}\|_2$ is nonincreasing and goes to zero as $N \rightarrow \infty$ (decreasing if the support of X is infinite). Optimal quantizers also exist with respect to the $L^r(\mathbb{P})$ -norm, $r \neq 2$. The rate of convergence is ruled by the Zador Theorem.

Theorem 1 (a) Sharp rate [9]. Let $X \in L^{r+\eta}(\mathbb{P})$ for some $r, \eta > 0$. Let $\mathbb{P}_x(d\xi) = \varphi(\xi)d\xi + \nu(d\xi)$ be the canonical decomposition of the distribution of X (ν and the Lebesgue measure are singular). Then there exists a real constant $\tilde{J}_{r,d} \in (0, \infty)$ such that

$$\lim_N N^{\frac{1}{d}} \min_{\Gamma \subset \mathbb{R}^d, \text{card}(\Gamma) \leq N} \|X - \widehat{X}^\Gamma\|_r \sim \tilde{J}_{r,d} \times \left(\int_{\mathbb{R}^d} \varphi^{\frac{d}{d+r}}(u) du \right)^{\frac{1}{d} + \frac{1}{r}} \quad \text{as } N \rightarrow +\infty \quad (3)$$

(b) Nonasymptotic upper bound [13]. Let $d \in \mathbb{N}$. Let $r, \eta > 0$. There exists $C_{d,r,\eta} \in (0, \infty)$ such that, for every $\mathbb{R}^d$ -valued random vector X ,

$$\forall N \geq 1,$$

$$\min_{\Gamma \subset \mathbb{R}^d, |\Gamma| \leq N} \|X - \widehat{X}^\Gamma\|_r \leq C_{d,r,\eta} \|X\|_{r+\eta} N^{-\frac{1}{d}} \quad (4)$$

The real constant $\tilde{J}_{r,d}$ (which depends on the underlying norm on $\mathbb{R}^d$) corresponds to the case of the uniform distribution over $[0, 1]^d$ for which the above “ $\lim_N$ ” also holds as an “ $\inf_N$ ” as well. When $d = 1$, $\tilde{J}_{r,1} = (r+1)^{-\frac{1}{r}}/2$. When $d = 2$, with the canonical Euclidean norm, $\tilde{J}_{2,d} = \sqrt{\frac{5}{18\sqrt{3}}}$. For $d \geq 3$ one only knows that $\tilde{J}_{2,d} \sim \sqrt{\frac{d}{2\pi e}} \approx \sqrt{\frac{d}{17.08}}$ as $d \rightarrow$

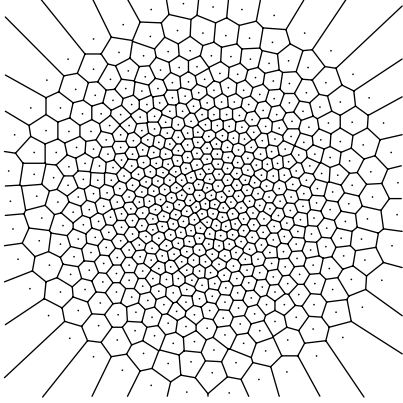


Figure 1 N -quantizer (and its Voronoi diagram) of the normal distribution $\mathcal{N}(0; I_2)$ on $\mathbb{R}^2$ with $N = 500$ (The Voronoi diagram never needs to be computed for numerics)

$+\infty$. For more results on the theoretical aspects of vector quantization we refer to [9] and the references therein. Figure 1 shows a quantization of the bivariate normal distribution of size $N = 500$.

Some Quantization-based Cubature Formulae

Let X be an $\mathbb{R}^d$ -valued random vector and Y an $\mathbb{R}^q$ -valued random vector; let $\Gamma_X = \{x_1, \dots, x_{N_X}\}$, $\Gamma_Y = \{y_1, \dots, y_{N_Y}\}$ be two quantizers of X and Y , respectively. Let $F : \mathbb{R}^d \rightarrow \mathbb{R}$ be a (continuous) function. It seems natural to approximate these quantities by their quantized version, that is,

$$\mathbb{E}(F(X)) \approx \mathbb{E}(F(\widehat{X}^{\Gamma_X})) \quad \text{and} \\ \mathbb{E}(F(X)|Y) \approx \mathbb{E}(F(\widehat{X}^{\Gamma_X})|\widehat{Y}^{\Gamma_Y})$$

$$\text{where } \mathbb{E}(F(\widehat{X}^{\Gamma_X})) = \sum_{1 \leq i \leq N_X} F(x_i) \mathbb{P}(\widehat{X}^{\Gamma_X} = x_i) \quad (5)$$

$$\mathbb{E}(F(\widehat{X}^{\Gamma_X})|\widehat{Y}^{\Gamma_Y}) \\ = \sum_{1 \leq i \leq N_X} F(x_i) \mathbb{P}(\widehat{X}^{\Gamma_X} = x_i | \widehat{Y}^{\Gamma_Y} = y_j), \\ 1 \leq j \leq N_Y \quad (6)$$

Numerical computation of $\mathbb{E}(F(\widehat{X}^{\Gamma_X}))$ and $\mathbb{E}(F(\widehat{X}^{\Gamma_X})|\widehat{Y}^{\Gamma_Y})$ is possible as soon as $F(\xi)$ is computable at any $\xi \in \mathbb{R}^d$ and both the distribution $(\mathbb{P}(\widehat{X}^{\Gamma_X} = x_i))_{1 \leq i \leq N_X}$ of $\widehat{X}^{\Gamma_X}$ and the conditional distribution of $\widehat{X}^{\Gamma_X}$ given $\widehat{Y}^{\Gamma_Y}$ are made explicit.

Likewise, one can consider *a priori* the $\sigma(\widehat{X}^{\Gamma_X})$ -measurable random variable $F(\widehat{X}^{\Gamma_X})$ as a good approximation of the conditional expectation $\mathbb{E}(F(X)|\widehat{X}^{\Gamma_X})$. One shows (see, e.g., [25]) that

$$\|X - \widehat{X}^{\Gamma_X}\|_1 = \sup_{[F]_{\text{Lip}} \leq 1} |\mathbb{E}F(X) - \mathbb{E}F(\widehat{X}^{\Gamma_X})| \quad (7)$$

where $[F]_{\text{Lip}}$ denotes the Lipschitz coefficient of F . If, furthermore, $\varphi_F : \mathbb{R}^q \rightarrow \mathbb{R}$, which is a (Borel) version of the conditional expectation, that is, satisfying $\mathbb{E}(F(X)|Y) = \varphi_F(Y)$ turns out to be Lipschitz, then

$$\|\mathbb{E}(F(X)|Y) - \mathbb{E}(F(\widehat{X}^{\Gamma_X})|\widehat{Y}^{\Gamma_Y})\|_2 \\ \leq [F]_{\text{Lip}} \|X - \widehat{X}^{\Gamma_X}\|_2 + [\varphi_F]_{\text{Lip}} \|Y - \widehat{Y}^{\Gamma_Y}\|_2 \quad (8)$$

When F is twice differentiable with a Lipschitz differential and Γ_X is a stationary quantizer, then

$$|\mathbb{E}F(X) - \mathbb{E}F(\widehat{X}^{\Gamma_X})| \leq [DF]_{\text{Lip}} \|X - \widehat{X}^{\Gamma_X}\|_2^2 \quad (9)$$

Similar cubature formulas can be established for locally Lipschitz functions such that $|F(x) - F(y)| \leq C|x - y|(1 + g(x) + g(y))$ where g is a nonnegative, nondecreasing, convex function (e.g., $g(x) = e^{|a||x|}$). Finally, when F is convex and $\widehat{X}^{\Gamma_X}$ is stationary, Jensen's inequality yields

$$\mathbb{E}(F(X)|\widehat{X}^{\Gamma_X}) \geq F(\mathbb{E}(X|\widehat{X}^{\Gamma_X})) \\ = F(\widehat{X}^{\Gamma_X}) \quad (\text{so that } \mathbb{E}(F(\widehat{X}^{\Gamma_X})) \leq \mathbb{E}(F(X))) \quad (10)$$

Example: Pricing a Bermuda Option Using a Quantization Tree

Let $(X_k)_{0 \leq k \leq n}$ be a Markov chain modeling the dynamics of d traded risky assets (interest rate is set to 0 for simplicity), assumed to be homogeneous for the sake of simplicity, with Lipschitz transition $P(x, dy) = \mathcal{L}(X_{k+1}|X_k = x)$, that is, satisfying the condition that for every Lipschitz continuous function f , $[Pf]_{\text{Lip}} \leq [P]_{\text{Lip}}[f]_{\text{Lip}}$. Let $(\widehat{X}_k)_{0 \leq k \leq n}$ be a sequence of quantizations $(\widehat{X}_k := \widehat{X}_k^{\Gamma_k})$ where the grids $\Gamma_k := \{x_1^k, \dots, x_{N_k}^k\} \subset \mathbb{R}^d$ are optimal, see the section How to Get Optimal Quantization below). These grids (and the related quantized transition probability weights defined below) are called a *quantization tree* of the chain.

Let $(h(X_k))_{0 \leq k \leq n}$ be a Bermuda (vanilla) payoff. Then, one can approximate the premium

$$\mathcal{V}_0 = \sup \{ \mathbb{E}(h(X_\tau)), \tau \mathcal{F}^X\text{-stopping time} \} \quad (11)$$

of the option by implementing a backward quantized dynamic programming formula as follows:

$$\begin{aligned} \widehat{\mathcal{V}}_n &= h(\widehat{X}_n), \\ \widehat{\mathcal{V}}_k &= \max(h(\widehat{X}_k), \mathbb{E}(\widehat{\mathcal{V}}_{k+1} | \widehat{X}_k)), \quad k = 0, \dots, n-1 \end{aligned} \quad (12)$$

in which the Markov property is “forced” since $(\widehat{X}_k)_{0 \leq k \leq n}$ has no reason to be a Markov chain. In practice, one shows that $\widehat{\mathcal{V}}_k = v_k(\widehat{X}_k)$ where the functions v_k defined on Γ_k satisfy the following backward induction:

$$v_n(x_i^n) = h(x_i^n), \quad i = 1, \dots, N_n \quad (13)$$

$$v_k(x_i^k) = \max \left(h(x_i^k), \sum_{x_j \in \Gamma_{k+1}} \widehat{p}_k^{ij} v_{k+1}(x_j^{k+1}) \right), \quad i = 1, \dots, N_k \quad (14)$$

$$\text{where } \widehat{p}_k^{ij} = \mathbb{P}(\widehat{X}_{k+1} = x_j^{k+1} | \widehat{X}_k = x_i^k) \quad (15)$$

The point is that once the transitions $\widehat{p}_k^{ij}$ have been computed (e.g., by a—possibly parallelized, see [5]—Monte Carlo simulation), the above backward induction can be applied to any (reasonable) payoff: the quantization-based approach is not payoff dependent as the regression-based simulation methods (*see Bermudan Options*) are. The resulting error bound, combining equation (8) and the Zador Theorem(b) is

$$\begin{aligned} |\mathcal{V}_0 - \mathbb{E}v_0(X_0)| &\leq C_{[P]_{\text{Lip}}, d} \sum_{k=0}^n \|X_k - \widehat{X}_k\|_2 \\ &= O\left(\frac{n}{\bar{N}^{\frac{1}{d}}}\right) \end{aligned} \quad (16)$$

where $\bar{N} := \frac{1}{n+1} \sum_{k=0}^n N_k$. First-order schemes have been devised involving the approximation $\widehat{D}v_k$ of the (space) differential of Dv_k of v_k in [3]. Other quantization-based schemes have been devised for many other problems (stochastic control, nonlinear filtering [20], etc.).

How to Get Optimal Quantization

For this aspect, which is clearly critical for applications, we mainly refer to [17, 21, 25] and the references therein. We just say that the two main procedures are both based on the stationary equation $\widehat{X}^\Gamma = \mathbb{E}(X | \widehat{X}^\Gamma)$. The randomized Lloyd’s I is the induced fixed-point procedure whereas the *competitive learning vector quantization* algorithm is a recursive stochastic gradient zero search procedure. Both are based on the massive simulation of independent and identically distributed (i.i.d.) copies of X and nearest neighbor search. Recent developments in the field of fast versions of such procedures [7, 14] clearly open new perspectives to the online implementation of quantization-based methods. Regarding the Gaussian distribution, a quantization process has been completed and some optimal grids are available on the website [23]: www.quantize.maths-fi.com.

New Directions

Although optimal quantization is an autonomous field of research at the intersection of approximation theory, information theory, and probability theory, which has its own life, it seems that it generates many ideas that can easily and efficiently be applied to numerical probability and computational finance.

One important direction, not developed here, is *functional quantization* where a stochastic process—for example, the Brownian motion, a Lévy process, or a diffusion—is quantized as a random variable taking values in its path space (see [10–12]) or [17] for a survey, and the references therein). This has been applied to the pricing of path-dependent options in [22]. See also the website [23] to download optimal quadratic functional quantizers of the Brownian motion.

Another direction is variance reduction where quantization can be used either as a control variate or a stratification method, with, in both cases, the specificity being an optimal way to proceed among Lipschitz continuous functions/functionals [22, 24, 25].

References

- [1] Bally, V. & Pagès, G. (2003). A quantization algorithm for solving discrete time multidimensional optimal stopping problems, *Bernoulli* **9**(6), 1003–1049.

- [2] Bally, V., Pagès, G. & Printems, J. (2001). A stochastic quantization method for non-linear problems, *Monte Carlo Methods and Applications* **7**(1), 21–34.
- [3] Bally, V., Pagès, G. & Printems, J. (2003). First order schemes in the numerical quantization method, *Mathematical Finance* **13**(1), 1–16.
- [4] Bally, V., Pagès, G. & Printems, J. (2005). A quantization tree method for pricing and hedging multidimensional American options, *Mathematical Finance* **15**(1), 119–168.
- [5] Bardou, O., Bouthemy, S. & Pagès, G. (2007). Pricing swing options using optimal quantization, pre-print LPMA-1146, to appear in *Applied Mathematical Finance*.
- [6] Bardou, O., Bouthemy, S. & Pagès, G. (2007). When are swing option bang-bang and how to use it? pre-print LPMA-1141, submitted.
- [7] Friedman, J.H., Bentley, J.L. & Finkel, R.A. (1977). An algorithm for finding best matches in logarithmic expected time, *ACM Transactions on Mathematical Software* **3**(3), 209–226.
- [8] Gobet, E., Pagès, G., Pham, H. & Printems, J. (2007). Discretization and simulation of the Zakai equation, *SIAM Journal on Numerical Analysis* **44**(6), 2505–2538. See also, Discretization and simulation for a class of SPDEs with applications to Zakai and McKean-Vlasov equations, Pré-pub. PMA-958, 2005.
- [9] Graf, S. and Luschgy, H. (2000). *Foundations of Quantization for Probability Distributions*, Lecture Notes in Mathematics 1730, Springer, Berlin, 230.
- [10] Luschgy, H. & Pagès, G. (2002). Functional quantization of Gaussian processes, *Journal of Functional Analysis* **196**(2), 486–531.
- [11] Luschgy, H. & Pagès, G. (2004). Sharp asymptotics of the functional quantization problem for Gaussian processes, *The Annals of Probability* **32**(2), 1574–1599.
- [12] Luschgy, H. & Pagès, G. (2006). Functional quantization of a class of Brownian diffusions: A constructive approach, *Stochastic Processes and Applications* **116**, 310–336.
- [13] Luschgy, H. & Pagès, G. (2008). Functional quantization rate and mean regularity of processes with an application to Lévy processes, *Annals of Applied Probability* **18**(2), 427–469.
- [14] McNames, J. (2001). A fast nearest-neighbor algorithm based on a principal axis search tree, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(9), 964–976.
- [15] Pagès, G. (1993). Voronoi tessellation, space quantization algorithm and numerical integration, in *Proceedings of the ESANN'93*, M. Verleysen, ed, Editions D Facto, Bruxelles, p. 221–228.
- [16] Pagès, G. (1998). A space vector quantization method for numerical integration, *Journal of Computational and Applied Mathematics* **89**, 1–38.
- [17] Pagès, G. (2007). Quadratic optimal functional quantization methods and numerical applications, in *Proceedings of MCQMC, Ulm'06*, Springer, Berlin, p. 101–142.
- [18] Pagès, G. & Pham, H. (2005). Optimal quantization methods for non-linear filtering with discrete-time observations, *Bernoulli* **11**(5), 893–932.
- [19] Pagès, G., Pham, H. & Printems, J. (2003). Optimal quantization methods and applications to numerical methods in finance in *Handbook of Computational and Numerical Methods in Finance*, S.T. Rachev, ed, Birkhäuser, Boston, p. 429.
- [20] Pagès, G., Pham, H. & Printems, J. (2004). An optimal Markovian quantization algorithm for multidimensional stochastic control problems, *Stochastics and Dynamics* **4**(4), 501–545.
- [21] Pagès, G. & Printems, J. (2003). Optimal quadratic quantization for numerics: the Gaussian case, *Monte Carlo Methods and Applications* **9**(2), 135–165.
- [22] Pagès, G. & Printems, J. (2005). Functional quantization for numerics with an application to option pricing, *Monte Carlo Methods and Applications* **11**(4), 407–446.
- [23] Pagès, G. & Printems, J. (2005). Website devoted to vector and functional optimal quantization, www.quantize.maths-fi.com.
- [24] Pagès, G. & Printems, J. (2008). Reducing variance using quantization, pre-pub. LPMA, submitted.
- [25] Pagès, G. & Printems, J. (2008). Optimal quantization for finance: from random vectors to stochastic processes, in *Handbook of Numerical Analysis*, P. G. Ciarlet, ed, special volume: Mathematical Modelling and Numerical Methods in Finance, A. Bensoussan & Q. Zhang, (guest editors), North-Holland, Netherlands, Vol. XV pp. 595–648, ISBN: 978-0-444-51879-8.

Further Reading

- Bally, V. & Pagès, G. (2003). Error analysis of the quantization algorithm for obstacle problems, *Stochastic Processes & Their Applications*, **106**(1), 1–40.
- Pagès, G. & Sellami, A. (2007). Convergence of multi-dimensional quantized SDE's, Pré-print LPMA-1196, submitted.

Related Articles

American Options; Bermudan Options; Stochastic Mesh Method; Tree Methods.

GILLES PAGÈS

Monte Carlo Simulation for Stochastic Differential Equations

Weak Convergence Criterion

There exists a well-developed literature on classical Monte Carlo methods. We might mention, among others, Hammersley and Handscomb [3] and Fishman [1]. This literature, however, does not focus on approximating functionals of solutions of *stochastic differential equations* (SDEs). Exploiting the stochastic analytic structure of the underlying SDEs in discrete-time approximations allows to obtain deeper insight and significant benefits in a Monte Carlo simulation. One can construct more efficient methods than usually would be obtained under the classical Monte Carlo approach. Monographs on Monte Carlo methods for SDEs include [6, 7, 10]. The area of Monte Carlo methods in finance is the focus of the well-written books by Jackel [5] and Glasserman [2]. In [12], one can find a brief survey of such methods. In many circumstances, when other numerical methods fail or are difficult to implement, Monte Carlo methods can still provide a reasonable answer. This applies, in particular, to problems involving a large number of factors.

First, let us introduce a criterion that provides a classification of different simulation schemes. In Monte Carlo simulation, one concentrates on the approximation of probabilities and expectations of payoffs. Consider the process $X = \{X_t, t \in [0, T]\}$, which is the solution of the SDE

$$dX_t = a(X_t) dt + b(X_t) dW_t \quad (1)$$

for $t \in [0, T]$ with $X_0 \in \mathfrak{R}$. We shall say that a discrete-time approximation Y^Δ *converges with weak order* $\beta > 0$ to X at time T as $\Delta \rightarrow 0$ if for each polynomial $g : \mathfrak{R} \rightarrow \mathfrak{R}$ there exists a constant C_g , which does not depend on Δ , and a $\Delta_0 \in [0, 1]$ such that

$$\mu(\Delta) = |E(g(X_T)) - E(g(Y_T^\Delta))| \leq C_g \Delta^\beta \quad (2)$$

for each $\Delta \in (0, \Delta_0)$. We call this also the *weak convergence criterion*.

Systematic and Statistical Error

Under the weak convergence criterion (2) functionals of the form

$$u = E(g(X_T)) \quad (3)$$

are approximated *via weak approximations* Y^Δ of the solution of the SDE (1). One can form a *raw Monte Carlo estimate* by using the sample average

$$u_{N,\Delta} = \frac{1}{N} \sum_{k=1}^N g(Y_T^\Delta(\omega_k)) \quad (4)$$

with N independent simulated realizations $Y_T^\Delta(\omega_1), Y_T^\Delta(\omega_2), \dots, Y_T^\Delta(\omega_N)$ of a discrete-time weak approximation Y_T^Δ at time T , where $\omega_k \in \Omega$ for $k \in \{1, 2, \dots, N\}$. The corresponding *weak error* $\hat{\mu}_{N,\Delta}$ has the form

$$\hat{\mu}_{N,\Delta} = u_{N,\Delta} - E(g(X_T)) \quad (5)$$

which we decompose into a *systematic error* μ_{sys} and a *statistical error* μ_{stat} , such that

$$\hat{\mu}_{N,\Delta} = \mu_{\text{sys}} + \mu_{\text{stat}} \quad (6)$$

Here, we set

$$\begin{aligned} \mu_{\text{sys}} &= E(\hat{\mu}_{N,\Delta}) \\ &= E\left(\frac{1}{N} \sum_{k=1}^N g(Y_T^\Delta(\omega_k))\right) - E(g(X_T)) \\ &= E(g(Y_T^\Delta)) - E(g(X_T)) \end{aligned} \quad (7)$$

Thus, we have

$$\mu(\Delta) = |\mu_{\text{sys}}| \quad (8)$$

The absolute systematic error $|\mu_{\text{sys}}|$ can be interpreted as the absolute weak error and is a critical variable under the weak convergence criterion (2).

For a large number N of simulated independent realizations of Y^Δ , we can conclude from the central limit theorem that the statistical error μ_{stat} becomes asymptotically Gaussian with mean 0 and variance of the form

$$\text{Var}(\mu_{\text{stat}}) = \text{Var}(\hat{\mu}_{N,\Delta}) = \frac{1}{N} \text{Var}(g(Y_T^\Delta)) \quad (9)$$

Note that in equation (9) we used the independence of the realization for each ω_k . The expression (9) reveals a major disadvantage of the raw Monte Carlo method. One notes that the variance of the statistical error μ_{stat} decreases only with $1/N$. Consequently, the deviation

$$\text{Dev}(\mu_{\text{stat}}) = \sqrt{\text{Var}(\mu_{\text{stat}})} = \frac{1}{\sqrt{N}} \sqrt{\text{Var}(g(Y_T^\Delta))} \quad (10)$$

of the statistical error decreases at only the slow rate $1/\sqrt{N}$ as $N \rightarrow \infty$. This means that, unless one has to estimate the expectation of a random variable $g(Y_T^\Delta)$ with a small variance, one may need an extremely large number N of sample paths to achieve a reasonably small confidence interval. However, there exist various *variance reduction techniques* (see **Variance Reduction**), that deal with this problem.

We shall discuss now discrete-time approximations of solutions of SDEs that are appropriate for the Monte Carlo simulation of derivative prices or other functionals of diffusion processes. This means that we study the weak order of convergence of discrete-time approximations. By truncating appropriately the Wagner–Platen expansion (see **Stochastic Taylor Expansions**) one obtains *weak Taylor schemes*. The desired order of weak convergence determines the kind of truncation that must be used [6]. However, the truncations will be different from those required for the strong convergence of a comparable order, as described in **Stochastic Differential Equations: Scenario Simulation**. In general, weak Taylor schemes involve fewer terms compared with the strong Taylor schemes of the same order.

Euler and Simplified Weak Euler Scheme

The simplest weak Taylor scheme is the Euler scheme (see also **Stochastic Taylor Expansions**), which has the form

$$Y_{n+1} = Y_n + a \Delta + b \Delta W_n \quad (11)$$

with $\Delta W_n = W_{\tau_{n+1}} - W_{\tau_n}$ and initial value $Y_0 = X_0$, where $\tau_n = n\Delta$, $\Delta > 0$. Here and in the following, we suppress in our notation of coefficient functions, as a and b , the dependence on τ_n and Y_n .

The Euler scheme (11) corresponds to the truncated Wagner–Platen expansion that contains only the time integral and the single Itô integral with respect to the Wiener process. The Euler scheme has order of weak convergence $\beta = 1.0$ if the drift and diffusion coefficient are sufficiently smooth and regular.

For weak convergence, we only need to approximate the probability measure induced by the process X . Here, we can replace the Gaussian increments ΔW_n in (11) by other simpler random variables $\Delta \hat{W}_n$ with similar moment properties as ΔW_n . We can thus obtain a simpler scheme by choosing more easily generated random variables. This leads to the *simplified weak Euler scheme*

$$Y_{n+1} = Y_n + a \Delta + b \Delta \hat{W}_n \quad (12)$$

where the $\Delta \hat{W}_n$ are independent random variables with moments satisfying the moment-matching condition

$$\begin{aligned} & \left| E(\Delta \hat{W}_n) \right| + \left| E\left((\Delta \hat{W}_n)^3 \right) \right| \\ & + \left| E\left((\Delta \hat{W}_n)^2 \right) - \Delta \right| \leq K \Delta^2 \end{aligned} \quad (13)$$

for some constant K . The simplest example of such a simplified random variable $\Delta \hat{W}_n$ to be used in equation (12) is a two-point distributed random variable with

$$P(\Delta \hat{W}_n = \pm \sqrt{\Delta}) = 1/2 \quad (14)$$

When the drift and diffusion coefficients are only Hölder continuous, then it has been shown in [8] that the Euler scheme still converges weakly, but with some weak order $\beta < 1.0$.

Weak Order 2.0 Taylor Scheme

Now, let us consider the Taylor scheme of weak order $\beta = 2.0$. This scheme is obtained by adding all of the double stochastic integrals from the Wagner–Platen expansion to the terms of the Euler scheme, as shown in **Stochastic Taylor Expansions**. This scheme was

first proposed in [9]. The *weak order 2.0 Taylor scheme* has the form

$$\begin{aligned} Y_{n+1} = & Y_n + a \Delta + b \Delta W_n + \frac{1}{2} b b' \{(\Delta W_n)^2 - \Delta\} \\ & + a' b \Delta Z_n + \frac{1}{2} \left(a a' + \frac{1}{2} a'' b^2 \right) \Delta^2 \\ & + \left(a b' + \frac{1}{2} b'' b^2 \right) \{ \Delta W_n \Delta - \Delta Z_n \} \end{aligned} \quad (15)$$

The random variable $\Delta Z_n = \int_{\tau_n}^{\tau_{n+1}} \int_{\tau_n}^{s_2} dW_{s_1} ds_2$ represents a stochastic double integral. One can easily generate the pair of correlated Gaussian random variables ΔW_n and ΔZ_n from independent Gaussian random variables.

Under the weak convergence criterion, one has more freedom than under the strong convergence criterion (see **Stochastic Differential Equations: Scenario Simulation**) in constructing the random variables in a discrete-time weak approximation. For instance, from the above scheme, one can derive the *simplified weak order 2.0 Taylor scheme*

$$\begin{aligned} Y_{n+1} = & Y_n + a \Delta + b \Delta \hat{W}_n \\ & + 1/2 b b' \left\{ (\Delta \hat{W}_n)^2 - \Delta \right\} \\ & + 1/2 (a' b + a b' + 1/2 b'' b^2) \Delta \hat{W}_n \Delta \\ & + 1/2 (a a' + 1/2 a'' b^2) \Delta^2 \end{aligned} \quad (16)$$

Here the simplified random variable $\Delta \hat{W}_n$ must satisfy the moment-matching condition

$$\begin{aligned} & \left| E(\Delta \hat{W}_n) \right| + \left| E\left((\Delta \hat{W}_n)^3 \right) \right| \\ & + \left| E\left((\Delta \hat{W}_n)^5 \right) \right| + \left| E\left((\Delta \hat{W}_n)^2 \right) - \Delta \right| \\ & + \left| E\left((\Delta \hat{W}_n)^4 \right) - 3\Delta^2 \right| \leq K \Delta^3 \end{aligned} \quad (17)$$

for some constant K . For instance, an $N(0, \Delta)$ Gaussian distributed random variable satisfies the

condition (17). So also does a three-point distributed random variable $\Delta \hat{W}_n$ with

$$\begin{aligned} P(\Delta \hat{W}_n = \pm \sqrt{3\Delta}) &= \frac{1}{6} \quad \text{and} \\ P(\Delta \hat{W}_n = 0) &= \frac{2}{3} \end{aligned} \quad (18)$$

Under appropriate conditions on the drift and diffusion coefficients, the scheme converges with weak order 2.0 [6].

Weak Order 3.0 Taylor Scheme

As shown in [6], Taylor schemes of weak order $\beta = 3.0$ need to include from the Wagner–Platen expansion all of the multiple Itô integrals of up to multiplicity three. The following *simplified weak order 3.0 Taylor scheme* can be obtained

$$\begin{aligned} Y_{n+1} = & Y_n + a \Delta + b \Delta \tilde{W}_n + \frac{1}{2} L^1 b \left\{ (\Delta \tilde{W}_n)^2 - \Delta \right\} \\ & + L^1 a \Delta \tilde{Z}_n + \frac{1}{2} L^0 a \Delta^2 + L^0 b \\ & \times \left\{ \Delta \tilde{W}_n \Delta - \Delta \tilde{Z}_n \right\} \\ & + \frac{1}{6} (L^0 L^0 b + L^0 L^1 a + L^1 L^0 a) \Delta \tilde{W}_n \Delta^2 \\ & + \frac{1}{6} (L^1 L^1 a + L^1 L^0 b + L^0 L^1 b) \\ & \times \left\{ (\Delta \tilde{W}_n)^2 - \Delta \right\} \Delta \\ & + \frac{1}{6} L^0 L^0 a \Delta^3 + \frac{1}{6} L^1 L^1 b \\ & \times \left\{ (\Delta \tilde{W}_n)^2 - 3\Delta \right\} \Delta \tilde{W}_n \end{aligned} \quad (19)$$

Here $\Delta \tilde{W}_n$ and $\Delta \tilde{Z}_n$ are correlated Gaussian random variables with

$$\Delta \tilde{W}_n \sim N(0, \Delta), \quad \Delta \tilde{Z}_n \sim N(0, 1/3 \Delta^3) \quad (20)$$

and covariance

$$E(\Delta \tilde{W}_n \Delta \tilde{Z}_n) = 1/2 \Delta^2 \quad (21)$$

Furthermore, we use here the operators

$$L^0 = \frac{\partial}{\partial t} + a \frac{\partial}{\partial x} + \frac{1}{2} b^2 \frac{\partial^2}{\partial x^2} \quad (22)$$

and $L^1 = b \frac{\partial}{\partial x}$

Weak Order 4.0 Taylor Scheme

To construct the weak order 4.0 Taylor scheme, we also need to include all of the fourth-order multiple stochastic integrals from the Wagner–Platen expansion.

In the case of particular SDEs, for instance those with additive noise, one obtains highly accurate schemes in this manner. For accurate Monte Carlo simulation, the following *simplified weak order 4.0 Taylor scheme for additive noise* can be used:

$$\begin{aligned} Y_{n+1} = & Y_n + a \Delta + b \Delta \tilde{W}_n + \frac{1}{2} L^0 a \Delta^2 \\ & + L^1 a \Delta \tilde{Z}_n + L^0 b \left\{ \Delta \tilde{W}_n \Delta - \Delta \tilde{Z}_n \right\} \\ & + \frac{1}{3!} \left\{ L^0 L^0 b + L^0 L^1 a \right\} \Delta \tilde{W}_n \Delta^2 \\ & + L^1 L^1 a \left\{ 2 \Delta \tilde{W}_n \Delta \tilde{Z}_n - \frac{5}{6} \left(\Delta \tilde{W}_n \right)^2 \Delta \right. \\ & \quad \left. - \frac{1}{6} \Delta^2 \right\} \\ & + \frac{1}{3!} L^0 L^0 a \Delta^3 + \frac{1}{4!} L^0 L^0 L^0 a \Delta^4 \\ & + \frac{1}{4!} \left\{ L^1 L^0 L^0 a + L^0 L^1 L^0 a \right. \\ & \quad \left. + L^0 L^0 L^1 a + L^0 L^0 L^0 b \right\} \Delta \tilde{W}_n \Delta^3 \\ & + \frac{1}{4!} \left\{ L^1 L^1 L^0 a + L^0 L^1 L^1 a + L^1 L^0 L^1 a \right\} \\ & \times \left\{ \left(\Delta \tilde{W}_n \right)^2 - \Delta \right\} \Delta^2 \\ & + \frac{1}{4!} L^1 L^1 L^1 a \Delta \tilde{W}_n \left\{ \left(\Delta \tilde{W}_n \right)^2 - 3 \Delta \right\} \Delta \end{aligned} \quad (23)$$

Here $\Delta \tilde{W}_n$ and $\Delta \tilde{Z}_n$ are correlated Gaussian random variables with $\Delta \tilde{W}_n \sim N(0, \Delta)$, $\Delta \tilde{Z}_n \sim N(0, \frac{1}{3} \Delta^3)$ and $E(\Delta \tilde{W}_n \Delta \tilde{Z}_n) = \frac{1}{2} \Delta^2$. The weak order of convergence of the above schemes is derived in [6].

Explicit Weak Schemes

Higher order weak Taylor schemes require the evaluation of derivatives of various orders of the drift and diffusion coefficients. We can construct derivative-free discrete-time weak approximations, which avoid the use of such derivatives.

The following *explicit weak order 2.0 scheme* was suggested by Platen

$$\begin{aligned} Y_{n+1} = & Y_n + 1/2 \left(a \left(\tilde{\Upsilon}_n \right) + a \right) \Delta \\ & + 1/4 \left(b \left(\tilde{\Upsilon}_n^+ \right) + b \left(\tilde{\Upsilon}_n^- \right) + 2b \right) \Delta \hat{W}_n \\ & + 1/4 \left(b \left(\tilde{\Upsilon}_n^+ \right) - b \left(\tilde{\Upsilon}_n^- \right) \right) \\ & \times \left\{ \left(\Delta \hat{W}_n \right)^2 - \Delta \right\} \Delta^{1/2} \end{aligned} \quad (24)$$

with supporting values

$$\tilde{\Upsilon}_n = Y_n + a \Delta + b \Delta \hat{W}_n \quad (25)$$

and

$$\tilde{\Upsilon}_n^\pm = Y_n + a \Delta \pm b \sqrt{\Delta} \quad (26)$$

Here $\Delta \hat{W}_n$ is required to satisfy the moment condition (17). This means, $\Delta \hat{W}_n$ can be, for instance, Gaussian or three-point distributed with

$$\begin{aligned} P \left(\Delta \hat{W}_n = \pm \sqrt{3\Delta} \right) &= 1/6 \quad \text{and} \\ P \left(\Delta \hat{W}_n = 0 \right) &= 2/3 \end{aligned} \quad (27)$$

By comparing equation (24) with the corresponding simplified weak Taylor scheme (16), one notes that equation (24) avoids the derivatives that appear in equation (16) by using additional supporting values.

For *additive noise* the second-order weak scheme (24) reduces to the relatively simple algorithm

$$\begin{aligned} Y_{n+1} = & Y_n + 1/2 \left\{ a \left(Y_n + a \Delta + b \Delta \hat{W}_n^j \right) + a \right\} \Delta \\ & + b \Delta \hat{W}_n^j \end{aligned} \quad (28)$$

For the case with additive noise, one finds in [6] the *explicit weak order 3.0 scheme*

$$\begin{aligned}
 Y_{n+1} = & Y_n + a \Delta + b \Delta \hat{W}_n \\
 & + \frac{1}{2} \left(a_{\zeta_n}^+ + a_{\zeta_n}^- - \frac{3}{2} a - \frac{1}{4} (\tilde{a}_{\zeta_n}^+ + \tilde{a}_{\zeta_n}^-) \right) \Delta \\
 & + \sqrt{\frac{2}{\Delta}} \left(\frac{1}{\sqrt{2}} (a_{\zeta_n}^+ - a_{\zeta_n}^-) \right. \\
 & \left. - \frac{1}{4} (\tilde{a}_{\zeta_n}^+ - \tilde{a}_{\zeta_n}^-) \right) \zeta_n \Delta \hat{Z}_n \\
 & + \frac{1}{6} \left[a \left(Y_n + (a + a_{\zeta_n}^+) \Delta + (\zeta_n + \varrho_n) b \sqrt{\Delta} \right) \right. \\
 & \left. - a_{\zeta_n}^+ - a_{\varrho_n}^+ + a \right] \\
 & \times \left[(\zeta_n + \varrho_n) \Delta \hat{W}_n \sqrt{\Delta} + \Delta \right. \\
 & \left. + \zeta_n \varrho_n \left\{ (\Delta \hat{W}_n)^2 - \Delta \right\} \right] \quad (29)
 \end{aligned}$$

with

$$a_{\phi}^{\pm} = a \left(Y_n + a \Delta \pm b \sqrt{\Delta} \phi \right) \quad (30)$$

and

$$\tilde{a}_{\phi}^{\pm} = a \left(Y_n + 2a \Delta \pm b \sqrt{2\Delta} \phi \right) \quad (31)$$

where ϕ is either ζ_n or ϱ_n . Here, one can use two correlated Gaussian random variables $\Delta \hat{W}_n \sim N(0, \Delta)$ and $\Delta \hat{Z}_n \sim N(0, 1/3 \Delta^3)$ with $E(\Delta \hat{W}_n \Delta \hat{Z}_n) = 1/2\Delta^2$, together with two independent two-point distributed random variables ζ_n and ϱ_n with

$$P(\zeta_n = \pm 1) = P(\varrho_n = \pm 1) = 1/2 \quad (32)$$

Extrapolation Methods

Extrapolation provides an efficient, yet simple way of obtaining a higher order weak approximation when using only lower order weak schemes. Only equidistant time discretizations of the time interval $[0, T]$ with $\tau_{n_T} = T$ are used in what follows. As before, we denote the considered discrete-time approximation with time step size $\Delta > 0$ by Y^Δ , with value $Y_{\tau_n}^\Delta = Y_n^\Delta$ at the discretization time τ_n , and the corresponding approximation with twice this step size by $Y^{2\Delta}$, and so on.

Suppose that we have evaluated *via* simulation the functional

$$E(g(Y_T^\Delta))$$

of a weak order 1.0 approximation using, say, the Euler scheme (11) or the simplified Euler scheme (12) with step size Δ . Let us repeat this Monte Carlo simulation with the double step size 2Δ to obtain a Monte Carlo estimate of the functional

$$E(g(Y_T^{2\Delta}))$$

We can then combine these two functionals to obtain the *weak order 2.0 extrapolation*

$$V_{g,2}^\Delta(T) = 2E(g(Y_T^\Delta)) - E(g(Y_T^{2\Delta})) \quad (33)$$

which was proposed in [13]. It is a stochastic generalization of the well-known *Richardson extrapolation*.

As is shown in [6], if a weak method exhibits a certain representation of the leading error term, then a corresponding extrapolation method can be constructed. For instance, one can use a weak order $\beta = 2.0$ approximation Y^Δ and extrapolate it to obtain a fourth-order weak approximation of the targeted functional. A *weak order 4.0 extrapolation* has the form

$$\begin{aligned}
 V_{g,4}^\Delta(T) = & 1/21 \left[32 E(g(Y_T^\Delta)) \right. \\
 & \left. - 12 E(g(Y_T^{2\Delta})) + E(g(Y_T^{4\Delta})) \right] \quad (34)
 \end{aligned}$$

Suitable weak order 2.0 approximations include the weak order 2.0 Taylor scheme (15), the simplified weak order 2.0 Taylor scheme (16), and the explicit weak order 2.0 scheme (24).

The practical use of extrapolations of discrete time approximations depends strongly on the numerical stability of the underlying weak schemes. These weak methods need to have almost identical leading error coefficients for a wide range of step sizes and should yield numerically stable simulation results; see **Stochastic Differential Equations: Scenario Simulation**.

Implicit Methods

In Monte Carlo simulation, the numerical stability of a scheme has highest priority. Introducing some

type of implicitness into a scheme usually improves its numerical stability. The simplest implicit weak schemes can be found in the *family of drift implicit simplified Euler schemes*

$$Y_{n+1} = Y_n + \{(1 - \alpha) a(Y_n) + \alpha a(Y_{n+1})\} \Delta + b(Y_n) \Delta \hat{W}_n \quad (35)$$

where the random variables $\Delta \hat{W}_n$ are independent two-point distributed with

$$P(\Delta \hat{W}_n = \pm \sqrt{\Delta}) = 1/2 \quad (36)$$

The parameter α is the *degree of drift implicitness*. With $\alpha = 0$, the scheme (35) reduces to the simplified Euler scheme (12), whereas with $\alpha = 0.5$ it represents a stochastic generalization of the trapezoidal method. Under sufficient regularity conditions, one can show that the scheme (35) converges with weak order $\beta = 1.0$. The scheme (35) is A -stable for $\alpha \in [0.5, 1]$, whereas for $\alpha \in [0, 0.5]$ its region of A -stability, in the sense of what is discussed in **Stochastic Differential Equations: Scenario Simulation**, is the interior of the interval that begins at $-2(1 - 2\alpha)^{-1}$ and finishes at 0.

The possible use of bounded random variables in weak simplified schemes allows us to construct fully implicit weak schemes that is, algorithms where also the approximate diffusion term becomes implicit.

The *fully implicit weak Euler scheme* has the form

$$Y_{n+1} = Y_n + \bar{a}(Y_{n+1}) \Delta + b(Y_{n+1}) \Delta \hat{W}_n \quad (37)$$

where $\Delta \hat{W}_n$ is as in equation (35) and $\bar{a}$ is some adjusted drift coefficient defined by

$$\bar{a} = a - b b' \quad (38)$$

The drift adjustment is necessary, otherwise the approximation would not converge toward the correct solution of the given Itô SDE.

We also mention a *family of implicit weak Euler schemes*

$$Y_{n+1} = Y_n + \{\alpha \bar{a}_\eta(Y_{n+1}) + (1 - \alpha) \bar{a}_\eta(Y_n)\} \Delta + \{\eta b(Y_{n+1}) + (1 - \eta) b(Y_n)\} \Delta \hat{W}_n \quad (39)$$

where the random variables $\Delta \hat{W}_n$ are as in equation (35) and the corrected drift coefficient $\bar{a}_\eta$ is defined

as

$$\bar{a}_\eta = a - \eta b \frac{\partial b}{\partial x} \quad (40)$$

for $\alpha, \eta \in [0, 1]$.

One can avoid the calculation of derivatives in the above family of implicit schemes by using differences instead. The following *implicit weak order 2.0 scheme* can be found in [11], where

$$Y_{n+1} = Y_n + 1/2 (a + a(Y_{n+1})) \Delta + 1/4 (b(\tilde{Y}_n^+) + b(\tilde{Y}_n^-) + 2b) \Delta \hat{W}_n + 1/4 (b(\tilde{Y}_n^+) - b(\tilde{Y}_n^-)) \times \left\{ (\Delta \hat{W}_n)^2 - \Delta \right\} \Delta^{-1/2} \quad (41)$$

with supporting values

$$\tilde{Y}_n^\pm = Y_n + a \Delta \pm b \sqrt{\Delta} \quad (42)$$

Here, the random variable $\Delta \hat{W}_n$ can be chosen as in (18).

Note that the scheme (39) is A -stable. In [6], it is shown that the above second-order weak scheme converges under appropriate conditions with weak order $\beta = 2.0$.

Weak Predictor–Corrector Methods

In general, implicit schemes require an algebraic equation to be solved at each time step. This imposes an additional computational burden. However, without giving a weak scheme some kind of implicitness the simulation might not turn out to be of much practical use due to inherent numerical instabilities.

Predictor–corrector methods are similar to implicit methods but do not require the solution of an algebraic equation at each time step. They are used mainly because of their good numerical stability properties, which they inherit from the implicit counterparts of their corrector. The following predictor–corrector methods can be found in [11].

One has the following *family of weak order 1.0 predictor–corrector methods* with corrector

$$Y_{n+1} = Y_n + \{\alpha \bar{a}_\eta(\bar{Y}_{n+1}) + (1 - \alpha) \bar{a}_\eta(Y_n)\} \Delta + \{\eta b(\bar{Y}_{n+1}) + (1 - \eta) b(Y_n)\} \Delta \hat{W}_n \quad (43)$$

for $\alpha, \eta \in [0, 1]$, where

$$\bar{a}_\eta = a - \eta b \frac{\partial b}{\partial x} \quad (44)$$

and with predictor

$$\bar{Y}_{n+1} = Y_n + a \Delta + b \Delta \hat{W}_n \quad (45)$$

Here, the random variables $\Delta \hat{W}_n$ are as in (14). Note that the corrector (43) with $\eta > 0$ allows to include some implicitness in the diffusion terms. This scheme often provides efficient and numerically reliable methods for appropriate choices of α and η . By performing the Monte Carlo simulation with different parameter choices for α and η one can obtain useful information about the numerical stability of the scheme for the given application.

A *weak order 2.0 predictor–corrector method* is obtained by choosing as corrector

$$Y_{n+1} = Y_n + 1/2 \{a(\bar{Y}_{n+1}) + a\} \Delta + \Psi_n \quad (46)$$

with

$$\begin{aligned} \Psi_n = & b \Delta \hat{W}_n + 1/2 b b' \left\{ (\Delta \hat{W}_n)^2 - \Delta \right\} \\ & + 1/2 \left(a b' + \frac{1}{2} b^2 b'' \right) \Delta \hat{W}_n \Delta \end{aligned} \quad (47)$$

and as predictor

$$\begin{aligned} \bar{Y}_{n+1} = & Y_n + a \Delta + \Psi_n + 1/2 a' b \Delta \hat{W}_n \Delta \\ & + 1/2 (a a' + 1/2 a'' b^2) \Delta^2 \end{aligned} \quad (48)$$

Here the random variable $\Delta \hat{W}_n$ can be, for instance, $N(0, \Delta)$ Gaussian or three-point distributed as in equation (18).

Another *derivative-free weak order 2.0 predictor–corrector method* has corrector

$$Y_{n+1} = Y_n + 1/2 \{a(\bar{Y}_{n+1}) + a\} \Delta + \phi_n \quad (49)$$

where

$$\begin{aligned} \phi_n = & 1/4 (b(\bar{\Upsilon}_n^+) + b(\bar{\Upsilon}_n^-) + 2b) \Delta \hat{W}_n \\ & + 1/4 (b(\bar{\Upsilon}_n^+) - b(\bar{\Upsilon}_n^-)) \\ & \times \left\{ (\Delta \hat{W}_n)^2 - \Delta \right\} \Delta^{-1/2} \end{aligned} \quad (50)$$

with supporting values

$$\bar{\Upsilon}_n^\pm = Y_n + a \Delta \pm b \sqrt{\Delta} \quad (51)$$

and with predictor

$$\bar{Y}_{n+1} = Y_n + 1/2 \{a(\bar{\Upsilon}_n) + a\} \Delta + \phi_n \quad (52)$$

using the supporting value

$$\bar{\Upsilon}_n = Y_n + a \Delta + b \Delta \hat{W}_n \quad (53)$$

Here the random variable $\Delta \hat{W}_n$ can be chosen as in equation (18).

Predictor–corrector methods of the above kind have been successfully used in the Monte Carlo simulation of various asset price models, see, for instance, [4].

References

- [1] Fishman, G.S. (1996). *Monte Carlo: Concepts, Algorithms and Applications*, Springer.
- [2] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering, Applied Mathematics*, Springer, Vol. 53.
- [3] Hammersley, J.M. & Handscomb, D.C. (1964). *Monte Carlo Methods*, Methuen, London.
- [4] Hunter, C.J., Jäckel, P. & Joshi, M.S. (2001). Getting the drift, *Risk* **14**(7), 81–84.
- [5] Jäckel, P. (2002). *Monte Carlo Methods in Finance*, John Wiley & Sons.
- [6] Kloeden, P.E. & Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations, Applied Mathematics*, Springer, Vol. 23 (Third Printing).

- [7] Kloeden, P.E., Platen, E. & Schurz, H. (2003). *Numerical Solution of SDE's Through Computer Experiments*, Universitext, Springer (Third Corrected Printing).
- [8] Mikulevicius, R. & Platen, E. (1991). Rate of convergence of the Euler approximation for diffusion processes, *Mathematische Nachrichten* **151**, 233–239.
- [9] Milstein, G.N. (1978). A method of second order accuracy integration of stochastic differential equations, *Theory of Probability and its Applications* **23**, 396–401.
- [10] Milstein, G.N. (1995). *Numerical Integration of Stochastic Differential Equations*, Mathematics and Its Applications, Kluwer.
- [11] Platen, E. (1995). On weak implicit and predictor–corrector methods, *Mathematics and Computers in Simulation* **38**, 69–76.
- [12] Platen, E. & Heath, D. (2006). *A Benchmark Approach to Quantitative Finance*, Springer.
- [13] Talay, D. & Tubaro, L. (1990). Expansion of the global error for numerical schemes solving stochastic differential equations, *Stochastic Analysis and Applications* **8**(4), 483–509.

Related Articles

Backward Stochastic Differential Equations: Numerical Methods; LIBOR Market Model; LIBOR Market Models: Simulation; Pseudorandom Number Generators; Simulation of Square-root Processes; Stochastic Differential Equations: Scenario Simulation; Stochastic Differential Equations with Jumps: Simulation; Stochastic Integrals; Stochastic Taylor Expansions; Variance Reduction.

NICOLA BRUTI-LIBERATI & ECKHARD PLATEN

Stochastic Differential Equations: Scenario Simulation

Discrete-time Approximation

We have explicit solutions only in a few cases of *stochastic differential equation* (SDEs), including linear SDEs and their transformations. In finance, it is often helpful, and in some situations essential, to be able to simulate accurate trajectories of solutions of SDEs that model the financial quantities under consideration. These scenarios are typically generated by algorithms that use pseudorandom number generators, as introduced in **Pseudorandom Number Generators**. However, one can use also historical returns, log-returns, or increments from observed asset prices as input in a simulation that is then called a *historical simulation*. In the following, we give a basic introduction into *scenario simulation*. Here, we approximate the path of a solution of a given SDE by the path of a discrete-time approximation, simulated by using random number generators, *see Pseudorandom Number Generators*. For books on scenario simulation of solutions of SDEs we refer to [7, 8]. There exists also an increasing literature on simulation methods in finance. Important monographs that the reader can be referred to include [5, 6].

Consider a given discretization $0 = \tau_0 < \tau_1 < \dots < \tau_n < \dots < \tau_N = T$ of the time interval $[0, T]$. We shall approximate a diffusion process $X = \{X_t, t \in [0, T]\}$ satisfying the one-dimensional SDE

$$dX_t = a(t, X_t) dt + b(t, X_t) dW_t \quad (1)$$

on $t \in [0, T]$ with initial value $X_0 \in \mathbb{R}$.

Let us first introduce one of the simplest discrete-time approximations, the *Euler scheme* or *Euler-Maruyama scheme*, see [9]. An *Euler approximation* is a continuous-time stochastic process $Y^\Delta = \{Y_t^\Delta, t \in [0, T]\}$ satisfying the recursive scheme

$$Y_{n+1} = Y_n + a(\tau_n, Y_n) (\tau_{n+1} - \tau_n) + b(\tau_n, Y_n) (W_{\tau_{n+1}} - W_{\tau_n}) \quad (2)$$

for $n \in \{0, 1, \dots\}$ with initial value

$$Y_0 = Y_0^\Delta = X_0 \quad (3)$$

For convenience, we write

$$Y_n = Y_{\tau_n}^\Delta \quad (4)$$

for the value of the approximation at the discretization time τ_n . We write

$$\Delta_n = \tau_{n+1} - \tau_n \quad (5)$$

for the n th increment of the time discretization and call

$$\Delta = \max_{n \in \{0, 1, \dots, N-1\}} \Delta_n \quad (6)$$

the *maximum step size*. We consider here, for simplicity, *equidistant time discretizations* with

$$\tau_n = n \Delta \quad (7)$$

for $\Delta_n = \Delta = T/N$ and some integer N large enough so that $\Delta \in (0, 1)$.

The sequence $(Y_n)_{n \in \{0, 1, \dots, N\}}$ of values of the Euler approximation (2) at the instants of the time discretization $\tau_0, \tau_1, \dots, \tau_N$ can be recursively computed. For this purpose, we need to generate the random increments

$$\Delta W_n = W_{\tau_{n+1}} - W_{\tau_n} \quad (8)$$

for $n \in \{0, 1, \dots, N-1\}$ of the Wiener process $W = \{W_t, t \in [0, T]\}$. These increments represent independent Gaussian distributed random variables with mean

$$E(\Delta W_n) = 0 \quad (9)$$

and variance

$$E((\Delta W_n)^2) = \Delta \quad (10)$$

For the increments (8) of the Wiener process we can use a sequence of independent Gaussian pseudorandom numbers. These can be generated, for instance, as suggested in **Pseudorandom Number Generators**.

For describing a numerical scheme efficiently, we typically use the abbreviation

$$f = f(\tau_n, Y_n) \quad (11)$$

for a function f defined on $[0, T] \times \mathbb{R}^d$ and $n \in \{0, 1, \dots, N-1\}$, when no misunderstanding is possible. We can then rewrite the Euler scheme (2) in the form

$$Y_{n+1} = Y_n + a \Delta + b \Delta W_n \quad (12)$$

for $n \in \{0, 1, \dots, N-1\}$. Usually, we suppress the mentioning of the initial condition, where we typically set

$$Y_0 = X_0 \quad (13)$$

The recursive structure of the Euler scheme, which generates approximate values of the diffusion process X at the discretization times only, is the key to its implementation.

Interpolation

We emphasize that we shall consider a discrete-time approximation to be a stochastic process defined on the interval $[0, T]$. Although it will be often sufficient to generate its values at the discretization times, if required, values at the intermediate instants can be determined by an appropriate interpolation method. The simplest one is the *piecewise constant interpolation* with

$$Y_t^\Delta = Y_{n_t} \quad (14)$$

for $t \in [0, T]$. Here n_t is the integer defined by

$$n_t = \max\{n \in \{0, 1, \dots, N : \tau_n \leq t\}\} \quad (15)$$

which is the largest integer n for which τ_n does not exceed t .

Furthermore, the *linear interpolation*

$$Y_t^\Delta = Y_{n_t} + \frac{t - \tau_{n_t}}{\tau_{n_t+1} - \tau_{n_t}} (Y_{n_t+1} - Y_{n_t}) \quad (16)$$

for $t \in [0, T]$ is often used in the visualization of trajectories because it provides continuous sample paths in a convenient and realistic manner.

In general, the sample paths of a diffusion process inherit some properties of the trajectory of its driving Wiener process, in particular, its nondifferentiability. It is impossible to fully reproduce on a digital computer the microstructure of such path in a scenario simulation. Thus, we shall concentrate on simulating values of a discrete-time approximation at given discretization times and interpolate these appropriately if needed.

Simulating Geometric Brownian Motion

To illustrate various aspects of a *scenario simulation via* a discrete-time approximation of a diffusion process, it is useful to examine, in some detail, a simple

but important example with an explicit solution. In finance, one often faces growth processes. These are under the standard market model usually interpreted as geometric Brownian motions. Let us consider the geometric Brownian motion $X = \{X_t, t \in [0, T]\}$ satisfying the linear SDE

$$dX_t = a X_t dt + b X_t dW_t \quad (17)$$

for $t \in [0, T]$ with initial value $X_0 > 0$. This model is also known as the *Black–Scholes model*. The geometric Brownian motion X is a diffusion process with *drift coefficient*

$$a(t, x) = a x \quad (18)$$

and *diffusion coefficient*

$$b(t, x) = b x \quad (19)$$

Here a denotes the *appreciation rate* and $b \neq 0$ the *volatility*. It is well known that the SDE (17) has the explicit solution

$$X_t = X_0 \exp\{(a - 1/2 b^2)t + b W_t\} \quad (20)$$

for $t \in [0, T]$ and given Wiener process $W = \{W_t, t \in [0, T]\}$. The availability of an explicit solution makes the simulation an easy task because the solution of the SDE (17) is simply the exponential function (20) of the corresponding Wiener process value.

Simulation of Approximate Trajectories

Knowing the solution (20) of the SDE (17) explicitly gives us the possibility of comparing the Euler approximation Y , see equation (12), with the exact solution X . Of course, in general, this is not possible.

To simulate a trajectory of the Euler approximation for a given time discretization, we start from the initial value $Y_0 = X_0$, and proceed recursively by generating the next value

$$Y_{n+1} = Y_n + a Y_n \Delta + b Y_n \Delta W_n \quad (21)$$

for $n \in \{0, 1, \dots, N-1\}$ according to the Euler scheme (12) with drift coefficient (18) and diffusion coefficient (19). Here

$$\Delta W_n = W_{\tau_{n+1}} - W_{\tau_n} \quad (22)$$

see equation (8), is the $N(0, \Delta)$ Gaussian distributed increment of the Wiener process W over the interval $[\tau_n, \tau_n + \Delta]$, which we generate by a pseudorandom number generator, *see Pseudorandom Number Generators*.

For comparison, we can use the equation (20) for the given Black–Scholes dynamics to determine the corresponding exact values of the solution. This uses the same sample path of the Wiener process in equation (20), as employed by the Euler scheme (21), where one obtains as explicit solution at time τ_n the value

$$X_{\tau_n} = X_0 \exp \left\{ (a - 1/2 b^2) \tau_n + b \sum_{i=1}^n \Delta W_{i-1} \right\} \quad (23)$$

for $n \in \{0, 1, \dots, N-1\}$.

One needs always to be careful when using a discrete-time approximation such as the Euler scheme (21). This is clearly a different object than the exact solution of the underlying SDE. For instance, inconsistencies may arise because the increments in the noise part of the Euler scheme can take extremely large values of either sign. Even though this can occur only with small probability, large negative stochastic increments can, for instance, make the Euler approximation (21) negative. This would be not consistent with our aim of simulating positive asset prices via the Black–Scholes model. The possibility of generating negative asset prices is quite a serious defect in the Euler method when directly applied for the Black–Scholes dynamics. For extremely small step sizes, this phenomenon becomes highly unlikely; however, it is not excluded unless one designs the simulation in a way that guarantees positivity. For the above Black–Scholes dynamics, this can be achieved by simulating first the logarithm of the geometric Brownian motion by using a corresponding Euler scheme and deriving then the approximate value of X as the exponential of the approximated logarithm. In the case of constant volatility and appreciation rate, this even yields the exact solution of the SDE at the discretization points.

Order of Strong Convergence

So far we have not specified a criterion for the classification of the overall accuracy of a discrete-time

approximation with respect to vanishing time step size. Such a criterion should reflect the main goal of a scenario simulation, which applies in situations where a *pathwise approximation* is required. For instance, scenario simulation is needed for the visualization of typical trajectories of financial models, the testing of calibration methods and statistical estimators, and also in filtering hidden variables. Monte Carlo simulation is discussed in **Monte Carlo Simulation for Stochastic Differential Equations**, which is different since it focuses on *approximating probabilities* and functionals of the underlying processes.

One can estimate the error of an approximation using the following *absolute error criterion*. For a given step size Δ , it is defined as the expectation of the absolute difference between the discrete-time approximation $Y_N = Y_T^\Delta$ and the solution X_T of the SDE at the time horizon $T = N\Delta$, that is,

$$\varepsilon(\Delta) = E(|X_T - Y_T^\Delta|) \quad (24)$$

This gives some measure for the pathwise closeness at the end of the time interval $[0, T]$ as a function of the maximum step size Δ of the time discretization.

Although the Euler approximation is one of the simplest discrete-time approximations, it is, generally, not very efficient numerically. In the following we shall, therefore, derive and investigate also other discrete-time approximations.

To classify different discrete-time approximations, we introduce their *order of strong convergence*. We shall say that a discrete-time approximation Y^Δ *converges strongly with order $\gamma > 0$* at time T if there exists a positive constant C , which does not depend on Δ , and a $\delta_0 > 0$ such that

$$\varepsilon(\Delta) = E(|X_T - Y_T^\Delta|) \leq C \Delta^\gamma \quad (25)$$

for each $\Delta \in (0, \delta_0)$. We call equation (25) the strong convergence criterion.

We emphasize that this criterion has been constructed for the classification of, so called, *strong approximations*. There exist in the literature results, see for instance, [7], which provide uniform error estimates and involve also higher moments than used by the criterion (25). For instance, the supremum of the squared difference between X_t and Y_t^Δ for $t \in [0, T]$ has been estimated. The criterion (25) for the order of strong convergence appears

to be natural and sufficient for the classification of schemes.

Euler Scheme

Let us now use Wagner–Platen expansions (*see Stochastic Taylor Expansions*), to derive discrete-time approximations with respect to the criterion (25). By appropriate truncation of the expansions in each discretization step, we obtain corresponding strong Taylor schemes of given strong order of convergence. For details, we refer to [7]. Recall that we usually suppress the dependence on τ_n and Y_n in the coefficients.

Let us begin with the already mentioned *Euler scheme* (21), which represents the simplest useful strong Taylor approximation. The *Euler scheme* has the form

$$Y_{n+1} = Y_n + a \Delta + b \Delta W_n \quad (26)$$

where $\Delta = \tau_{n+1} - \tau_n$ is the length of the time discretization subinterval $[\tau_n, \tau_{n+1}]$ and $\Delta W_n = W_{\tau_{n+1}} - W_{\tau_n}$ is an $N(0, \Delta)$ independent Gaussian distributed increment of the Wiener process W on $[\tau_n, \tau_{n+1}]$.

The Euler scheme corresponds to a truncated Wagner–Platen expansion that contains only single time and Wiener integrals. Assuming Lipschitz and linear growth conditions on the coefficient functions a and b , it can be shown that the Euler approximation is of strong order $\gamma = 0.5$.

Note that in special cases, this means for specific SDEs, the Euler scheme may achieve a higher order of strong convergence. For example, if the noise is *additive*, that is when the diffusion coefficient is a differentiable, deterministic function of time of the form

$$b(t, x) = b_t \quad (27)$$

then the Euler scheme achieves the order of strong convergence $\gamma = 1.0$.

The Euler scheme gives reasonable numerical results when the drift and diffusion coefficients are almost constant and the time step size is chosen to be sufficiently small. In general, however, it is not very satisfactory. The use of higher order and numerically stable schemes is recommended, which we discuss later.

Milstein Scheme

We now introduce an important scheme, the *Milstein scheme*, which was first suggested in [10] and turns out to be of strong order $\gamma = 1.0$. It is obtained by using one additional term from the Wagner–Platen expansions, *see Stochastic Taylor Expansions*.

The Milstein scheme has the form

$$Y_{n+1} = Y_n + a \Delta + b \Delta W_n + b b' 1/2 \times \{(\Delta W_n)^2 - \Delta\} \quad (28)$$

We remark that for more general SDEs, involving several driving Wiener processes, multiple stochastic integrals

$$I_{(j_1, j_2)} = \int_{\tau_n}^{\tau_{n+1}} \int_{\tau_n}^{s_1} dW_{s_2}^{j_1} dW_{s_1}^{j_2} \quad (29)$$

with $j_1 \neq j_2$ appear in the general version of the Milstein scheme. They cannot be simply expressed in terms of the increments $\Delta W_n^{j_1}$ and $\Delta W_n^{j_2}$ of the corresponding Wiener processes. Still, approximations are possible [4, 8]. However, in the special case $j_1 = j_2$ we have

$$I_{(j_1, j_1)} = 1/2 \{(\Delta W_n^{j_1})^2 - \Delta\} \quad (30)$$

which makes this double Wiener integral easy to generate.

For the Black–Scholes model and some other important dynamics in financial modeling, the corresponding SDEs have special properties which simplify the Milstein scheme. This avoids, in some cases, the use of double Wiener integrals of the type (29) involving two different Wiener processes. For instance, in the case of an SDE with *additive noise*, that is, when the diffusion coefficients depend at most on time t and not on the state variable, the Milstein scheme (28) reduces to the Euler scheme (26). Another important special case is that of *commutative noise*, *see* [7]. For commutative noise, the coefficient functions of the double Wiener integrals $I_{(j_1, j_2)}$ and $I_{(j_2, j_1)}$ are the same and one can exploit the fact that $I_{(j_1, j_2)} + I_{(j_2, j_1)} = \Delta W_n^{j_1} \Delta W_n^{j_2}$.

Order 1.5 Strong Taylor Scheme

There exist simulation tasks that require more accurate schemes than the Milstein scheme. For

instance, if one wants to capture extreme asset price movements or simply needs to be more accurate in a scenario simulation, then one may use higher order strong schemes. In general, we obtain more accurate strong Taylor schemes by including into the scheme further multiple stochastic integrals from the Wagner–Platen expansion (see **Stochastic Taylor Expansions**). Each of these multiple stochastic integrals contains additional information about the sample path of the driving Wiener process. The necessity of the inclusion of further multiple stochastic integrals for achieving higher orders of convergence is a fundamental feature of stochastic numerical methods solving SDEs.

The *order 1.5 strong Taylor scheme* is of the form

$$\begin{aligned} Y_{n+1} = & Y_n + a \Delta + b \Delta W_n \\ & + b b' 1/2 \{(\Delta W_n)^2 - \Delta\} + a' b \Delta Z_n \\ & + 1/2 \left(a a' + \frac{1}{2} b^2 a'' \right) \Delta^2 \\ & + (a b' + 1/2 b^2 b'') \{ \Delta W_n \Delta - \Delta Z_n \} \\ & + b (b b'' + (b')^2) 1/2 \\ & \times \left\{ \frac{1}{3} (\Delta W_n)^2 - \Delta \right\} \Delta W_n \end{aligned} \quad (31)$$

Here the additional random variable ΔZ_n is required, which represents the double integral

$$\Delta Z_n = I_{(1,0)} = \int_{\tau_n}^{\tau_{n+1}} \int_{\tau_n}^{s_2} dW_{s_1} ds_2 \quad (32)$$

One can show that ΔZ_n is Gaussian distributed with mean $E(\Delta Z_n) = 0$, variance $E((\Delta Z_n)^2) = \frac{1}{3} \Delta^3$, and covariance $E(\Delta Z_n \Delta W_n) = 1/2 \Delta^2$. Note that a pair of appropriately correlated Gaussian distributed random variables $(\Delta W_n, \Delta Z_n)$ can be easily constructed by linear combination of two independent $N(0, 1)$ standard Gaussian distributed random variables. All multiple stochastic integrals that appear in equation (31) can be expressed in terms of Δ , ΔW_n , and ΔZ_n . In particular, the last term in equation (31) contains the triple Wiener integral

$$I_{(1,1,1)} = 1/2 \{ 1/3 (\Delta W_n^1)^2 - \Delta \} \Delta W_n^1 \quad (33)$$

which reflects the third-order Hermite polynomial in ΔW_n^1 . The Taylor schemes of any desired strong order are described in [7].

An Explicit Order 1.0 Strong Scheme

Let us now consider strong schemes that avoid the use of derivatives similar to Runge–Kutta schemes for ordinary differential equations. Even though the resulting schemes look similar to Runge–Kutta methods, they cannot be simply constructed as heuristic adaptations of deterministic Runge–Kutta schemes. Appropriate *strong derivative-free schemes* have been systematically designed in [1, 2, 7].

Various first-order derivative-free schemes can be obtained from the Milstein scheme (28) simply by replacing the derivatives in the Milstein scheme by corresponding differences. The inclusion of these differences requires the computation of supporting values of the coefficients at additional supporting points. An example is given by the following scheme, called the *order 1.0 Platen scheme*, see [7], which is of the form

$$\begin{aligned} Y_{n+1} = & Y_n + a \Delta + b \Delta W_n + \frac{1}{2\sqrt{\Delta}} \\ & \times \{b(\tau_n, \tilde{\gamma}_n) - b\} \{(\Delta W_n)^2 - \Delta\} \end{aligned} \quad (34)$$

and uses the supporting value

$$\tilde{\gamma}_n = Y_n + a \Delta + b \sqrt{\Delta} \quad (35)$$

Multidimensional versions of this and most other schemes exist also for the case of several driving Wiener processes.

An Explicit Order 1.5 Strong Scheme

One can construct also derivative-free schemes of strong order $\gamma = 1.5$ by replacing the derivatives in the order 1.5 strong Taylor scheme by corresponding finite differences. The following *explicit order 1.5 strong scheme*, also called *order 1.5 Platen scheme*, has the form

$$\begin{aligned} Y_{n+1} = & Y_n + b \Delta W_n + \frac{1}{2\sqrt{\Delta}} \\ & \times \{a(\tilde{\gamma}_+) - a(\tilde{\gamma}_-)\} \Delta Z_n \\ & + \frac{1}{4} \{a(\tilde{\gamma}_+) + 2a + a(\tilde{\gamma}_-)\} \Delta \\ & + \frac{1}{4\sqrt{\Delta}} \{b(\tilde{\gamma}_+) - b(\tilde{\gamma}_-)\} \{(\Delta W_n)^2 - \Delta\} \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{2\Delta} \{b(\bar{\Upsilon}_+) - 2b + b(\bar{\Upsilon}_-)\} \\
& \times \{\Delta W_n \Delta - \Delta Z_n\} \\
& + \frac{1}{4\Delta} \left[b(\bar{\Phi}_+) - b(\bar{\Phi}_-) - b(\bar{\Upsilon}_+) + b(\bar{\Upsilon}_-) \right] \\
& \times \left\{ \frac{1}{3} (\Delta W_n)^2 - \Delta \right\} \Delta W_n
\end{aligned} \tag{36}$$

with

$$\bar{\Upsilon}_{\pm} = Y_n + a \Delta \pm b \sqrt{\Delta} \tag{37}$$

and

$$\bar{\Phi}_{\pm} = \bar{\Upsilon}_{\pm} \pm b(\bar{\Upsilon}_{\pm}) \sqrt{\Delta} \tag{38}$$

Here ΔZ_n is the double integral $I_{(1,0)}$ defined in equation (32). Further higher order explicit schemes with corresponding strong order of convergence can be found in [7].

Numerical Stability

Numerical stability is very important in simulation. It can be understood as the ability of a scheme to control the propagation of errors. Such errors naturally occur in almost any simulation on a computer due to its limited precision and as a result of truncation errors in the above schemes. However, numerical methods differ in their ability to dampen the arising errors. What really matters for a scheme is that it must be numerically stable when generating sufficiently accurate results over longer time periods. Before any properties of higher order convergence can be reasonably exploited, the question of numerical stability has to be satisfactorily answered. In particular, for solutions of SDEs with multiplicative noise, which are martingales, simulation studies have shown [11] that, with the above-described explicit schemes, numerical instabilities can easily occur.

The propagation of errors depends significantly on the specific nature of the diffusion coefficient of the SDE. For general SDEs and a given scheme, it is a delicate matter to provide reasonably useful answers with respect to their numerical stability. However, it is very helpful to study representative classes of test equations. Each particular type of SDEs usually has its own numerical challenges. Test SDEs can reveal typical numerical instabilities and allow the design of appropriate schemes. For SDEs with additive noise,

that is, SDEs with deterministic diffusion coefficients, the following concept of *A-stability* can be applied. We use here a real-valued test SDE with additive noise of the form

$$dX_t = \lambda X_t dt + dW_t \tag{39}$$

for $\lambda \in \mathbb{R}$. Obviously, X_t forms an Ornstein–Uhlenbeck process, which has a stationary density for $\lambda < 0$. A discrete-time approximation Y , when applied to the test equation (39), typically yields a recursive relation of the form

$$Y_{n+1} = G^A(\lambda \Delta) Y_n + Z_n \tag{40}$$

Here, the random term Z_n is assumed not to depend on $Y_0, Y_1, \dots, Y_n, Y_{n+1}$, and we call $G^A(\lambda \Delta)$ the *transfer function* of the scheme.

Now, the *A-stability region* of a given scheme is defined as the subset of the real axis consisting of those numbers $\lambda \Delta$, which are mapped by the *transfer function* $G^A(\cdot)$ into the unit interval $(-1, 1)$. These are those $\lambda \Delta$ for which

$$|G^A(\lambda \Delta)| < 1 \tag{41}$$

If, for a scheme, the *A-stability region* covers the left half of the real axis, that is the part formed by all $\lambda \Delta$ with $\lambda < 0$, then we say that the scheme is *A-stable*. In this case, the scheme achieves numerical stability for those parameters where the test dynamics themselves show stability. Since we have used additive noise in the test equation (39) we interpret *A-stability* here as *additive noise stability*. Let us now examine some implicit strong schemes.

Drift Implicit Euler and Milstein Schemes

The simplest implicit strong scheme is the *drift implicit Euler scheme*, which has strong order $\gamma = 0.5$. It has the form

$$Y_{n+1} = Y_n + a(\tau_{n+1}, Y_{n+1}) \Delta + b \Delta W_n \tag{42}$$

Here we follow our convention in writing $b = b(\tau_n, Y_n)$. There is also a *family of drift implicit Euler schemes*

$$\begin{aligned}
Y_{n+1} = & Y_n + \{\alpha a(\tau_{n+1}, Y_{n+1}) \\
& + (1 - \alpha) a\} \Delta + b \Delta W
\end{aligned} \tag{43}$$

where the parameter $\alpha \in [0, 1]$ characterizes the *degree of implicitness*. Note that for $\alpha = 0$ we have the explicit Euler scheme (26) and for $\alpha = 1$ the drift implicit Euler scheme (42). For $\alpha = 0.5$, we obtain from (43) a stochastic generalization of the deterministic trapezoidal method.

For the test equation (39), the family of drift implicit Euler schemes with degree of implicitness $\alpha \in [0, 1]$ yields the recursion formula

$$Y_{n+1} = Y_n + \{\alpha \lambda Y_{n+1} + (1 - \alpha) \lambda Y_n\} \Delta + \Delta W_n \quad (44)$$

and thus

$$Y_{n+1} = G^A(\lambda \Delta) Y_n + \Delta W_n (1 - \alpha \lambda \Delta)^{-1} \quad (45)$$

with transfer function

$$G^A(\lambda \Delta) = (1 - \alpha \lambda \Delta)^{-1} (1 + (1 - \alpha) \lambda \Delta) \quad (46)$$

If we denote by $\bar{Y}_{n+1}$, the corresponding discrete-time approximation that starts at the initial value $\bar{Y}_0$ instead of Y_0 , then we obtain for the difference

$$\begin{aligned} Y_{n+1} - \bar{Y}_{n+1} &= G^A(\lambda \Delta) (Y_n - \bar{Y}_n) \\ &= (G^A(\lambda \Delta))^n (Y_0 - \bar{Y}_0) \end{aligned} \quad (47)$$

Obviously, as long as the absolute value of the transfer function $G^A(\lambda \Delta)$ is smaller than 1, that is

$$|G^A(\lambda \Delta)| < 1 \quad (48)$$

the impact of the initial error $|Y_0 - \bar{Y}_0|$ is decreased during the simulation of the approximate trajectory over time, as can be seen from equation (47). However, for values

$$|G^A(\lambda \Delta)| > 1 \quad (49)$$

this is not the case, and errors are propagated. For the above test equation, it turns out that the region of A -stability ranges from $-2/(1 - 2\alpha)$ up to 0. This means for a degree of implicitness $\alpha \geq 1/2$ the corresponding drift implicit scheme is A -stable.

We call the following counterpart of the Milstein scheme (28) a *drift implicit Milstein scheme*. It has the form

$$\begin{aligned} Y_{n+1} &= Y_n + a(\tau_{n+1}, Y_{n+1}) \Delta + b \Delta W_n \\ &\quad + b b' 1/2 \{(\Delta W_n)^2 - \Delta\} \end{aligned} \quad (50)$$

There exist further families of drift implicit strong schemes of higher order. Some of these are A -stable and can be found in [7].

Drift Implicit Order 1.0 Strong Runge–Kutta Scheme

Let us now discuss a family of drift implicit schemes which avoid the use of derivatives. We shall call these drift implicit strong Runge–Kutta schemes. We emphasize that these schemes are not a simple heuristic adaptation of any deterministic Runge–Kutta scheme.

A family of drift implicit order 1.0 strong Runge–Kutta schemes is given as

$$\begin{aligned} Y_{n+1} &= Y_n + \{\alpha a(\tau_{n+1}, Y_{n+1}) + (1 - \alpha) a\} \Delta \\ &\quad + b \Delta W_n + \frac{1}{\sqrt{\Delta}} \{b(\tau_n, \bar{Y}_n) - b\} \\ &\quad \times \frac{1}{2} ((\Delta W_n)^2 - \Delta) \end{aligned} \quad (51)$$

with supporting value

$$\bar{Y}_n = Y_n + a \Delta + b \sqrt{\Delta} \quad (52)$$

and degree of implicitness parameter $\alpha \in [0, 1]$. Further implicit schemes and the derivations of their strong order of convergence can be also found in [7].

Balanced Implicit Method

A major difficulty arises from the fact that all the above strong schemes do not provide implicit expressions in the diffusion terms. Only drift terms are made implicit. Unfortunately, one cannot make diffusion terms in the Euler scheme simply implicit as it was possible for the drift. An *ad hoc* implicit discrete-time approximation with implicit diffusion term would result in a scheme that includes the inverse of a Gaussian random variable. Such an algorithm may explode and, thus, does not provide a reasonable scheme.

Drift implicit methods are well adapted for systems with small noise or additive noise. However, when the diffusion part plays an essential role in the dynamics, as it is the case for martingales,

the application of fully implicit methods, involving also implicit diffusion terms, could bring numerical stability.

An illustration for such a situation is provided by the one-dimensional SDE of an exponential martingale

$$dX_t = \sigma X_t dW_t \quad (53)$$

for $t \in [0, T]$, starting at $X_0 = x_0$. Here $W = \{W_t, t \in [0, T]\}$ is a standard Wiener process. The volatility σ is, for simplicity, chosen to be constant. The SDE (53) describes the standard Black–Scholes dynamics of a discounted asset price under the risk-neutral probability measure. Alternatively, it could describe the dynamics of a security denominated in units of the growth optimal portfolio under the real-world probability measure applying the benchmark approach [12]. These are typical dynamics that one faces when simulating asset dynamics in finance. Obviously, in such cases one cannot apply drift implicit schemes to improve the numerical stability of the approximate solution of an SDE that does not have any drift term, as is the case for the martingale satisfying the SDE (53). However, we can use implicit methods that introduce implicitness in the diffusion terms in such a way that the scheme still makes sense and converges to the correct limit.

Let us now describe such an implicit method that allows to overcome a range of numerical instabilities. In [11] a family of *balanced implicit methods* has been proposed that resolves, in many cases, the problem of numerical stability. Such a balanced implicit method can be written in the form

$$Y_{n+1} = Y_n + a \Delta + b \Delta W_n + (Y_n - Y_{n+1}) C_n, \quad (54)$$

where

$$C_n = c^0(Y_n) \Delta + c^1(Y_n) |\Delta W_n| \quad (55)$$

and c^0, c^1 represent positive, real-valued uniformly bounded functions. The freedom of choosing c^0 and c^1 can be exploited to construct a numerically stable scheme tailored for the dynamics of the given SDE. Note, however, the balanced implicit method is only of strong order $\gamma = 0.5$ since it is a variation of the Euler scheme with some additional higher order terms designed to control the propagation of errors. The low order of strong convergence is a

price that is paid here for obtaining better numerical stability.

The balanced implicit method can be interpreted as a family of strong methods, providing a kind of balance between the approximating and the higher order diffusion terms in a scenario simulation. In a range of applications in finance and filtering, see [3], balanced implicit methods have shown better numerical stability than most other methods presented above. We emphasize any numerically stable scheme is better than an unstable one. An eventually theoretically higher order of strong convergence of a scheme is only of secondary importance.

References

- [1] Burrage, K. & Burrage, P.M. (1998). General order conditions for stochastic Runge–Kutta methods for both commuting and non-commuting stochastic ordinary differential equation systems, *Applied Numerical Mathematics* **28**, 161–177.
- [2] Burrage, K. & Platen, E. (1994). Runge–Kutta methods for stochastic differential equations, *Annals of Numerical Mathematics* **1**(1–4), 63–78.
- [3] Fischer, P. & Platen, E. (1999). Applications of the balanced method to stochastic differential equations in filtering, *Monte Carlo Methods and Applications* **5**(1), 19–38.
- [4] Gaines, J.G. & Lyons, T.J. (1994). Random generation of stochastic area integrals, *SIAM Journal on Applied Mathematics* **54**(4), 1132–1146.
- [5] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*, Applied Mathematics, Springer, Vol. 53.
- [6] Jäckel, P. (2002). *Monte Carlo Methods in Finance*, Wiley.
- [7] Kloeden, P.E. & Platen E. (1999). *Numerical Solution of Stochastic Differential Equations*, Applied Mathematics, Springer, Vol. 23, Third printing.
- [8] Kloeden, P.E., Platen, E. & Schurz, H. (2003). *Numerical Solution of SDE's Through Computer Experiments*, Universitext, Springer, Third corrected printing.
- [9] Maruyama, G. (1955). Continuous Markov processes and stochastic equations, *Rendiconti del Circolo Matematico di Palermo* **4**, 48–90.
- [10] Milstein, G.N. (1974). Approximate integration of stochastic differential equations. *Theory of Probability and Its Applications* **19**, 557–562.
- [11] Milstein, G.N., Platen, E. & Schurz, H. (1998). Balanced implicit methods for stiff stochastic systems, *SIAM Journal on Numerical Analysis* **35**(3), 1010–1019.
- [12] Platen, E. & Heath, D. (2006). *A Benchmark Approach to Quantitative Finance*, Springer Finance.

Related Articles

LIBOR Market Model; LIBOR Market Models: Simulation; Monte Carlo Simulation for Stochastic Differential Equations; Pseudorandom Number Generators; Simulation of Square-root Processes; Stochastic Integrals; Stress Testing; Variance Reduction.

ECKHARD PLATEN & NICOLA BRUTI-LIBERATI

Pseudorandom Number Generators

Stochastic models of quantitative finance are defined in the abstract framework of probability theory. To apply the Monte Carlo method to these models, it suffices, in principle, to sample independent realizations of the underlying random variables or random vectors. This can be achieved by sampling independent random variables uniformly distributed over the interval $(0, 1)$ (i.i.d. $\mathcal{U}(0, 1)$, for short) and applying appropriate transformations to these uniform random variables. Nonuniform variate generation techniques develop such transformations and provide efficient algorithms that implement them [3, 6]. A simple general way to obtain independent random variables $X_1, X_2, \dots$ with distribution function F from a sequence of i.i.d. $\mathcal{U}(0, 1)$ random variables $U_1, U_2, \dots$ is to define

$$X_j = F^{-1}(U_j) \stackrel{\text{def}}{=} \min\{x \mid F(x) \geq U_j\}; \quad (1)$$

this is the *inversion method*. This technique can provide a sequence of independent standard normal random variables, for example, which can, in turn, be used to generate the sample path of a geometric Brownian motion or other similar type of stochastic process. There is no closed-form expression for the inverse standard normal distribution function, but very accurate numerical approximations are available.

But how do we get the i.i.d. $\mathcal{U}(0, 1)$ random variables? Realizing these random variables exactly is very difficult, perhaps, practically impossible. With current knowledge, this can be realized only *approximately*. Fortunately, the approximation seems good enough for all practical applications of the Monte Carlo method in financial engineering and in other areas as well.

A first class of methods to realize approximations of these random variables are based on real physical noise coming from hardware devices. There is a large variety of such devices; they include gamma ray counters, fast oscillators sampled at low and slightly random frequencies, amplifiers of heat noise produced in electric resistances, photon counting and photon trajectory detectors, and so on. Some of these devices sample a signal at successive epochs and return 0 if the signal is below a given threshold,

and 1 if it is above the threshold, at each sampling epoch. Others return the parity of a counter. Most of them produce sequences of bits that are slightly correlated and often slightly biased, but the bias and correlation can be reduced to a negligible amount, that becomes practically undetectable by statistical tests in reasonable time, by combining the bits in a clever way. For example, a simple technique to eliminate the bias when there is no correlation, proposed long ago by John von Neumann, places the successive bits in nonoverlapping pairs, discards all the pairs 00 and 11, and replaces the pairs 01 and 10 by 1 and 0, respectively. Generalizations of this technique can eliminate both the bias and correlation [2]. Simpler techniques such as Xoring (adding modulo 2) the bits by blocks of 2 or more, or Xoring several bit streams from different sources, are often used in practice. Reliable devices to generate random bits and numbers, based on these techniques, are available on the market. These types of devices are needed for applications such as cryptography, lotteries, and gambling machines, for example, where some amount of real randomness (or entropy) is essential to provide the required unpredictability and security.

For Monte Carlo methods, however, these devices are unnecessary and unpractical. They are unnecessary because simple *deterministic algorithms* are available that require no other hardware than a standard computer and provide good enough imitations of i.i.d. $\mathcal{U}(0, 1)$ random variables from a statistical viewpoint, in the sense that the statistical behavior of the simulation output is pretty much the same (for all practical purposes) if we use (*pseudo*)*random numbers* produced by these algorithms in place of true i.i.d. $\mathcal{U}(0, 1)$ random variables. These deterministic algorithmic methods are much more convenient than hardware devices.

A (*pseudo*)*random number generator* (RNG, for short) can be defined as a structure comprised of the following ingredients [9]: a finite *set of states* $\mathcal{S}$; a probability distribution on $\mathcal{S}$ to select the *initial state* s_0 (also called the *seed*); a *transition function* $f : \mathcal{S} \rightarrow \mathcal{S}$; an *output space* $\mathcal{U}$; and an *output function* $g : \mathcal{S} \rightarrow \mathcal{U}$. Here, we assume that $\mathcal{U}$ is the interval $(0, 1)$. The state evolves according to the recurrence $s_i = f(s_{i-1})$, for $i \geq 1$, and the *output* at step i is $u_i = g(s_i) \in \mathcal{U}$. These u_i 's are the successive *random numbers* produced by the RNG. (Following common usage in the simulation community, here we leave out

the qualifier “pseudo”. In the area of cryptology, the term *pseudo-RNG* refers to a stronger notion, with polynomial-time unpredictability properties [20]).

Because $\mathcal{S}$ is finite, the RNG must eventually return to a previously visited state, that is, $s_{l+j} = s_l$ for some $l \geq 0$ and $j > 0$. Then, $s_{i+j} = s_i$ and $u_{i+j} = u_i$ for all $i \geq l$; that is, the output sequence eventually repeats itself. The smallest $j > 0$ for which this happens is the *period length* ρ . Clearly, ρ cannot exceed $|\mathcal{S}|$, the number of distinct states. If the state can be represented with b bits of memory, then $\rho \leq 2^b$. For good RNGs, ρ is usually close to 2^b , as it is not difficult to construct recurrences with this property. Typical values of b range from 31 to around 20 000 or even higher [18]. In our opinion, ρ should never be less than 2^{100} and preferably more than 2^{200} . Values of b that exceed 1000 are unnecessary if the RNG satisfies the quality criteria described in what follows.

A key advantage of algorithmic RNGs is their ability to repeat exactly the same sequence of random numbers without storing them. Repeating the same sequence several times is essential for the proper implementation of variance reduction techniques such as using common random numbers for comparing similar systems, sensitivity analysis, sample-path optimization, external control variates, antithetic variates, and so on [1, 5] (see also **Variance Reduction**). It is also handy for program verification and debugging. On the other hand, some real randomness can be used for selecting the seed s_0 of the RNG.

Streams and Substreams

Modern high-quality simulation software often offers the possibility to declare and create virtual RNGs just like for any other type of variable or object, in practically unlimited amount. In an implementation adopted by several simulation software vendors, these virtual RNGs are called *streams*, and each stream is split into multiple *substreams* long enough to prevent potential overlap [14, 19]. For any given stream, there are methods to generate the next number, to rewind to the beginning of the stream or to the beginning of the current substream, or to the beginning of the next substream.

To illustrate its usefulness, consider a simple model of a financial option whose payoff is a function of a geometric Brownian motion observed at fixed

points in time. We want to estimate $d = \mathbb{E}[X_2 - X_1]$ where X_1 and X_2 are the payoffs with two slightly different parameter settings, such as different volatilities or different strike prices, for example. This is often useful for sensitivity analysis (estimating the *greeks*; see **Monte Carlo Greeks**). To estimate d , we would simulate the model with the two different settings using *common random numbers* across the two versions [1, 5] (see also **Variance Reduction**), repeat this n times independently and compute a confidence interval on d from the n independent copies of $X_2 - X_1$. To implement this, we take a stream of random numbers that contains multiple substreams, use the same substream to simulate both X_1 and X_2 for each replication, and n different substreams for the n replications. At the beginning of a replication, the stream is placed to the beginning of a new substream and the model is simulated to compute X_1 . Then the stream is reset to the beginning of its current substream before simulating the model again to compute X_2 . This ensures that exactly the same random numbers are used to generate the Brownian motion increments at the same time points for both X_1 and X_2 . Then the stream is moved to the beginning of the next substream for the next pair of runs.

There are many situations where the number of calls to the RNG during a simulation depends on the model parameters and may not be the same for X_1 and X_2 . Even in that case, the above scheme ensures that the RNG restarts at the same place for both parameter settings, for each replication. In more complicated models, to ensure a good synchronization of the random numbers across the two settings (i.e., make sure that the same random numbers are used for the same purposes in both cases), it is typically convenient to have several different streams, each stream being dedicated to one specific aspect of the model. For instance, in the previous example, if we also need to simulate external events that occur according to a Poisson process and influence the payoff in some way (e.g., they could trigger jumps in the Brownian motion), it is better to use a separate stream to simulate this process, to guarantee that no random number is used for the Brownian motion increment in one setting and for the Poisson process in the other setting.

Quality Criteria and Testing

A good RNG must obviously have a very long period, to make sure that there is no chance of wrapping around. It should also be *repeatable* (able to reproduce exactly the same sequence several times), *portable* (be easy to implement and behave the same way in different software/hardware environments), and it should be easy to split its sequence into several disjoint streams and substreams, and implement efficient tools to move between those streams and substreams. The latter requires the availability of efficient jump-ahead methods, that can quickly compute s_{i+v} given s_i , for any large v . The number b of bits required to store the state should not be too large, because the computing time for jumping ahead typically increases faster than linearly with b , and also because there can be a large number of streams and substreams in a given simulation, especially for large complex models. Another key performance measure is the speed of the generator itself. Fast generators can produce up to 100 million $\mathcal{U}(0, 1)$ random numbers per second on current personal computers [18].

All these nice properties are not sufficient, however. For example, an RNG that returns $u_i = (i/10^{1000}) \bmod 1$ at step i satisfies these properties but is definitely not recommendable, because its successive output values have an obvious strong correlation. Ideally, if we select a random seed s_0 uniformly in $\mathcal{S}$, we would like the vector of the first s output values, $(u_0, \dots, u_{s-1})$, to be uniformly distributed over the s -dimensional unit hypercube $[0, 1]^s$ for each $s > 0$. This would guarantee both uniformity and independence. Formally, we cannot have this, because these s -dimensional vectors must take their values from the finite set $\Psi_s = \{(u_0, \dots, u_{s-1}) : s_0 \in \mathcal{S}\}$, whose cardinality cannot exceed $|\mathcal{S}|$. If s_0 is random, Ψ_s can be viewed as the *sample space* from which vectors of successive output values are drawn randomly. Then, to approximate the uniformity and independence, we want the finite set Ψ_s to provide a dense and uniform coverage of the hypercube $[0, 1]^s$, at least for small and moderate values of s . This is possible only if $\mathcal{S}$ has large cardinality, and it is, in fact, a more important reason for having a long period than the danger of exhausting the cycle.

Hence, the uniformity of Ψ_s in $[0, 1]^s$ is a key quality criterion. But how do we measure it? There are many ways of measuring the uniformity (or the discrepancy from the uniform distribution) for

a point set in the unit hypercube [16, 22] (*see also Quasi-Monte Carlo Methods*). To be practical, the uniformity measure must be selected so that it can be effectively computed without generating explicitly the points of Ψ_s . For this reason, the theoretical *figures of merit* that measure the uniformity usually depend on the mathematical structure of the RNG. This is also the main reason for having RNGs based on linear recurrences: their point sets Ψ_s are easier to analyze mathematically, because they have a simpler structure. One could argue that nonlinear and more complex structures give rise to point sets Ψ_s that look more random, and some of them behave very well in empirical statistical tests, but their structure is much harder to analyze. They could leave large holes in $[0, 1]^s$ that are difficult to detect.

To design a good RNG, one typically selects an algorithm together with the size of the state space, and constraints on the parameters that ensure a fast implementation. Then one makes a computerized search in the space of parameters to find a set of values that give (i) the maximal period length within this class of generators and then (ii) the largest figure of merit than can be found. RNGs are thus selected and constructed primarily based on theoretical criteria. Then, they are implemented and tested empirically.

A large variety of *empirical statistical tests* have been designed and implemented for RNGs [8, 18]. All these tests try to detect empirical evidence against the hypothesis $\mathcal{H}_0$ that the u_i are i.i.d. $\mathcal{U}[0, 1]$. A test can be any function Y of a finite set of u_i 's, which can be computed in reasonable time, and whose distribution under $\mathcal{H}_0$ can be approximated well enough. There is an unlimited number of such tests. When applying the test, one computes the realization of Y , say y , and then the probability $p^+ = \mathbb{P}[Y \geq y \mid \mathcal{H}_0]$, called the right p value. If Y takes a much larger value than expected, then p^+ will be very close to 0, and we declare that the RNG fails the test. We may also examine the left p value $p^- = \mathbb{P}[Y \leq y \mid \mathcal{H}_0]$, or both p^+ and p^- , depending on the design of the test. When a generator really fails a test, it is not unusual to find p values as small as 10^{-15} or less.

Specific batteries that contain a variety of standard tests, which detect problems often encountered in poorly designed or too simple RNGs, have been proposed and implemented [18]. The bad news is that a majority of the RNGs available in popular commercial software fail these tests unequivocally,

with p values smaller than 10^{-15} . These generators should be discarded, unless we have very good reasons to believe that for our specific simulation models, the problems detected by these failed tests will not affect the results. The good news is that some freely available high-quality generators pass all the tests in these batteries. Of course, passing all these tests is not a proof that the RNG is reliable for all the possible simulations, but it certainly improves our confidence in the generator. In fact, no RNG can pass all conceivable statistical tests. In some sense, the good RNGs fail only very complicated tests that are hard to find and implement, whereas bad RNGs fail simple tests.

Linear Recurrences

Most RNGs used for simulation are based on linear recurrences of the general form

$$x_i = (a_1 x_{i-1} + \dots + a_k x_{i-k}) \bmod m \quad (2)$$

where k and m are positive integers, and the coefficients $a_1, \dots, a_k$ are in $\{0, 1, \dots, m-1\}$, with $a_k \neq 0$. Some use a large value of m , preferably a prime number, and define the output as $u_i = x_i/m$, so the state at step i can be viewed as $s_i = \mathbf{x}_i = (x_{i-k+1}, \dots, x_i)$. The RNG is then called a *multiple recursive generator* (MRG). For $k = 1$, we obtain the classical linear congruential generator (LCG). In practice, the output transformation is modified slightly to make sure that u_i is always strictly between 0 and 1, for example, by taking $u_i = (x_i + 1)/(m + 1)$ or $u_i = (x_i + 1/2)/m$. Jumping ahead from $\mathbf{x}_i$ to $\mathbf{x}_{i+v}$ for an arbitrary large v can be implemented easily: because of the linearity, one can write $\mathbf{x}_{i+v} = \mathbf{A}_v \mathbf{x}_i \bmod m$, where $\mathbf{A}_v$ is a $k \times k$ matrix that can be precomputed once for all [13]. When m is prime, one can choose the coefficients a_j so that the period length reaches $m^k - 1$, its maximum [8].

The point set Ψ_s produced by an MRG is known to have a *lattice structure*, and its uniformity is measured *via* a figure of merit for the quality of that lattice, for several values of s . This is known as the *spectral test* [4, 8, 10].

Typically, m is chosen as one of the largest prime integers representable on the target computer, for example, $m = 2^{31} - 1$ on a 32-bit computer. Then,

a direct implementation of equation (2) with integer numbers would cause overflow, so more clever implementation techniques are needed. These techniques require that we impose additional conditions on the coefficients a_j . We have to be careful that these conditions do not oversimplify the structure of the point set Ψ_s . One extreme example of this is to take only two nonzero coefficients, say a_r and a_k , both equal to ± 1 . Implementation is then easy and fast. However, all triples of the form (u_i, u_{i-r}, u_{i-k}) produced by such a generator, for $i = 0, 1, \dots$, lie in only two planes in the three-dimensional unit cube. Despite this awful behavior, these types of generators (or variants thereof) can be found in many popular software products [18]. They should be avoided. All simple LCGs, say with $m \leq 2^{64}$, should be discarded; they have too much structure and their period length is too short for present computers.

One effective way of implementing high-quality MRGs is to combine two (or more) of them by adding their outputs modulo 1. (There are also other slightly different ways of combining.) If the components have distinct prime moduli, the combination turns out to be just another MRG with (nonprime) modulus m equal to the product of the moduli of the components, and the period can be up to half the product of the component's periods when we combine two of them. The idea is to select the components so that (i) a fast implementation is easy to construct for each individual component and (ii) the combined MRG has a more complicated structure and highly uniform sets Ψ_s , as measured by the spectral test [10]. Specific MRG constructions can be found in [10, 13, 18] and the references given therein.

A different approach uses a linear recurrence as in equation (2), but with $m = 2$. All operations are then performed modulo 2, that is, in the finite field $\mathbb{F}_2$ with elements $\{0, 1\}$. This allows very fast implementations by exploiting the binary nature of computers. A general framework for this is the matrix linear recurrence [13, 17]:

$$\mathbf{x}_i = \mathbf{A} \mathbf{x}_{i-1} \quad (3)$$

$$\mathbf{y}_i = \mathbf{B} \mathbf{x}_i \quad (4)$$

$$u_i = \sum_{\ell=1}^w y_{i,\ell-1} 2^{-\ell} \quad (5)$$

where $\mathbf{x}_i = (x_{i,0}, \dots, x_{i,k-1})^t$ is the k -bit *state vector* at step i , $\mathbf{y}_i = (y_{i,0}, \dots, y_{i,w-1})^t$ is the w -bit *output*

vector at step i , k , and w are the positive integers, $\mathbf{A}$ is a $k \times k$ binary *transition matrix*, $\mathbf{B}$ is a $w \times k$ binary *output transformation matrix*, and $u_i \in [0, 1)$ is the *output* at step i . All operations in equations (3) and (4) are performed in $\mathbb{F}_2$. These RNGs are called $\mathbb{F}_2$ -linear generators.

The theoretical analysis usually assumes the simple output definition (5), but, in practice, this definition is modified slightly to avoid returning 0 or 1. This framework covers several types of generators, including the Tausworthe, polynomial LCG, generalized feedback shift register (GFSR), twisted GFSR, Mersenne twister, Well, Xorshift, linear cellular automaton, and combinations of these [13, 17, 21]. With a carefully selected matrix $\mathbf{A}$ (its characteristic polynomial must be a primitive polynomial over $\mathbb{F}_2$), the period length can reach $2^k - 1$. In practice, the matrices $\mathbf{A}$ and $\mathbf{B}$ are chosen so that the products (3) and (4) can be implemented very efficiently on a computer by a few simple binary operations such as or, exclusive-or, shift, and rotation, on blocks of bits. The idea is to find a compromise between the number of such operations (which affects the speed) and a good uniformity of the point sets Ψ_s (which is easier to reach with more operations). The uniformity of these point sets is measured *via* their *equidistribution*; essentially, the hypercube $[0, 1]^s$ is partitioned into small subcubes (or subrectangles) of equal sizes, and for several such partitions, we check if all the subcubes contain exactly the same number of points from Ψ_s . This can be computed efficiently by computing the ranks of certain binary matrices [17]. Combined generators of this type, defined by Xoring the output vectors $\mathbf{y}_i$ of the components, are equivalent to yet another $\mathbb{F}_2$ -linear generator. Such combinations have the same motivation as for MRGs.

Nonlinear Generators

Linear RNGs have many nice properties, but they also fail certain specialized statistical tests focused at detecting linearity. When the simulation itself applies nonlinear transformations to the uniform random numbers, which is typical, one should not worry about the linearity, unless the structure of Ψ_s is not very good. However, there are cases where the linearity can matter. For example, to generate a large random binary matrix, one should not use an $\mathbb{F}_2$ -linear generator, because the rank of the matrix is

likely to be much smaller than expected, due to the excessive linear dependence [18].

There are many ways of constructing nonlinear generators. For example, one can simply add a nonlinear output transformation to a linear RNG, or permute (shuffle) the output values with the help of another generator. Another way is to combine an MRG with an $\mathbb{F}_2$ -linear generator, either by addition modulo 1 or by Xoring the outputs. An important advantage of this technique is that the uniformity of the resulting combined generator can be assessed theoretically, at least to a certain extent [15]. They can also be fast.

When combining generators, it is important to understand what we do and we should be careful to examine the structure not only of the combination but also of the quality of the components. By blindly combining two good components, it is indeed possible (and not too difficult) to obtain a bad (worst) RNG.

Generators whose underlying recurrence is nonlinear are generally harder to analyze and are slower. These are the types of generators used for cryptographic applications. Empirically, well-designed nonlinear generators tend to perform better in statistical tests than the linear ones [18], but from the theoretical perspective, their structure is not understood as well. RNGs based on chaotic dynamical systems have often been proposed in the literature, but these generators have several major drawbacks, including the fact that their s -dimensional uniformity is often very poor [7].

What to Look For and What to Avoid

A quick look at the empirical results in [12, 18] shows that many widely used RNGs are seriously deficient, including the default generators of several highly popular software products. So before running important simulation experiments, one should always check what is the default RNG, and be ready to replace it if needed. Note that the generators that pass the tests in [18] are not all recommended. Before adoption, one should verify that the RNG has solid theoretical support, that it is fast enough, and that multiple streams and substreams are available, for example. Convenient software packages with multiple streams and substreams are described in [14, 19] and are available freely from the web page of this

author. These packages are based on combined MRGs of [10], combined Tausworthe generators of [11], the Well generators [23] (which are improvements over the Mersenne twister in terms of equidistribution), and some additional nonlinear generators, among others. No uniform RNG can be guaranteed against all possible defects, but one should at least avoid those that fail simple statistical tests miserably and go for the more robust ones, for which no serious problem has been detected after years of usage and testing.

Acknowledgments

This work has been supported by the Natural Sciences and Engineering Research Council of Canada Grant No. ODGP0110050 and a Canada Research Chair to the author.

References

- [1] Bratley, P., Fox, B.L. & Schrage, L.E. (1987). *A Guide to Simulation*, 2nd Edition, Springer-Verlag, New York.
- [2] Chor, B. & Goldreich, O. (1988). Unbiased bits from sources of weak randomness and probabilistic communication complexity, *SIAM Journal on Computation* **17**(2), 230–261.
- [3] Devroye, L. (1986). *Non-Uniform Random Variate Generation*, Springer-Verlag, New York.
- [4] Fishman, G.S. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*, Series in Operations Research, Springer-Verlag, New York.
- [5] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*, Springer-Verlag, New York.
- [6] Hörmann, W., Leydold, J. & Derflinger, G. (2004). *Automatic Nonuniform Random Variate Generation*, Springer-Verlag, Berlin.
- [7] Jäckel, P. (2002). *Monte Carlo Methods in Finance*, John Wiley & Sons, Chichester.
- [8] Knuth, D.E. (1998). *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, 3rd Edition, Addison-Wesley, Reading.
- [9] L'Ecuyer, P. (1994). Uniform random number generation, *Annals of Operations Research* **53**, 77–120.
- [10] L'Ecuyer, P. (1999). Good parameters and implementations for combined multiple recursive random number generators, *Operations Research* **47**(1), 159–164.
- [11] L'Ecuyer, P. (1999). Tables of maximally equidistributed combined LFSR generators, *Mathematics of Computation* **68**(225), 261–269.
- [12] L'Ecuyer, P. (2001). Software for uniform random number generation: distinguishing the good and the bad, in *Proceedings of the 2001 Winter Simulation Conference*, IEEE Press, Piscataway, pp. 95–105.
- [13] L'Ecuyer, P. (2006). Uniform random number generation, in *Simulation, Handbooks in Operations Research and Management Science*, S.G. Henderson & B.L. Nelson, eds, Elsevier, Amsterdam, Chapter 3, pp. 55–81.
- [14] L'Ecuyer, P. & Buist, E. (2005). Simulation in Java with SSJ, in *Proceedings of the 2005 Winter Simulation Conference*, IEEE Press, pp. 611–620.
- [15] L'Ecuyer, P. & Granger-Piché, J. (2003). Combined generators with components from different families, *Mathematics and Computers in Simulation* **62**, 395–404.
- [16] L'Ecuyer, P. & Lemieux, C. (2002). Recent advances in randomized quasi-Monte Carlo methods, in *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, M. Dror & P. L'Ecuyer, F. Szidarovszky, eds, Kluwer Academic, Boston, pp. 419–474.
- [17] L'Ecuyer, P. & Panneton, F. (2009). F_2 -Linear random number generators, *Advancing the Frontiers of Simulation: A Festschrift in Honor of George S. Fishman*, Springer-Verlag.
- [18] L'Ecuyer, P. & Simard, R. (2007). TestU01: A C library for empirical testing of random number generators, *ACM Transactions on Mathematical Software* **33**(4), Article 22, 5.
- [19] L'Ecuyer, P., Simard, R., Chen, E.J. & Kelton, W.D. (2002). An object-oriented random-number package with many long streams and substreams, *Operations Research* **50**(6), 1073–1075.
- [20] Luby, M. (1996). *Pseudorandomness and Cryptographic Applications*, Princeton University Press, Princeton.
- [21] Matsumoto, M. & Nishimura, T. (1998). Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, *ACM Transactions on Modeling and Computer Simulation* **8**(1), 3–30.
- [22] Niederreiter, H. (1992). Random number generation and quasi-Monte Carlo methods, in *SIAM CBMS-NSF Regional Conference Series in Applied Mathematics*, SIAM, Philadelphia, Vol. 63.
- [23] Panneton, F., L'Ecuyer, P. & Matsumoto, M. (2006). Improved long-period generators based on linear recurrences modulo 2, *ACM Transactions on Mathematical Software* **32**(1), 1–16.

Related Articles

Monte Carlo Greeks; Monte Carlo Simulation for Stochastic Differential Equations; Quasi-Monte Carlo Methods; Stochastic Differential Equations; Scenario Simulation; Variance Reduction.

PIERRE L'ECUYER

Monte Carlo Greeks

Monte Carlo simulation (*see Monte Carlo Simulation*) allows not only to compute option prices but can also be used to estimate their sensitivity parameters (delta, gamma, vega, etc.). The baseline estimator, the one which is to be improved, is known as the *resimulation estimator*. Resimulation consists of simply rerunning the simulation (using the same sequence of pseudorandom numbers) after perturbing the parameter or variable of interest. The corresponding estimator is then taken as the change in the value of the derivative divided by the magnitude of the perturbation. The purpose of most of the sophisticated methods that are discussed is primarily that resimulation can be very demanding computationally, while the desired information is in some way already contained in the original simulation data and can be extracted with much more modest amounts of computation than with resimulation. The focus of this article is on the two main such techniques, known as the *pathwise* and *likelihood ratio methods*.

The standard integration, or expectation, that represents the value of a derivative takes the form

$$V = \int \pi(S) \psi(S) dS \quad (1)$$

where $\pi(\bullet)$ represents the payoff function and $\psi(\bullet)$ is the probability density of the underlying. The pathwise estimator is based on how the relevant parameter functionally enters into the payoff, whereas the likelihood ratio estimator is based on how the relevant parameter enters into the density function of the underlying(s), given the appropriate change of variable. We specialize this in what follows by assuming that normal variates are the drivers for all price changes and by including a vector of parameters α , which may represent the initial prices of the underlying securities or rates, volatilities, time to expiration, dividend rates, and so on. In what follows, we give the specifics of the computation of these methods and give some examples of each.

Pathwise Method

The pathwise estimator is obtained with

$$V = \int \pi[S(z, \alpha)] \varphi(z) dz \quad (2)$$

where $\varphi(z)$ is the standard normal density. The desired sensitivities are then

$$\frac{dV}{d\alpha} = \frac{d}{d\alpha} \left\{ \int \pi[S(z, \alpha)] \varphi(z) dz \right\} \quad (3)$$

or, exchanging the differentiation and integral

$$\frac{dV}{d\alpha} = \int \frac{d}{d\alpha} (\pi[S(z, \alpha)]) \varphi(z) dz \quad (4)$$

The corresponding (unbiased) pathwise estimator can then be computed as

$$\left\{ \frac{dV}{d\alpha} \right\}_P = \frac{1}{N} \sum_{i=1}^N \frac{d}{d\alpha} (\pi[S^i(z_i, \alpha)]) \quad (5)$$

where the summation is taken over N simulation paths indexed by i and the z_i are independent and identically distributed standard normal variates. As an application of the pathwise methodology, we consider computing an estimate of vega for a European call option under the Black–Scholes model where

$$dS_t = S_t [(r - \delta) dt + \sigma dB_t] \quad (6)$$

and S_t is the underlying asset price at time t , r is the domestic interest rate, δ is the dividend rate or foreign interest rate, σ is the volatility, and B_t is standard Brownian motion. A European call option with expiration T and strike price K has value equal to

$$V = E [e^{-rT} (S_T - K)^+] \quad (7)$$

where $E(\bullet)$ is the risk-neutral expectation operator. In the Black–Scholes model we have

$$S_T = S_0 \exp \left[\left(r - \delta - \frac{\sigma^2}{2} \right) T + \sigma \sqrt{T} Z \right] \quad (8)$$

where Z is a realization of a standard normal variate. Therefore

$$\begin{aligned} \frac{dS_T}{d\sigma} &= S_T (-\sigma T + \sqrt{T} Z) \\ &= \frac{S_T}{\sigma} \left[\ln(S_T/S_0) - \left(r - \delta + \frac{\sigma^2}{2} \right) T \right] \end{aligned} \quad (9)$$

Now we simply apply the chain rule

$$\frac{dV}{d\sigma} = E \left[\frac{d\pi}{dS_T} \frac{dS_T}{d\sigma} \right] \quad (10)$$

where

$$\frac{d\pi}{dS_T} = e^{-rT} I(S_T > K) \quad (11)$$

and $I(\bullet)$ is the indicator function that takes on the value 1 if the argument is true and 0 otherwise. That is, the call payoff responds one-to-one with the underlying at expiration provided that the option finishes in-the-money. We then replace the expectation with a summation, as before, to obtain the pathwise estimator

$$\left\{ \frac{dV}{d\sigma} \right\}_P = \frac{1}{N} \sum_{i=1}^N e^{-rT} 1(S_T^i > K) \frac{S_T^i}{\sigma} \times \left[\ln(S_T^i/S_0) - \left(r - \delta + \frac{\sigma^2}{2} \right) T \right] \quad (12)$$

Likelihood Ratio Method

For the likelihood ratio method, we perform a change of variables such that the dependence on α in the integrand is put into the density function *via*

$$\frac{dV}{d\alpha} = \int \pi(S) \frac{d}{d\alpha} \varphi(S; \alpha) dS \quad (13)$$

or

$$\frac{dV}{d\alpha} = \int \pi(S) \frac{\frac{d}{d\alpha} \varphi(S; \alpha)}{\varphi(S; \alpha)} \varphi(S; \alpha) dS \quad (14)$$

In this form, the calculation resembles the estimation of the price of a derivative with payoff

$$\chi(S; \alpha) \equiv \pi(S) \omega(S; \alpha) \quad (15)$$

where

$$\omega \equiv \frac{\frac{d}{d\alpha} \varphi(S; \alpha)}{\varphi(S; \alpha)} \quad (16)$$

so that we have

$$\frac{dV}{d\alpha} = \int \chi(S; \alpha) \varphi(S, \alpha) dS \quad (17)$$

As with the pathwise method, we can convert this into a discrete sum suitable for a Monte Carlo

implementation to obtain an unbiased estimator of

$$\left\{ \frac{dV}{d\alpha} \right\}_L = \frac{1}{N} \sum_{i=1}^N \chi(S_i(z_i); \alpha) \quad (18)$$

In order to illustrate this technique, we again consider the vega calculation. For the normal density we get

$$\omega = \frac{\frac{d}{d\alpha} \varphi(S; \alpha)}{\varphi(S; \alpha)} \quad (19)$$

$$= \frac{\ln(x/S_0) - (r - \delta - \sigma^2/2) T}{S_0 \sigma^2 T} \quad (20)$$

and therefore the likelihood ratio estimator is

$$\left\{ \frac{dV}{d\sigma} \right\}_L = \sum_{i=1}^N e^{-rT} (S_T^i(z_i) - K)^+ \times \left[\frac{\ln(x/S_0) - (r - \delta - \sigma^2/2) T}{S_0 \sigma^2 T} \right] \quad (21)$$

Comments

The interchange of the differentiation and integral operators that both the pathwise and likelihood ratio methods entail requires smoothness in the payoff functions and density functions, respectively. As financial derivative payoffs are often not smooth while densities generally are, the likelihood ratio method is more advantageous. For example, the call option payoff is not differentiable in the underlying asset price. The problem with the likelihood ratio method, however, is that the parameter of interest may not appear in the density function even after transformations are considered, although this is a fairly rare occurrence. For programming systems, a very significant advantage of the likelihood ratio method is that it can be coded in a modular way in that prior knowledge of the payoff is not required as made clear by equations (15–17).

It should be noted that these methods can also be applied to path-dependent payoffs (which can complicate the application of the pathwise method) as well as non-Gaussian cases (which can complicate the application of the likelihood ratio method).

It may be mentioned that there are other methods, such as the equivalent entropy projection and Malliavin calculus approaches. The equivalent entropy projection method, instead of perturbing the parameters of interest, involves perturbing the probabilities of the simulated paths. The Malliavin calculus approach (see **Sensitivity Computations: Integration by Parts**) is a more elaborate or general version of, but arguably not superior, to the likelihood ratio method [1].

References

- [1] Chen, N. & Glasserman, P. (2007). Malliavin Greeks without Malliavin calculus, *Stochastic Processes and their Applications* **117**, 1723.

Further Reading

Avellaneda, M. & Gamba, R. (2001). Conquering the Greeks in Monte Carlo: Efficient Calculation of the Market Sensitivities and Hedge Ratios of Financial Assets by Direct Numerical Simulation in *Quantitative Analysis in Financial Markets*, M. Avellaneda, ed., World Scientific, 336–364, Vol. III,

[www.math.nyu.edu/faculty/avellane/Conquering TheGreeks.pdf](http://www.math.nyu.edu/faculty/avellane/Conquering%20TheGreeks.pdf).

Broadie, M. & Glasserman, P. (1996). Estimating security price derivatives using simulation, *Management Science* **42**(2), 269–285.

Dupire, B. (ed) (1998). *Monte Carlo: Methodologies and Applications for Pricing and Risk Management*, Risk Publications.

Fournie, E., Lasry, J.M., Lebuchoux, J., Lions, P.L. & Touzi, N. (1999). Applications of Malliavin Calculus to Monte Carlo methods in finance, *Finance and Stochastics* **3**(4), 391–412.

Jäckel, P. (2002). *Monte Carlo methods in finance*, John Wiley & Sons.

Jäckel, P. (2005). *More Likely than Not*, www.jaeckel.org.

Related Articles

Delta Hedging; Gamma Hedging; Monte Carlo Simulation; Monte Carlo Simulation for Stochastic Differential Equations; Sensitivity Computations: Integration by Parts; Stochastic Differential Equations: Scenario Simulation; Variance Reduction;

MICHAEL CURRAN

Bermudan Options

A Bermudan option allows its holder to exercise the contract before maturity, and this feature makes its pricing significantly more difficult in comparison with the corresponding European option. Even for simple put options, currently there are no explicit formulae for their prices and therefore numerical methods must be employed. Although there are several such methods available for pricing Bermudan and American (see **American Options**) options depending on a single asset (see **Finite Difference Methods for Early Exercise Options**), these methods become typically ineffective for options on multiple assets. In such cases, we have to usually resort to methods based on Monte Carlo simulations. In this context, approaches that combine simulations with regression techniques have proven to be particularly effective. An attractive feature of these methods is that, in principle, they can be applied to any situation where trajectories of the underlying process can be simulated, since no other information about the process is required (unlike, e.g., the stochastic mesh method) (see **Stochastic Mesh Method**). In particular, they can be applied to path-dependent options. Since their introduction by several authors, especially Carrière [2], Thitsiklis and Van Roy [7], and Longstaff and Schwartz [4], the area of their applications has been extended beyond the pricing of Bermudan options, and currently it also includes the optimal dynamic asset allocation problem [1] and hedging [6].

Regression methods use the dynamic programming representation to determine the price of the option and also the optimal exercise strategy. To simplify the presentation of these methods, suppose that our objective is to price a Bermudan style option whose payoff depends only on the current value of the underlying security. The option can be exercised at $M + 1$ time points (including the initial time), which we shall denote by $t_0, t_1, \dots, t_M = T$. At the time of exercise, τ , the value of the option is equal to $G(\tau, S_\tau)$, where G is a payoff function. The dynamic of the price of the underlying security is described by a process $\{S_t\}_{t=0,1,\dots,M}$, which we assume to be a Markov chain with values in $\mathcal{R}^b$. This process may be obtained, for example, as a result of sampling a continuous process $\{S_t\}_{0 \leq t \leq T}$ that solves a stochastic differential equation.

From the general theory of arbitrage-free pricing (see **Risk-neutral Pricing**), it follows that an arbitrage-free price of the option can be represented as the optimal expected discounted payoff:

$$P(t_0, S_0) := \max_{\tau} E[B(t_0, \tau)G(\tau, S_\tau)] \quad (1)$$

where the expectation is taken with respect to a given risk-neutral measure Q , and $B(s, t)$ denotes the discount factor for the period (s, t) . The maximum in equation (1) is taken over all stopping times taking values in the set $\{t_0, t_1, \dots, t_M\}$.

Since we do not know the optimal exercise strategy, a direct calculation of the price from equation (1) is not feasible. However, the price of the option can also be obtained by using the dynamic programming representation. For this, we use the following backward recursion to find functions $P(t_i, \cdot)$, $i = 1, \dots, M$,

$$P(T, x) = G(T, x) \quad (2)$$

$$P(t_i, x) = \max\{G(t_i, x), C(t_i, x)\}, \quad i = M - 1, \dots, 0 \quad (3)$$

where the continuation value, $C(t_i, x)$, is defined as

$$C(t_i, x) := B_i E[P(t_{i+1}, S_{t_{i+1}}) | S_{t_i} = x] \quad (4)$$

Then the price of the option at t_0 is given by $P(t_0, S_0)$. In the last equation, we assume that the discount factor $B_i \equiv B(t_i, t_{i+1})$ is deterministic. Equation (3) lends itself to a very intuitive explanation. At the i th exercise opportunity, the owner of the option makes the decision about early exercise by comparing the immediate exercise value with the present value of continuing. The larger of these two determines the present value of the option and the optimal action.

To use this method, in practice, we must be able to calculate efficiently the conditional expectations

$$E[P(t_{i+1}, S_{t_{i+1}}) | S_{t_i} = x] \quad (5)$$

for $i = 0, \dots, M - 1$ and some selected set of points $x \in \mathcal{R}^b$ from the state space. Regression-based methods accomplish this through the regression of option values at the next time step on a set of regressors that depend on the current state. The main assumption behind these methods is that the conditional expectation (5) as a function of the state variable x

can be represented in the form of an infinite series expansion, meaning that we have

$$C(t_i, x) = \sum_{j=1}^{\infty} a_{ij} \beta_j(x) \quad (6)$$

for some basis functions $\beta_j : \mathcal{R}^b \rightarrow \mathcal{R}$ and constants $\{a_{ij}\}$. Then an approximation to the continuation value can be obtained by using only a finite number of basis functions, say L . This can be accomplished, for example, by projecting $C(t_i, \cdot)$ onto the span of the basis functions β_j , $j = 1, \dots, L$. If the projection space is equipped with a measure that corresponds to the distribution of S_{t_i} , then the coefficients in this approximation, $a_{i1}^*, \dots, a_{iL}^*$, solve the following optimization problem

$$\begin{aligned} & E \left[\left(C(t_i, S_{t_i}) - \sum_{j=1}^L a_{ij}^* \beta_j(S_{t_i}) \right)^2 \right] \\ &= \min_{\{a_{i1}, \dots, a_{iL}\}} E \left[\left(C(t_i, S_{t_i}) - \sum_{j=1}^L a_{ij} \beta_j(S_{t_i}) \right)^2 \right] \end{aligned} \quad (7)$$

Thus, here the continuation value is approximated by a member of a parametric family of functions. This method, however, cannot be implemented directly since the continuation value $C(t_i, \cdot)$ is unknown. From the definition of $C(t_i, \cdot)$, it follows that for a single realization $(s_{i1}, s_{(i+1)1})$ of the vector $(S_{t_i}, S_{t_{i+1}})$ the continuation value $C(t_i, s_{i1})$ can be approximated by $B_{t_i} P(t_{i+1}, s_{(i+1)1})$. This observation leads to the selection of the coefficients $a_{i1}^*, \dots, a_{iL}^*$ that minimize the following criterion (see [8]):

$$E \left[\left(B_{t_i} P(t_{i+1}, S_{t_{i+1}}) - \sum_{j=1}^L a_{ij} \beta_j(S_{t_i}) \right)^2 \right] \quad (8)$$

where the expectation is taken with respect to the joint distribution of $(S_{t_i}, S_{t_{i+1}})$.

The argument that motivates the use of equation (8) as a method of approximation may suggest that the method will not be accurate, since we are approximating the continuation value at a given state by using only one successor. We should observe, however, that in this method we are not approximating continuation values individually at each state but

rather we approximate the continuation value $C(t_i, \cdot)$ treated as a function. Because of this and the assumed smoothness of this function, the resulting estimate of the continuation value at any state “borrows” also information about continuation values from points in a neighborhood of this state.

A factor that indeed determines the effectiveness of this approach is the selection of the basis functions. The assumption that guarantees the existence of the infinite series expansion (6) is rather a weak one, as any sufficiently smooth function can be approximated, for example, by polynomials. In practice, however, we have to truncate this expansion to a finite sum and hence we have to decide which terms we need to keep. The choice of a finite number of basis functions determines the success of the method and often must be crafted to the problem at hand. It becomes especially difficult for options on multiple assets, since then the number of required basis functions grows quickly with the dimension of the underlying price process.

On the basis of this method of approximation of continuation values, we can define an implementable procedure for pricing Bermudan options. For this, in equation (8) we have to substitute for $P(t_{i+1}, \cdot)$ its approximation, $\hat{P}(t_{i+1}, \cdot)$, determined from the backward induction (2)–(4). In addition, to find the coefficients $a_{i1}^*, \dots, a_{iM}^*$ that minimize equation (8), typically we have to approximate the expectation by using a sample mean. The resulting algorithm can be summarized in the following way. In the first phase, we simulate N independent trajectories $\{s_{1j}, \dots, s_{Mj}\}$, $j = 1, \dots, N$, of the Markov chain $\{S_{t_i}\}_{0 \leq i \leq M}$. At maturity of the contract, we set $\hat{P}_{Mj} = G(T, s_{Mj})$ and then start the backward induction. Given the estimated values $\hat{P}_{(i+1)j}$, $j = 1, \dots, N$, we approximate the continuation value at s_{ij} by $\hat{C}_{ij} = \sum_{k=1}^M a_{ik}^* \beta_k(s_{ij})$, where $a_{i1}^*, \dots, a_{iM}^*$ minimize

$$\sum_{j=1}^N \left[B_{t_i} \hat{P}_{(i+1)j} - \sum_{k=1}^M a_{ik} \beta_k(s_{ij}) \right]^2 \quad (9)$$

Then, we set $\hat{P}_{ij} = \max\{G(t_i, s_{ij}), \hat{C}_{ij}\}$. Finally, the price of the option is calculated as $\max\{B_0(\hat{P}_{11} + \dots + \hat{P}_{1N})/N, G(t_0, s_0)\}$.

Carrière [2] has proposed a similar approach, but instead of approximating the continuation value by a member of a parametric family he suggests using

nonparametric regression techniques based on splines and a local polynomial smoother. The approach proposed by Longstaff and Schwartz [4] is similar to the one presented here except that the authors use a different formula for calculating $\hat{P}_{ij}$. They also recommend using the least-squares method (9) only for options in the money.

Convergence properties of regression-based methods have been studied in [3, 5, 7, 8]. In particular, Clément *et al.* [3] prove convergence of the method proposed by Longstaff and Schwartz as the number of simulated trajectories, N , tends to infinity. Stentoft [5] presents a detailed numerical analysis of the Longstaff and Schwartz approach. By considering alternative families of polynomials and different numbers of basis functions, the author provides a guidance into the problem of proper selection of basis functions. He also finds that in problems with high number of assets the least-squares approach is superior to the binomial model method in terms of the trade-off between computational time and precision.

References

- [1] Brandt, M.W., Goyal, A., Santa-Clara, P. & Stroud, J.R. (2005). A simulation approach to dynamic portfolio choice with an application to learning about return predictability, *Review of Financial Studies* **18**, 831–873.
- [2] Carrière, J. (1996). Valuation of early-exercise price of options using simulations and non-parametric regression, *Insurance: Mathematics and Economics* **19**, 19–30.
- [3] Clément, E., Lamberton, D. & Protter, P. (2002). An analysis of a least squares regression algorithm for American option pricing, *Finance and Stochastics* **6**, 449–471.
- [4] Longstaff, F.A. & Schwartz, E.S. (2001). Valuing American options by simulation: a simple least-squares approach, *Review of Financial Studies* **14**, 113–147.
- [5] Stentoft, L. (2004). Assessing the least-squares Monte-Carlo approach to American option valuation, *Review of Derivatives Research* **7**, 129–168.
- [6] Tebaldi, C. (2005). Hedging using simulation: a least squares approach, *Journal of Economic Dynamics and Control* **29**, 1287–1312.
- [7] Thitsiklis, J. & Van Roy, B. (1999). Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives, *IEEE Transactions on Automatic Control* **44**, 1840–1851.
- [8] Thitsiklis, J. & Van Roy, B. (2001). Regression methods for pricing complex American-style options, *IEEE Transactions on Neural Networks* **12**, 694–703.

Related Articles

American Options; Finite Difference Methods for Early Exercise Options; Integral Equation Methods for Free Boundaries; Monte Carlo Simulation for Stochastic Differential Equations.

ADAM KOLKIEWICZ

Simulation of Square-root Processes

Square-root diffusion processes are popular in many branches of quantitative finance. Guaranteed to stay nonnegative, yet almost as tractable as a Gaussian process, the mean-reverting square-root process has found applications ranging from short-rate modeling of the term structure of interest rates, to credit derivative pricing, and to stochastic volatility models, just to name a few.

A thorough description of the theoretical properties of square-root processes as well as their generalization into multifactor affine jump-diffusion processes can be found in [7] where a full literature survey is also available. While we shall rely on some of these results in the remainder of this article, our focus here is on the problem of generating Monte Carlo paths for the one-factor square-root process, first in isolation and later in combination with a lognormal asset process (as required in most stochastic volatility applications). As we shall see, such path generation can, under many relevant parameter settings, be surprisingly challenging. Indeed, despite the popularity and the longevity of the square-root diffusion—the first uses in finance date back several decades—it is only in the last few years that a satisfactory palette of Monte Carlo algorithms has been established.

Problem Definition and Key Theoretical Results

Let $x(t)$ be a scalar random variable satisfying a stochastic differential equation (SDE) of the mean-reverting square-root type, that is

$$\begin{aligned} dx(t) &= \kappa (\theta - x(t)) dt + \epsilon \sqrt{x(t)} dW(t), \\ x(0) &= x_0 \end{aligned} \quad (1)$$

where κ, θ, ϵ are positive constants and $W(t)$ is a Brownian motion in a given probability measure. Applications of equation (1) in finance include the seminal CIR (Cox–Ingersoll–Ross) model for interest rates (see **Cox–Ingersoll–Ross (CIR) Model**) and the Heston stochastic volatility model (see **Heston Model**). In practical usage of such models (e.g., to price options), we are often faced with the problem of generating Monte Carlo paths of x on some

discrete timeline. To devise a simulation scheme, it suffices to contemplate the more fundamental question of how to generate, for an arbitrary increment Δ , a random sample of $x(t + \Delta)$ given $x(t)$; repeated application of the resulting one-period scheme produces a full path of x on an arbitrary set of discrete dates.

To aid in the construction of simulation algorithms, let us quickly review a few well-known theoretical results for equation (1).

Proposition 1 *Let $F(z; v, \lambda)$ be the cumulative distribution function for the noncentral chi-square distribution with v degrees of freedom and noncentrality parameter λ :*

$$\begin{aligned} F(z; v, \lambda) &= e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j! 2^{v/2+j} \Gamma(v/2 + j)} \\ &\quad \times \int_0^z y^{v/2+j-1} e^{-y/2} dy \end{aligned} \quad (2)$$

For the process (1) define

$$\begin{aligned} d &= 4\kappa\theta/\epsilon^2; \\ n(t, T) &= \frac{4\kappa e^{-\kappa(T-t)}}{\epsilon^2 (1 - e^{-\kappa(T-t)})}, \quad T > t \end{aligned} \quad (3)$$

Let $T > t$. Conditional on $x(t)$, $x(T)$ is distributed as $e^{-\kappa(T-t)}/n(t, T)$ times a noncentral chi-square distribution with d degrees of freedom and noncentrality parameter $x(t)n(t, T)$. That is,

$$\begin{aligned} \Pr(x(T) < x | x(t)) \\ &= F\left(\frac{x \cdot n(t, T)}{e^{-\kappa(T-t)}}; d, x(t) \cdot n(t, T)\right) \end{aligned} \quad (4)$$

From the known properties of the noncentral chi-square distribution, the following corollary easily follows.

Corollary 1 *For $T > t$, $x(T)$ has the following first two conditional moments:*

$$\begin{aligned} E(x(T) | x(t)) &= \theta + (x(t) - \theta) e^{-\kappa(T-t)} \\ \text{Var}(x(T) | x(t)) &= \frac{x(t)\epsilon^2 e^{-\kappa(T-t)}}{\kappa} \\ &\quad \times \left(1 - e^{-\kappa(T-t)}\right) + \frac{\theta\epsilon^2}{2\kappa} \left(1 - e^{-\kappa(T-t)}\right)^2 \end{aligned} \quad (5)$$

Proposition 2 Assume that $x(0) > 0$. If $2\kappa\theta \geq \epsilon^2$, then the process for x can never reach zero. If $2\kappa\theta < \epsilon^2$, the origin is accessible and strongly reflecting.

The condition $2\kappa\theta \geq \epsilon^2$ in Proposition 2 is often known as the *Feller condition* (see [12]) for equation (1). When equation (1) is used as a model for interest rates or credit spreads, market-implied model parameters are typically such that the Feller condition is satisfied. However, when equation (1) represent a stochastic variance process (as in the section Stochastic Volatility Simulation), the Feller condition rarely holds. As it turns out, a violation of the Feller condition may increase the difficulty of Monte Carlo path generation considerably.

Simulation Schemes

Exact Simulation

According to Proposition 1, the distribution of $x(t + \Delta)$ given $x(t)$ is known in closed form. Generation of a random sample of $x(t + \Delta)$ given $x(t)$ can therefore be done entirely bias-free by sampling from a noncentral chi-square distribution. Using the fact that a noncentral chi-square distribution can be seen as a regular chi-square distribution with Poisson-distributed degrees of freedom (see [9]), the following algorithm can be used.

1. Draw a Poisson random variable N , with mean $\frac{1}{2}x(t)n(t, t + \Delta)$.
2. Given N , draw a regular chi-square random variable χ_v^2 , with $v = d + 2N$ degrees of freedom.
3. Set $x(t + \Delta) = \chi_v^2 \cdot \exp(-\kappa\Delta) / n(t, t + \Delta)$.

Steps 1 and 3 of this algorithm are straightforward, but Step 2 is somewhat involved. In practice, generation of chi-squared variables would most often use one of several available techniques for the gamma distribution, a special case of which is the chi-square distribution. A standard algorithm for the generation of gamma variates of acceptance–rejection type is the Cheng–Feast algorithm [5], and a number of others are listed in [9], though direct generation by the aid of the inverse cumulative distribution function [6] is also a practically viable option.

We should note that if $d > 1$, it may be numerically advantageous to use a different algorithm, based

on the convenient relation

$$\chi_d^2(\lambda) \stackrel{d}{=} \left(Z + \sqrt{\lambda}\right)^2 + \chi_{d-1}^2, \quad d > 1 \quad (6)$$

where $\stackrel{d}{=}$ denotes equality in distribution, $\chi_d^2(\lambda)$ is a noncentral chi-square variable with d degrees of freedom and noncentrality parameter λ , and Z is an ordinary $\mathcal{N}(0, 1)$ Gaussian variable. We trust that the reader can complete the details on application of equation (6) in a simulation algorithm for $x(t + \Delta)$.

One might think that the existence of an exact simulation scheme for $x(t + \Delta)$ would settle once and for all the question of how to generate paths of the square-root process. In practice, however, several complications may arise with the application of the algorithm mentioned earlier. Indeed, the scheme is quite complex compared with many standard SDE discretization schemes and may not fit smoothly into existing software architecture for SDE simulation routines. Also, computational speed may be an issue, and the application of acceptance–rejection sampling will potentially cause a “scrambling effect” when process parameters are perturbed, resulting in poor sensitivity computations. While caching techniques can be designed to overcome some of these issues, storage, look-up, and interpolation of such a cache pose their own challenges. Further, the basic scheme above provides no explicit link between the paths of the Brownian motion $W(t)$ and that of $x(t)$, complicating applications in which, say, multiple correlated Brownian motions need to be advanced through time.

In light of the discussion earlier, it seems reasonable to also investigate the application of simpler simulation algorithms. These will typically exhibit a bias for finite values of Δ , but convenience and speed may more than compensate for this, especially if the bias is small and easy to control by reduction of step-size. We proceed to discuss several classes of such schemes.

Biased Taylor-type Schemes

Euler Schemes. Going forward, let us use $\hat{x}$ to denote a discrete-time (biased) approximation to x . A classical approach to constructing simulation schemes for SDEs involves the application of Itô–Taylor expansions, suitably truncated. See **Monte Carlo Simulation for Stochastic Differential Equations, Stochastic Differential Equations: Scenario Simulation, and Stochastic Taylor Expansions** for details.

The simplest of such schemes is the *Euler scheme*, a direct application of which would write

$$\begin{aligned}\hat{x}(t + \Delta) &= \hat{x}(t) + \kappa(\theta - \hat{x}(t))\Delta \\ &\quad + \epsilon\sqrt{\hat{x}(t)} Z\sqrt{\Delta}\end{aligned}\quad (7)$$

where Z is $\mathcal{N}(0, 1)$ Gaussian variable. One immediate (and fatal) problem with equation (7) is that the discrete process for x can become negative with nonzero probability, making computation of $\sqrt{\hat{x}(t)}$ impossible and causing the time-stepping scheme to fail. To get around this problem, several remedies have been proposed in the literature, starting with the suggestion in [13] that one simply replaces $\sqrt{\hat{x}(t)}$ in equation (7) with $\sqrt{|\hat{x}(t)|}$. Lord *et al.*, [14] review a number of such “fixes”, concluding that the following works best:

$$\begin{aligned}\hat{x}(t + \Delta) &= \hat{x}(t) + \kappa(\theta - \hat{x}(t)^+)\Delta \\ &\quad + \epsilon\sqrt{\hat{x}(t)^+} Z\sqrt{\Delta}\end{aligned}\quad (8)$$

where we use the notation $x^+ = \max(x, 0)$. In [14] this scheme is denoted “full truncation”; its main characteristic is that the process for $\hat{x}$ is allowed to go below zero, at which point the process for x becomes deterministic with an upward drift of $\kappa\theta$.

Higher-order Schemes. The scheme (8) has first-order weak convergence, in the sense that expectations of functions of x will approach their true values as $\mathcal{O}(\Delta)$. To improve convergence, it is tempting to apply a *Milstein scheme*, the most basic of which is

$$\begin{aligned}\hat{x}(t + \Delta) &= \hat{x}(t) + \kappa(\theta - \hat{x}(t))\Delta \\ &\quad + \epsilon\sqrt{\hat{x}(t)} Z\sqrt{\Delta} + \frac{1}{4}\epsilon^2\Delta (Z^2 - 1)\end{aligned}\quad (9)$$

As was the case for equation (7), this scheme has a positive probability of generating negative values of $\hat{x}$ and therefore cannot be used without suitable modifications. Kahl and Jäkel [11] list several other Milstein-type schemes, some of which allow for a certain degree of control over the likelihood of generating negative values. One particularly appealing variation is the *implicit Milstein scheme*, defined as

$$\begin{aligned}\hat{x}(t + \Delta) &= (1 + \kappa\Delta)^{-1} \cdot (\hat{x}(t) + \kappa\theta\Delta \\ &\quad + \epsilon\sqrt{\hat{x}(t)} Z\sqrt{\Delta} + \frac{1}{4}\epsilon^2\Delta (Z^2 - 1))\end{aligned}\quad (10)$$

It is easy to verify that this discretization scheme results in strictly positive paths for the x process if $4\kappa\theta > \epsilon^2$. For cases where this bound does not hold, it will be necessary to modify equation (10) to prevent problems with the computation of $\sqrt{\hat{x}(t)}$. For instance, whenever $\hat{x}(t)$ drops below zero, we could use equation (8) rather than equation (10).

Under certain sufficient regularity conditions, Milstein schemes have second-order weak convergence. Owing to the presence of a square root in equation (1), these sufficient conditions are violated here, and one should not expect equation (10) to have second-order convergence for all parameter values, even the ones that satisfy $4\kappa\theta > \epsilon^2$. Numerical tests of Milstein schemes for square-root processes can be found in [9] and [11]; overall these schemes perform fairly well in certain parameter regimes, but are typically less robust than the Euler scheme.

Moment-matching Schemes

Lognormal Approximation. The simulation schemes introduced in the section Biased Taylor-type Schemes all suffer to various degrees from an inability to keep the path of x nonnegative throughout. One, rather obvious, way around this is to draw $\hat{x}(t + \Delta)$ from a user-selected probability distribution that (i) is reasonably close to the true distribution of $x(t + \Delta)$ and (ii) is certain not to produce negative values. To ensure that (i) is satisfied, it is natural to select the parameters of the chosen distribution to match one or more of the true moments for $x(t + \Delta)$, conditional upon $x(t) = \hat{x}(t)$. For instance, if we assume that the true distribution of $x(t + \Delta)$ is well approximated by a lognormal distribution with parameters μ and σ , we write (see [2])

$$\hat{x}(t + \Delta) = e^{\mu + \sigma Z} \quad (11)$$

where Z is a Gaussian random variable, and μ, σ are chosen to satisfy

$$e^{\mu + \frac{1}{2}\sigma^2} = \mathbb{E}(x(t + \Delta)|x(t) = \hat{x}(t)) \quad (12)$$

$$e^{2\left(\mu + \frac{1}{2}\sigma^2\right)} (e^{\sigma^2} - 1) = \text{Var}(x(t + \Delta)|x(t) = \hat{x}(t)) \quad (13)$$

The results in Corollary 1 can be used to compute the right-hand sides of this system of equations, which can then easily be solved analytically for μ and σ .

As is the case for many other schemes, equation (11) works best if the Feller condition is satisfied. If not, the lower tail of the lognormal distribution is often too thin to capture the true distribution shape of $\hat{x}(t + \Delta)$ —see Figure 1.

Truncated Gaussian. Figure 1 demonstrates that the density of $\hat{x}(t + \Delta)$ may sometimes be nearly singular at the origin. To accommodate this, one could contemplate inserting an actual singularity through outright truncation at the origin of a distribution that may otherwise go negative. Using a Gaussian distribution for this, say, one could write

$$\hat{x}(t + \Delta) = (\mu + \sigma Z)^+ \quad (14)$$

where μ and σ are determined by moment-matching, along the same lines as in the section Lognormal Approximation. While this moment-matching exercise cannot be done in an entirely analytical fashion, a number of caching tricks outlined in [3] can be used to make the determination of μ and σ essentially instantaneous. As documented in [3], the scheme (14) is robust and generally has attractive convergence properties when applied to standard option pricing problems. Being fundamentally Gaussian when $\hat{x}(t)$ is far from the origin, equation (14) is somewhat similar to the Euler scheme (8), although the performance

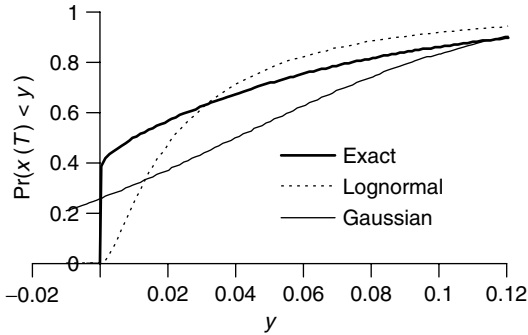


Figure 1 Cumulative distribution function for $x(T)$ given $x(0)$, with $T = 0.1$. Model parameters were $x(0) = \theta = 4\%$, $\kappa = 50\%$, and $\epsilon = 100\%$. The lognormal and Gaussian distributions in the graph were parameterized by matching mean and variances to the exact distribution of $x(T)$

of equation (14) is typically better than equation (8). Unlike equation (8), the truncated Gaussian scheme (14) also ensures, by construction, that negative values of $\hat{x}(t + \Delta)$ cannot be attained.

Quadratic-exponential. We finish our discussion of biased schemes for equation (1) with a more elaborate moment-matched scheme, based on a combination of a squared Gaussian and an exponential distribution. In this scheme, for large values of $\hat{x}(t)$, we write

$$\hat{x}(t + \Delta) = a(b + Z)^2 \quad (15)$$

where Z is a standard Gaussian random variable, and a and b are certain constants, to be determined by moment-matching. These constants a and b depend on the time step Δ and $\hat{x}(t)$, as well as the parameters in the SDE for x . While based on the well-established asymptotics for the noncentral chi-square distribution (see [3]), formula (15) does not work well for low values of $\hat{x}(t)$ —in fact, the moment-matching exercise fails to work—so we supplement it with a scheme to be used when $\hat{x}(t)$ is small. Andersen [3] shows that a good choice is to approximate the density of $\hat{x}(t + \Delta)$ with

$$\begin{aligned} &\Pr(\hat{x}(t + \Delta) \in [x, x + dx]) \\ &\approx (p\delta(x) + \beta(1 - p)e^{-\beta x}) dx, \quad x \geq 0 \end{aligned} \quad (16)$$

where δ is a Dirac delta-function, and p and β are nonnegative constants to be determined. As in the scheme in the section Truncated Gaussian, we have a probability mass at the origin, but now the strength of this mass (p) is explicitly specified, rather than implied from other parameters. The mass at the origin is supplemented with an exponential tail. It can be verified that if $p \in [0, 1]$ and $\beta \geq 0$, then equation (16) constitutes a valid density function.

Assuming that we have determined a and b , Monte Carlo sampling from equation (15) is trivial. To draw samples in accordance with equation (16), we can generate a cumulative distribution function

$$\begin{aligned} \Psi(x) &= \Pr(\hat{x}(t + \Delta) \leq x) \\ &= p + (1 - p)(1 - e^{-\beta x}), \quad x \geq 0 \end{aligned} \quad (17)$$

the inverse of which is readily computable, allowing for efficient generation of random draws by the inverse distribution method.

What remains is the determination of the constants a , b , p , and β , as well as a rule for when

to switch from equation (15) to sampling from equation (17). The first problem is easily settled by moment-matching techniques.

Proposition 3 Let $m \triangleq E(x(t + \Delta)|x(t) = \hat{x}(t))$ and $s^2 \triangleq \text{Var}(x(t + \Delta)|x(t) = \hat{x}(t))$ and set $\psi = s^2/m^2$. Provided that $\psi \leq 2$, set

$$b^2 = 2\psi^{-1} - 1 + \sqrt{2\psi^{-1} - 1} \sqrt{2\psi^{-1} - 1} \geq 0 \quad (18)$$

and

$$a = \frac{m}{1 + b^2} \quad (19)$$

Let $\hat{x}(t + \Delta)$ be as defined in equation (15); then $E(\hat{x}(t + \Delta)) = m$ and $\text{Var}(\hat{x}(t + \Delta)) = s^2$.

Proposition 4 Let m , s , and ψ be as defined in Proposition 3. Assume that $\psi \geq 1$ and set

$$p = \frac{\psi - 1}{\psi + 1} \in [0, 1) \quad (20)$$

and

$$\beta = \frac{1 - p}{m} = \frac{2}{m(\psi + 1)} > 0 \quad (21)$$

Let $\hat{x}(t + \Delta)$ be sampled from equation (17); then $E(\hat{x}(t + \Delta)) = m$ and $\text{Var}(\hat{x}(t + \Delta)) = s^2$.

The terms m , s , ψ defined in the two propositions above are explicitly computable from the result in Corollary 1. For any ψ_c in $[1, 2]$, a valid *switching rule* is to use equation (15) if $\psi \leq \psi_c$ and to sample equation (17) otherwise. The exact choice for ψ_c is noncritical; $\psi_c = 1.5$ is a good choice.

The quadratic-exponential (QE) scheme outlined above is typically the most accurate of the biased schemes introduced in this article. Indeed, in most practical applications, the bias introduced by the scheme is statistically undetectable at the levels of Monte Carlo noise acceptable in practical applications; see [3] for numerical tests under a range of challenging conditions. Variations on the QE scheme without an explicit singularity in zero can also be found in [3].

Stochastic Volatility Simulation

As mentioned earlier, square-root processes are commonly used to model stochastic movements in the volatility of some financial asset. A popular example

of such an application is the *Heston model* [10], defined by a vector SDE of the form^a

$$dY(t) = Y(t)\sqrt{x(t)} dW_Y(t) \quad (22)$$

$$dx(t) = \kappa(\theta - x(t)) dt + \epsilon\sqrt{x(t)} dW(t) \quad (23)$$

with $dW_Y(t) \cdot dW(t) = \rho dt$, $\rho \in [-1, 1]$. For numerical work, it is useful to recognize that the process for $Y(t)$ is often relatively close to geometric Brownian motion, making it sensible to work with logarithms of $Y(t)$, rather than $Y(t)$ itself. An application of Itô's Lemma shows that equations (22)–(23) are equivalent to

$$d \ln Y(t) = -\frac{1}{2}x(t) dt + \sqrt{x(t)} dW_Y(t) \quad (24)$$

$$dx(t) = \kappa(\theta - x(t)) dt + \epsilon\sqrt{x(t)} dW(t) \quad (25)$$

We proceed to consider the joint simulation of equations (24)–(25).

Broadie–Kaya Scheme

As demonstrated in [4], it is possible to simulate equations (24)–(25) bias-free. To show this, first integrate the SDE for $x(t)$ and rearrange.

$$\begin{aligned} & \int_t^{t+\Delta} \sqrt{x(u)} dW(u) \\ &= \epsilon^{-1} \left(x(t + \Delta) - x(t) - \kappa\theta\Delta + \kappa \int_t^{t+\Delta} x(u) du \right) \end{aligned} \quad (26)$$

Performing a Cholesky decomposition we can also write

$$\begin{aligned} d \ln Y(t) &= -\frac{1}{2}x(t) dt + \rho\sqrt{x(u)} dW(u) \\ &\quad + \sqrt{1 - \rho^2}\sqrt{x(u)} dW^*(u) \end{aligned} \quad (27)$$

where W^* is a Brownian motion independent of W . An integration yields

$$\begin{aligned} \ln Y(t + \Delta) &= \ln Y(t) + \frac{\rho}{\epsilon} (x(t + \Delta) - x(t) - \kappa\theta\Delta) \\ &\quad + \left(\frac{\kappa\rho}{\epsilon} - \frac{1}{2} \right) \int_t^{t+\Delta} x(u) du \\ &\quad + \sqrt{1 - \rho^2} \int_t^{t+\Delta} \sqrt{x(u)} dW^*(u) \end{aligned} \quad (28)$$

where we have used equation (26). Conditional on $x(t + \Delta)$ and $\int_t^{t+\Delta} x(u) du$, it is clear that the distribution of $\ln Y(t + \Delta)$ is Gaussian with easily computable moments. After first sampling $x(t + \Delta)$ from the noncentral chi-square distribution (as described in the section Exact Simulation), one then performs the following steps:

1. Conditional on $x(t + \Delta)$ (and $x(t)$) draw a sample of $I = \int_t^{t+\Delta} x(u) du$.
2. Conditional on $x(t + \Delta)$ and I , use equation (28) to draw a sample of $\ln Y(t + \Delta)$ from a Gaussian distribution.

While execution of the second step is straightforward, the first one is decidedly not, as the conditional distribution of the integral I is not known in closed form. In [4], the authors instead derive a characteristic function, which they numerically Fourier-invert to generate the cumulative distribution function for I , given $x(t + \Delta)$ and $x(t)$. Numerical inversion of this distribution function over a uniform random variable finally allows for generation of a sample of I . The total algorithm requires great care in numerical discretization to prevent introduction of noticeable biases and is further complicated by the fact that the characteristic function for I contains two modified Bessel functions.

The Broadie–Kaya algorithm is bias-free by construction, but its complexity and lack of speed is problematic in some applications. At the cost of introducing a (small) bias, [15] improves computational efficiency by introducing certain approximations to the characteristic function of time-integrated variance, enabling efficient caching techniques.

Other Schemes

Taylor-type Schemes. In their examination of “fixed” Euler schemes, Lord *et al.* [14] suggest simulation of the Heston model by combining equation (8) with the following scheme for $\ln Y$:

$$\begin{aligned} \ln \hat{Y}(t + \Delta) = & \ln \hat{Y}(t) - \frac{1}{2} \hat{x}(t)^+ \Delta \\ & + \sqrt{\hat{x}(t)^+} Z_Y \sqrt{\Delta} \end{aligned} \quad (29)$$

where Z_Y is a Gaussian $\mathcal{N}(0, 1)$ draw, correlated to Z in equation (8) with correlation coefficient ρ . For

the periods where $\hat{x}$ drops below zero in equation (8), the process for $\hat{Y}$ comes to a standstill.

Kahl and Jäckel [11] consider several alternative schemes for Y , the most prominent being the “IJK” scheme, defined as

$$\begin{aligned} \ln \hat{Y}(t + \Delta) = & \ln \hat{Y}(t) \\ & - \frac{\Delta}{4} (\hat{x}(t + \Delta) + \hat{x}(t)) + \rho \sqrt{\hat{x}(t)} Z \sqrt{\Delta} \\ & + \frac{1}{2} (\sqrt{\hat{x}(t + \Delta)} + \sqrt{\hat{x}(t)}) (Z_Y \sqrt{\Delta} - \rho Z \sqrt{\Delta}) \\ & + \frac{1}{4} \epsilon \rho \Delta (Z^2 - 1) \end{aligned} \quad (30)$$

Here, $\hat{x}(t + \Delta)$ and $\hat{x}(t)$ are meant to be simulated by the implicit Milstein scheme (5); again the correlation between the Gaussian samples Z_Y and Z is ρ .

Simplified Broadie–Kaya. We recall from the discussion earlier that the complicated part of the Broadie–Kaya algorithm was the computation of $\int_t^{t+\Delta} x(u) du$, conditional on $x(t)$ and $x(t + \Delta)$. Andersen [3] suggests a naive, but effective, approximation based on the idea that

$$\int_t^{t+\Delta} x(u) du \approx \Delta [\gamma_1 x(t) + \gamma_2 x(t + \Delta)] \quad (31)$$

for certain constants γ_1 and γ_2 . The constants γ_1 and γ_2 can be found by moment-matching techniques (using results from [8], p. 16), but [3] presents evidence that it will often be sufficient to use either an Euler-like setting ($\gamma_1 = 1, \gamma_2 = 0$) or a central discretization ($\gamma_1 = \gamma_2 = \frac{1}{2}$). In any case, equation (30) combined with equation (27) gives rise to a scheme for Y -simulation that can be combined with any basic algorithm that can produce $\hat{x}(t)$ and $\hat{x}(t + \Delta)$. Andersen [3] provides numerical results for the case where $\hat{x}(t)$ and $\hat{x}(t + \Delta)$ are simulated by the algorithms in the sections Truncated Gaussian and Quadratic-exponential; the results are excellent, particularly when the QE algorithm in the section Quadratic-exponential is used to sample x .

Martingale Correction. Finally, let us note that some of the schemes outlined above (including equation (29) and the one in the section Simplified Broadie–Kaya) generally do not lead to martingale-behavior of $\hat{Y}$; that is, $E(\hat{Y}(t + \Delta)) \neq E(\hat{Y}(t))$. For

the cases where the error $e = E(\hat{Y}(t + \Delta)) - E(\hat{Y}(t))$ is analytically computable, it is, however, straightforward to remove the bias by simply adding $-e$ to the sample value for $\hat{Y}(t + \Delta)$. Andersen [3] gives several examples of this idea.

Further Reading

In this article, we restricted ourselves to the presentation of relatively simple methods, which in the two-dimensional Heston model setting only require two variates per time step. Such schemes are often the most convenient in actual trading systems and for implementations that rely on Wiener processes built from low discrepancy numbers. More complicated high-order Taylor schemes, which often require extra variates, are described in [13]. The efficacy of such methods are, however, unproven in the specific setting of the Heston model.

In recent work, Alfonsi [1] constructs a second-order scheme for the CIR process, using a switching idea similar to that of the QE scheme. For the Heston process, Alfonsi develops a “second-order scheme candidate” involving three variates per time step; the numerical performance of the scheme compares favorably with Euler-type schemes.

End Notes

^aWe assume that Y is a martingale in the chosen measure; adding a drift is straightforward.

References

- [1] Alfonsi, A. (2008). *A Second-Order Discretization Scheme for the CIR Process: Application to the Heston Model*, Working Paper, Institut für Mathematik, TU, Berlin.
- [2] Andersen, L. & Brotherton-Ratcliffe, R. (2005). Extended Libor market models with stochastic volatility, *Journal of Computational Finance* **9**(1), 1–40.
- [3] Andersen, L. (2008). Simple and efficient simulation of the Heston stochastic volatility model, *Journal of Computational Finance* **11**(3), 1–42.
- [4] Broadie, M. & Kaya, Ö. (2006). Exact simulation of stochastic volatility and other affine jump diffusion processes, *Operations Research* **54**(2), 217–231.
- [5] Cheng, R. & Feast, G. (1980). Gamma variate generators with increased shape parameter range, *Communications of the ACM* **23**(7), 389–394.
- [6] DiDonato, A.R. & Morris, A.H. (1987). Incomplete gamma function ratios and their inverse, *ACM TOMS* **13**, 318–319.
- [7] Duffie, D., Pan, J. & Singleton, K. (2000). Transform analysis and asset pricing for affine jump diffusions, *Econometrica* **68**, 1343–1376.
- [8] Dufresne, D. (2001). *The Integrated Square-Root Process*, Working Paper, University of Montreal.
- [9] Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*, Springer Verlag, New York.
- [10] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**(2), 327–343.
- [11] Kahl, C. & Jäckel, P. (2006). Fast strong approximation Monte Carlo schemes for stochastic volatility models, *Journal of Quantitative Finance* **6**(6), 513–536.
- [12] Karlin, S. & Taylor, H. (1981). *A Second Course in Stochastic Processes*, Academic Press.
- [13] Kloeden, P. & Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations*, 3rd Edition, Springer Verlag, New York.
- [14] Lord, R., Koekkoek, R. & van Dijk, D. (2006). *A Comparison of Biased Simulation Schemes for Stochastic Volatility Models*, Working Paper, Tinbergen Institute, Amsterdam.
- [15] Smith, R. (2007). An almost exact simulation method for the Heston model, *Journal of Computational Finance* **11**(1), 115–125.

Related Articles

Affine Models; Cox–Ingersoll–Ross (CIR) Model; Heston Model; Monte Carlo Simulation for Stochastic Differential Equations; Stochastic Differential Equations: Scenario Simulation; Stochastic Taylor Expansions.

LEIF B.G. ANDERSEN, PETER JÄCKEL &
CHRISTIAN KAHL

Variance Reduction

Classical convergence results for the Monte Carlo method show that the ratio $\sigma/\sqrt{n}$ governs its accuracy, n being the number of drawings and σ the variance of the random variable of which we compute the expectation. Variance reduction techniques consist in modifying the classical Monte Carlo method to reduce the order of magnitude of the simulation error. The basic idea behind variance reduction techniques consists in rewriting the quantity to be computed as the expectation of a random variable that has a smaller variance.

In other words, if the quantity to be computed is the expectation $E[X]$ of a real square integrable random variable X , variance reduction methods aim at finding an alternative representation $E(X) = E(Y) + C$, using another square integrable random variable Y such that $\text{Var}(Y) \leq \text{Var}(X)$, C being a computable constant.

The most widely used variance reduction methods are “importance sampling” and “control variate” methods. As such, distinct sections are devoted to them. We also describe other classical variance reduction methods: antithetic variables, stratified sampling, and conditioning. For each method, we give simple examples related to option pricing.

Control Variates

Let X be a real-valued random variable and assume that we want to compute its expectation using a Monte Carlo method. In this method we use Y , another square integrable random variable, called the *control variate*, to write $E(X)$ as

$$E(X) = E(X - Y) + E(Y) \quad (1)$$

When $E(Y)$ can be computed using an explicit formula and $\text{Var}(X - Y)$ is smaller than $\text{Var}(X)$, we can use a Monte Carlo method to estimate $E(X - Y)$, and add the known value of $E(Y)$. Note that a variance reduction can be obtained only if X and Y are *not* independent. In fact, the more dependent are X and Y or the nearer is Y to X , the better the control variate performs.

Let us illustrate this principle by simple financial examples.

Using Call–put Arbitrage Formula for Variance Reduction

In a financial context, the price of the underlying assets is usually a good source for control variate as under a risk neutral probability the expected value of the actualized price remains constant with time.

This idea is used when taking into account the call–put arbitrage relation. Let S_t be the price at time t of an asset, and denote by C the price of the European call option

$$C = E(e^{-rT} (S_T - K)_+) \quad (2)$$

and by P the price of the European put option

$$P = E(e^{-rT} (K - S_T)_+) \quad (3)$$

There exists a relation between the price of the put and the call, which does not depend on the models for the price of the asset, namely, the “call–put arbitrage formula”:

$$C - P = E(e^{-rT} (S_T - K)) = S_0 - Ke^{-rT} \quad (4)$$

This arbitrage formula, which remains true whatever the model, can be used to replace the computation of a call option price by a put option price.

Remark 1 For the Black–Scholes model explicit formulas for the variance of the put and the call options can be obtained. Often, the variance of the put option is smaller than the variance of the call option. Note that this is not always true but since the payoff of the put is bounded, whereas the payoff of the call is not, this is certainly true when volatility is large enough.

Remark 2 Observe that call–put relations can also be obtained for Asian options or index options.

For Asian options, set $\bar{S}_T = 1/T \int_0^T S_s ds$. We have

$$\begin{aligned} E((\bar{S}_T - K)_+) - E((K - \bar{S}_T)_+) \\ = E(\bar{S}_T) - K \end{aligned} \quad (5)$$

and, in the Black–Scholes model,

$$\begin{aligned} E(\bar{S}_T) &= \frac{1}{T} \int_0^T E(S_s) ds \\ &= \frac{1}{T} \int_0^T S_0 e^{rs} ds = S_0 \frac{e^{rT} - 1}{rT} \end{aligned} \quad (6)$$

The Kemna and Vorst Method for Asian Options

A variance reduction method based on the control variate is proposed in [11] for computing the value of a fixed-strike Asian option. The price of an average (or Asian) put option with fixed strike is

$$\mathbf{E} \left(e^{-rT} \left(K - \frac{1}{T} \int_0^T S_s ds \right)_+ \right) \quad (7)$$

where $(S_t, t \geq 0)$ is the Black–Scholes model

$$S_t = x \exp \left(\left(r - \frac{\sigma^2}{2} \right) t + \sigma W_t \right) \quad (8)$$

If σ and r are small enough, an expansion of the exponential function suggest that $1/T \int_0^T S_s ds$ can be approximated by $\exp(1/T \int_0^T \log(S_s) ds)$. This heuristic argument suggests to use Y , where

$$Y = e^{-rT} (K - \exp(Z))_+ \quad (9)$$

and $Z = 1/T \int_0^T \log(S_s) ds$, as a control variate. As the random variable Z is Gaussian, we can explicitly compute $\mathbf{E}(Y)$ using the (Black–Scholes-type) formula

$$\begin{aligned} \mathbf{E} \left((K - e^Z)_+ \right) &= K N(-d) \\ &- e^{\mathbf{E}(Z) + \frac{1}{2} \text{Var}(Z)} N \left(-d - \sqrt{\text{Var}(Z)} \right) \end{aligned} \quad (10)$$

where $d = (\mathbf{E}(Z) - \log(K)) / \sqrt{\text{Var}(Z)}$.

To have a working algorithm, it remains to sample

$$e^{-rT} \left(K - \frac{1}{T} \int_0^T S_s ds \right)_+ - Y \quad (11)$$

This method can be very efficient when $\sigma \approx 0.3$ by year, $r \approx 0.1$ by year and $T \approx 1$ year. Of course, for larger values of σ and r , the gain obtained with this control variate is less significant but this method still remains useful.

Index Options

A very similar idea can be used for pricing index options. Assume that S_t is given by the multidimensional Black–Scholes model. Let σ be a $p \times d$

matrix and $W^1, \dots, W^d$ be d independent Brownian motions. Denote by $(S_t, t \geq 0)$ the solution of

$$\begin{cases} dS_t^1 = S_t^1 (r dt + [\sigma dW_t]_1), & S_0^1 = x_1 \\ \dots \\ dS_t^p = S_t^p (r dt + [\sigma dW_t]_p), & S_0^p = x_p \end{cases} \quad (12)$$

where $[\sigma dW_t]_i = \sum_{j=1}^d \sigma_{ij} dW_t^j$. Note that this equation can be solved to get, for $i = 1, \dots, p$,

$$S_T^i = x_i e^{(r-1/2 \sum_{j=1}^d \sigma_{ij}^2)T + \sum_{j=1}^d \sigma_{ij} W_T^j} \quad (13)$$

Moreover, denote by I_t the value of an index $I_t = \sum_{i=1}^p a_i S_t^i$, where $a_1, \dots, a_p$ is a given set of positive numbers such that $\sum_{i=1}^p a_i = 1$. Suppose that we want to compute the price of a European index put option with payoff at time T given by $(K - I_T)_+$.

Consider I_T/m where $m = a_1 x_1 + \dots + a_p x_p$. Because $I_0/m = 1$, an expansion of the exponential function suggests approximation of I_T/m by Y_T/m , where Y_T is the lognormal random variable

$$Y_T = m e^{\sum_{i=1}^p a_i x_i / m \left(\left\{ r - \frac{1}{2} \sum_{j=1}^d \sigma_{ij}^2 \right\} T + \sum_{j=1}^d \sigma_{ij} W_T^j \right)} \quad (14)$$

As we can explicitly compute $\mathbf{E}[(K - Y_T)_+]$ using a Black–Scholes formula, this suggests to use the control variate $Z = (K - Y_T)_+$ and to sample $(K - I_T)_+ - (K - Y_T)_+$. We refer to Figure 1 to see the improvement in variance obtained when using this control variate in a multidimensional Black–Scholes model.

A Random Volatility Model

Consider the pricing of an option in a Black–Scholes model with stochastic volatility. The price $(S_t, t \geq 0)$ is the solution of the stochastic differential equation

$$dS_t = S_t (r dt + \sigma(Y_t) dW_t), \quad S(0) = x \quad (15)$$

where σ is a bounded function and Y_t is the solution of another stochastic differential equation

$$dY_t = b(Y_t) dt + c(Y_t) dW_t', \quad Y_0 = y \quad (16)$$

where $(W_t, t \geq 0)$ and $(W_t', t \geq 0)$ are two, not necessarily independent, Brownian motions. We want to

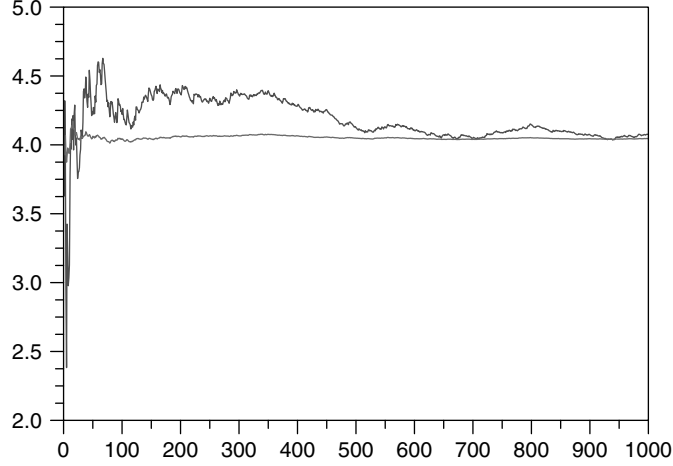


Figure 1 At the money index call option : with and without control variate, $d = 10$, $\sigma = 0.3/\sqrt{\text{year}}$ for each asset, every covariance equal to 0.5, $T = 1$

compute the price of a European option with payoff $f(S_T)$ at time T given by

$$\mathbf{E}(e^{-rT} f(S_T)) \quad (17)$$

If the volatility of the volatility (i.e., $c(Y_t)$) is not too large or if Y_t has an invariant law (as for the Ornstein–Uhlenbeck process) with mean σ_0 , we can expect σ_0 to be an acceptable approximation of σ_t .

This suggests the use of the control variate $e^{-rT} f(\bar{S}_T)$, where $\bar{S}_T$ is the solution of a Black–Scholes equation:

$$d\bar{S}_t = \bar{S}_t (r dt + \sigma_0 dW_t), \quad S(0) = x \quad (18)$$

For standard payoff f , $\mathbf{E}(e^{-rT} f(\bar{S}_T))$ can be obtained using a Black–Scholes-type formula; hence, it remains to sample

$$e^{-rT} f(S_T) - e^{-rT} f(\bar{S}_T) \quad (19)$$

and to check on simulations, using the standard estimate for the variance, that this procedure actually reduces the variance.

Using the Hedge as a Control Variate

In most standard financial models, a hedging strategy is available. This hedge can be used as a hint to construct a control variate.

Let $(S_t, t \geq 0)$ be the price of the asset. Assume that the price of the option at time t can be expressed

as $C(t, S_t)$ (this fact is satisfied for any Markovian model). When an explicit approximation $\bar{C}(t, x)$ of $C(t, x)$ is known, we can use the control variate

$$Y = \sum_{k=1}^N \frac{\partial \bar{C}}{\partial x}(t_k, S_{t_k}) \times ((S_{t_{k+1}} - S_{t_k}) - \mathbf{E}(S_{t_{k+1}} - S_{t_k})) \quad (20)$$

Note that $\mathbf{E}(Y) = 0$ by construction and so no correction is needed. If $\bar{C}$ is close to C and if N is large enough, a large reduction in the variance can be obtained.

Optimizing a Set of Control Variates

Assume that $Y = (Y^1, \dots, Y^n)$ is a given set of control variates with 0 expectation (or more generally having a known expectation) and finite variance. It is quite easy to optimize the control variate among all linear combinations of the coordinate of Y .

Let us denote by λ an $\mathbf{R}^n$ vector. As for every λ

$$\mathbf{E}(X) = \mathbf{E}(X - \langle \lambda, Y \rangle) \quad (21)$$

we can use $\langle \lambda, Y \rangle$ as a control variate and it is natural to choose λ to be the minimizer of the variance $\text{Var}(X - \langle \lambda, Y \rangle)$. A simple computation shows that this minimizer is given by

$$\lambda^* = \Gamma_Y^{-1} \text{Cov}(X, Y) \quad (22)$$

when Γ_Y , the covariance matrix of the vector Y , is invertible and where $\text{Cov}(X, Y) = (\text{Cov}(X, Y_i), 1 \leq i \leq n)$. Note that the optimizing λ^* can be estimated using independent samples of the law of (X, Y) , $((X_1, Y_1), \dots, (X_n, Y_n))$ and the standard estimators of the variances and the covariances of the random variables. This leads to a convergent estimator $\hat{\lambda}_n := \hat{\lambda}_n(X_1, Y_1, \dots, X_n, Y_n)$ of λ^* .

Using $\hat{\lambda}_n$ with an independent sample $((X'_1, Y'_1), \dots, (X'_n, Y'_n))$ leads to a convergent and *unbiased* estimator

$$E_n^1 = \frac{1}{n} \sum_{i=1}^n X'_i - \langle \hat{\lambda}_n, Y'_n \rangle \quad (23)$$

whereas using the same drawings leads to a convergent but *biased* estimator

$$E_n^2 = \frac{1}{n} \sum_{i=1}^n X_i - \langle \hat{\lambda}_n, Y_n \rangle \quad (24)$$

The bias of the second estimator is negligible, at least for large samples as it can be shown that the two estimators follow the same central limit theorem:

$$\sqrt{n} (E_n^i - \mathbf{E}(X)) \rightarrow \mathcal{N}(0, (\sigma^*)^2) \quad (25)$$

for $i = 1$ or 2 and where $(\sigma^*)^2$ is the best available variance

$$(\sigma^*)^2 = \min_{\lambda \in \mathbf{R}^d} \text{Var}(X - \langle \lambda, Y \rangle) \quad (26)$$

See [5, 12] for details and proofs on this technique known as *adaptive control variates*.

Perfect Control Variates for Diffusion Models

An interesting result is that perfect (zero-variance) control variates exist for diffusion models. Although the argumentation is mainly theoretical, it can give hints for implementation.

We want to compute $\mathbf{E}(Z)$ where $Z = \psi(X_s, 0 \leq s \leq T)$ and $(X_s, s \geq 0)$ is the solution of

$$dX_t = b(X_t) dt + \sigma(X_t) dW_t, \quad X(0) = x \quad (27)$$

We assume that $(X_t, t \geq 0)$ is $\mathbf{R}^n$ valued and $(W_t, t \geq 0)$ is an $\mathbf{R}^d$ -valued Brownian motion.

The *predictable representation theorem* shows that we are often able (at least theoretically) to cancel the variance using a stochastic integral as a control variate.

Theorem 1 (Predictable Representation Theorem). *Let Z be a random variable such that $\mathbf{E}(Z^2) < +\infty$. Assume that Z is measurable with respect to $\sigma(W_s, s \leq T)$. Then there exists a stochastic process $(H_t, t \leq T)$ adapted to the Brownian filtration, such that $\mathbf{E}\left(\int_0^T H_s^2 ds\right) < +\infty$ and*

$$Z = \mathbf{E}(Z) + \int_0^T H_s dW_s \quad (28)$$

For a proof we refer to [10, 15].

Remark 3 Note that Z needs to be measurable with respect to the σ -field generated by the Brownian motion.

The theorem shows that, in principle, we are able to cancel the variance of Z using a stochastic integral as a control variate. Nevertheless, the explicit computation of $(H_s, s \leq T)$ is much more complicated than the one of $\mathbf{E}(Z)$! The reader is referred to [14] for formulas for H_s involving Malliavin derivatives and conditional expectations. In financial applications, empirical methods are often used instead.

When the price of the underlying asset is described by a Markovian model X_t , the process $(H_t, t \leq T)$ can be written as $H_t = v(t, X_t)$, v being a function of t and x often related to the hedge in the context of financial models.

Theorem 2 *Let b and σ be two Lipschitz continuous functions. Let $(X_t, t \geq 0)$ be the unique solution of*

$$dX_t = b(X_t) dt + \sigma(X_t) dW_t, \quad X_0 = x \quad (29)$$

Denote by A the infinitesimal generator of this diffusion

$$\begin{aligned} Af(x) &= \frac{1}{2} \sum_{i,j=1}^n a_{ij}(x) \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \\ &\quad + \sum_{j=1}^n b_j(x) \frac{\partial f}{\partial x_j}(x) \end{aligned} \quad (30)$$

where $a_{ij}(x) = \sum_{k=1}^d \sigma_{ik}(x) \sigma_{jk}(x)$.

Assume that there exists a $C^{1,2}([0, T] \times \mathbf{R}^d)$ function, with bounded derivatives in x , as solution to the problem

$$\begin{cases} \text{For } (t, x) \in [0, T] \times \mathbf{R}^n, \\ \left(\frac{\partial u}{\partial t} + Au \right)(t, x) = 0, \\ u(T, x) = g(x), \quad x \in \mathbf{R}^n \end{cases} \quad (31)$$

Then if $Z = g(X_T)$ and

$$Y = \int_0^T \frac{\partial u}{\partial x}(s, X_s) \sigma(s, X_s) dW_s \quad (32)$$

we have

$$\mathbf{E}(Z) = Z - Y \quad (33)$$

The random variable Y is, thus, a perfect control variate for Z .

Proof Using Itô's formula, we obtain

$$\begin{aligned} du(t, X_t) &= \left(\frac{\partial u}{\partial t} + Au \right)(t, X_t) dt \\ &\quad + \frac{\partial u}{\partial x}(t, X_t) \sigma(s, X_s) dW_t \end{aligned} \quad (34)$$

Now, integrate between 0 and T and take the expectation of both sides of the equality. Using the facts that u is a solution of equation (31) and that the stochastic integral is a martingale, we get

$$u(0, x) = Z - Y = \mathbf{E}(Z) \quad (35)$$

Remark 4 Theorem 2 shows that we only need to look for H_t as a function of t and X_t when X is a diffusion process. However, the explicit formula involves partial derivatives with respect to x and it is numerically difficult to take advantage of it.

In a practical situation, we can use the following heuristic procedure. Assume that we know an approximation $\bar{u}$ for u . The previous theorem suggests to use

$$Y = \int_0^T \frac{\partial \bar{u}}{\partial x}(t, X_t) \sigma(s, X_s) dW_t \quad (36)$$

as a control variate. Note that for every function $\bar{u}$ (even a bad approximation of u), we obtain an unbiased estimator for $\mathbf{E}(Z)$ by setting $Z' = Z - Y$. For a reasonable choice of $\bar{u}$, we can expect an improvement of the variance of the estimator. Indeed, set

$$\bar{Z} := g(X_T) - \int_0^T \frac{\partial \bar{u}}{\partial x}(t, X_t) \sigma(X_t) dW_t \quad (37)$$

$\bar{Z}$ is an unbiased estimator of $\mathbf{E}(g(X_T))$ and

$$\begin{aligned} &\mathbf{E}(|\bar{Z} - \mathbf{E}g(X_T)|^2) \\ &= \mathbf{E} \left(\int_0^T \left| \left(\frac{\partial u}{\partial x} - \frac{\partial \bar{u}}{\partial x} \right)(t, X_t) \right|^2 \sigma(t, X_t)^2 dt \right) \end{aligned} \quad (38)$$

This variance can be expected to be small if $\partial \bar{u} / \partial x$ is a good approximation of $\partial u / \partial x$.

Importance Sampling

Importance sampling methods proceed by *changing the law* of the samples. Assume that X takes its values in $\mathbf{R}^d$, ϕ is a bounded function from $\mathbf{R}^d$ to $\mathbf{R}$, and we want to compute $\mathbf{E}(\phi(X))$. The aim is to find a new random variable Y following a different law and a function i such that $\mathbf{E}(\phi(X)) = \mathbf{E}(i(Y)\phi(Y))$. The function i , the *importance function*, is needed to maintain the equality of the expectations for every function f . Obviously, this method is interesting in a Monte Carlo method only if $\text{Var}(i(Y)\phi(Y))$ is smaller than $\text{Var}(\phi(X))$.

Consider X as an $\mathbf{R}^d$ -valued random variable following a law with density $f(x)$ for which we want to compute $\mathbf{E}(\phi(X))$

$$\mathbf{E}(\phi(X)) = \int_{\mathbf{R}^d} \phi(x) f(x) dx \quad (39)$$

If $\tilde{f}$ is any density on $\mathbf{R}^d$ such that $\tilde{f}(x) > 0$ and $\int_{\mathbf{R}^d} \tilde{f}(x) dx = 1$, clearly one can rewrite $\mathbf{E}(\phi(X))$ as

$$\begin{aligned} \mathbf{E}(\phi(X)) &= \int_{\mathbf{R}^d} \frac{\phi(x) f(x)}{\tilde{f}(x)} \tilde{f}(x) dx \\ &= \mathbf{E} \left(\frac{\phi(Y) f(Y)}{\tilde{f}(Y)} \right) \end{aligned} \quad (40)$$

where Y is a random variable with density law $\tilde{f}(x)$ under $\mathbf{P}$. Hence, $\mathbf{E}(\phi(X))$ can be approximated by an alternative estimator

$$\frac{1}{n} \left(\frac{\phi(Y_1) f(Y_1)}{\tilde{f}(Y_1)} + \dots + \frac{\phi(Y_n) f(Y_n)}{\tilde{f}(Y_n)} \right) \quad (41)$$

where $(Y_1, \dots, Y_n)$ are independent copies of Y . Denoting $Z = \phi(Y) f(Y) / \tilde{f}(Y)$, this estimator will have a smaller asymptotic variance than the standard one if $\text{Var}(Z) < \text{Var}(\phi(X))$. Note that the variance of Z is given by

$$\text{Var}(Z) = \int_{\mathbf{R}} \frac{g^2(x) f^2(x)}{\tilde{f}(x)} dx - \mathbf{E}(\phi(X))^2 \quad (42)$$

An easy computation shows that when $\phi(x) > 0$, for every $x \in \mathbf{R}^d$, the choice of $\tilde{f}(x) = \phi(x) f(x) / \mathbf{E}(\phi(X))$ leads to a zero-variance estimator as

$\text{Var}(Z) = 0$. Of course, this result can seldom be used in practice as it relies on the exact knowledge of $\mathbf{E}(\phi(X))$, which is exactly what we need to compute. Nevertheless, it can lead to a useful heuristic approach : choose for $\tilde{f}(x)$ a good approximation of $|\phi(x)f(x)|$ such that $\tilde{f}(x)/\int_{\mathbf{R}} \tilde{f}(x) dx$ can be sampled easily.

An Elementary Gaussian Example

In finance, importance sampling is especially useful when computing multidimensional Gaussian expectations as all computations and simulations are completely explicit.

Let G be a Gaussian random variable with mean zero and unit variance. We want to compute $\mathbf{E}(\phi(G))$, for a function ϕ . We choose for the new sampling law, its shifted valued $\tilde{G} = G + m$, m being a real constant to be determined later; hence,

$$\mathbf{E}(\phi(G)) = \mathbf{E}\left(\phi(\tilde{G}) \frac{f(\tilde{G})}{\tilde{f}(\tilde{G})}\right) \quad (43)$$

where f is the density of the law of G and $\tilde{f}$ the density of the law of $\tilde{G}$. Easy computation leads to

$$\begin{aligned} \mathbf{E}(\phi(G)) &= \mathbf{E}\left(\phi(\tilde{G}) e^{-m\tilde{G} + (m^2/2)}\right) \\ &= \mathbf{E}\left(\phi(G + m) e^{-mG - (m^2/2)}\right) \end{aligned} \quad (44)$$

As a simple example of the use of equation (44), assume that we want to compute a European call option in the Black–Scholes model; ϕ is then given by

$$\phi(G) = (\lambda e^{\sigma G} - K)_+ \quad (45)$$

where λ , σ , and K are positive constants. When $\lambda \ll K$, $\mathbf{P}(\lambda e^{\sigma G} > K)$ is very small and the option will, very unlikely, be exercised. This fact leads to a very large relative error when using a standard Monte Carlo method. To increase to exercise probability, we can use equality (44) to obtain

$$\begin{aligned} &\mathbf{E}\left((\lambda e^{\sigma G} - K)_+\right) \\ &= \mathbf{E}\left((\lambda e^{\sigma(G+m)} - K)_+ e^{-mG - (m^2/2)}\right) \end{aligned} \quad (46)$$

and choose $m = m_0$ with $\lambda e^{\sigma m_0} = K$, since

$$\mathbf{P}(\lambda e^{\sigma(G+m_0)} > K) = \frac{1}{2} \quad (47)$$

This choice of m is certainly not optimal; however, it drastically improves the efficiency of the Monte Carlo method when $\lambda \ll K$. See Figure 2 for an illustration of the improvement, which can be obtained using this idea.

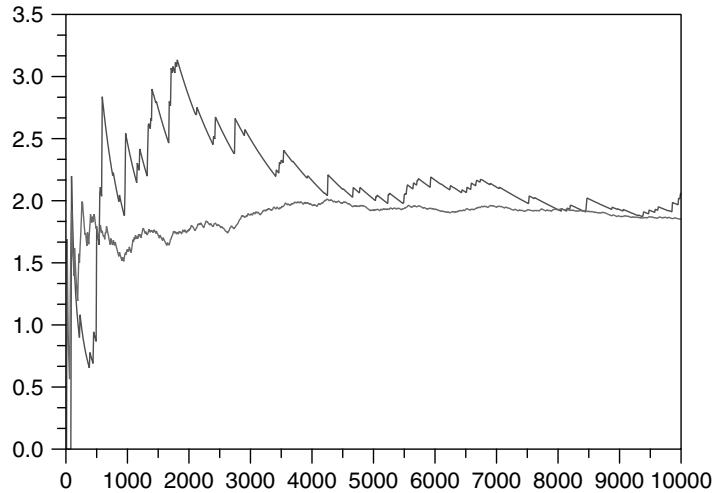


Figure 2 Call options: use of importance sampling for deep out-of-the-money call option $\sigma = 0.3/\sqrt{\text{year}}$, $S_0 = 70$, $K = 100$

A Multidimensional Gaussian Case: Index Options

The previous method can easily be extended to multidimensional Gaussian cases. Let us start by motivating this in the context of index option pricing. Denote by $(S_t, t \geq 0)$ a multidimensional Black–Scholes model solution of equation (13)

$$S_T^i = S_0^i \exp \left(\left(r - \frac{1}{2} \sum_{j=1}^d \sigma_{ij}^2 \right) T + \sum_{j=1}^d \sigma_{ij} W_T^j \right) \quad (48)$$

Moreover, denote I_t by the value of an index $I_t = \sum_{i=1}^n a_i S_t^i$, where $a_1, \dots, a_n$ is a given set of positive numbers such that $\sum_{i=1}^n a_i = 1$. Suppose that we want to compute the price of a European call or put option with payoff at time T given by $f(I_T)$. Obviously, there exists a function ϕ such that

$$f(I_T) = \phi(G_1, \dots, G_d) \quad (49)$$

where $G_j = W_T^j / \sqrt{T}$. The price of this option can be rewritten as $\mathbf{E}(\phi(G))$, where $G = (G_1, \dots, G_d)$ is a d -dimensional Gaussian vector with unit covariance matrix.

As in the one-dimensional case, it is easy (by a change of variable) to prove that if $m = (m_1, \dots, m_d)$, then

$$\mathbf{E}(\phi(G)) = \mathbf{E}(\phi(G + m) e^{-mG - (|m|^2/2)}) \quad (50)$$

where $m \cdot G = \sum_{i=1}^d m_i G_i$ and $|m|^2 = \sum_{i=1}^d m_i^2$. In view of equation (50), the second moment $V(m)$ of the random variable $X_m = \phi(G + m) e^{-m \cdot G - |m|^2/2}$ is

$$\begin{aligned} V(m) &= \mathbf{E}(\phi^2(G + m) e^{-2mG - |m|^2}) \\ &= \mathbf{E}(\phi^2(G) e^{-mG + |m|^2/2}) \end{aligned} \quad (51)$$

The choice of a minimizing m (or an approximation) is more difficult than in one dimension. From the previous formula, it follows that $V(m)$ is a strictly convex function, a property from which we can derive useful approximation methods for a minimizer of $V(m)$. The reader is referred to [6] for an almost optimal way to choose the parameter m or to [1, 2] for a use of stochastic algorithms to get convergent approximation of the m minimizing the variance $V(m)$.

The Girsanov Theorem and Path-dependent Options

We can extend further these techniques to path-dependent options, using the Girsanov theorem. Let $(S_t, t \geq 0)$ be the solution of

$$dS_t = S_t (r dt + \sigma dW_t), \quad S_0 = x \quad (52)$$

where $(W_t, t \geq 0)$ is a Brownian motion under a probability $\mathbf{P}$. We want to compute the price of a path-dependent option with payoff given by

$$\phi(S_t, t \leq T) = \psi(W_t, t \leq T) \quad (53)$$

Common examples of such a situation are

- Asian options whose payoff is given by $f(S_T, \int_0^T S_s ds)$ and
- Maximum options whose payoff is given by $f(S_T, \max_{s \leq T} S_s)$.

We start by considering the Brownian case that is a straightforward extension of the technique used in the preceding paragraph. For every real number λ , define the process $(W_t^\lambda, t \leq T)$ as

$$W_t^\lambda := W_t + \lambda t \quad (54)$$

According to the Girsanov theorem, $(W_t^\lambda, t \leq T)$ is a Brownian motion under the probability law $\mathbf{P}^\lambda$ defined by

$$\mathbf{P}^\lambda(A) = \mathbf{E}(L_T^\lambda \mathbf{1}_A), \quad A \in \mathcal{F}_T \quad (55)$$

where $L_T^\lambda = e^{-\lambda W_T - \lambda^2 T/2}$. Denote by $\mathbf{E}^\lambda$ the expectation under this new probability $\mathbf{P}^\lambda$. For every bounded function ψ we have

$$\begin{aligned} \mathbf{E}(\psi(W_t, t \leq T)) &= \mathbf{E}^\lambda(\psi(W_t^\lambda, t \leq T)) \\ &= \mathbf{E}(L_T^\lambda \psi(W_t^\lambda, t \leq T)) \end{aligned} \quad (56)$$

and thus

$$\begin{aligned} \mathbf{E}(\psi(W_t, t \leq T)) &= \mathbf{E}(e^{-\lambda W_T - (\lambda^2 T/2)} \psi(W_t + \lambda t, t \leq T)) \end{aligned} \quad (57)$$

For example, if we want to compute the price of a fixed-strike Asian option P given by

$$\mathbf{E} \left(e^{-rt} \left(\frac{1}{T} \int_0^T x e^{(r - (\sigma^2/2))s + \sigma W_s} ds - K \right) \right) \quad (58)$$

we can use the previous equality to obtain

$$P = \mathbf{E} \left(e^{-rt - \lambda W_T - \lambda^2 T/2} \left(\frac{1}{T} \int_0^T x e^{(r - \sigma^2/2)s + \sigma(W_s + \lambda s)} ds - K \right)_+ \right) \quad (59)$$

This representation can be used for deep out-of-the-money options (that is to say, $x \ll K$). Then λ can be chosen such that

$$\frac{x}{T} \int_0^T e^{(r - \sigma^2/2)s + \sigma \lambda s} ds = K \quad (60)$$

in order to increase the exercise probability.

Importance Sampling for the Poisson Process

Similar results can also be obtained for Poisson processes and can be useful to construct variance reduction methods for financial models with jumps.

Let $(N_s^\lambda, s \geq 0)$ be a Poisson process with intensity λ . Denote by

$$L_T^{\mu \rightarrow \lambda} = e^{T(\mu - \lambda)} \left(\frac{\lambda}{\mu} \right)^{N_T^\mu} \quad (61)$$

We have for every bounded functional f

$$\mathbf{E} \left(f(N_s^\lambda, s \leq T) \right) = \mathbf{E} \left(L_T^{\mu \rightarrow \lambda} f(N_s^\mu, s \leq T) \right) \quad (62)$$

See [4] for a proof and extensions. Note that the variance is given by $V(\lambda) = \mathbf{E}(X_\mu^2) - \mathbf{E}(X_\mu)^2$ where $X_\mu = L_T^{\mu \rightarrow \lambda} f(N_s^\mu, s \leq T)$ and

$$\begin{aligned} \mathbf{E}(X_\mu^2) &= \mathbf{E} \left(e^{T(\mu - \lambda)} \left(\frac{\lambda}{\mu} \right)^{N_T^\lambda} \right. \\ &\quad \left. \times f^2(N_t^\lambda, 0 \leq t \leq T) \right) \end{aligned} \quad (63)$$

From this formula, a convexity property in μ can be derived and optimization algorithms deduced from the Euler equation associated with the variance minimization problem.

This methodology can be useful for jump models in finance (e.g., the Merton model) by mixing a change of law on the underlying Brownian motion and on Poisson process.

Importance Sampling for Diffusion Processes

Following [14], we now present a result which proves that variance can be canceled using importance sampling for a diffusion process. The reader is also referred to [13] for the necessary background on simulation of diffusion processes.

Proposition 1 *Let Z be a random variable such that $Z = \psi(W_s, 0 \leq s \leq T)$, $\mathbf{E}(Z^2) < +\infty$ and $\mathbf{P}(Z \geq \epsilon) = 1$, for an $\epsilon > 0$.*

Then, there exists an adapted process $(h_t, 0 \leq t \leq T)$, such that, if

$$L_T = \exp \left(- \int_0^T h_s dW_s - \frac{1}{2} \int_0^T |h_s|^2 ds \right) \quad (64)$$

$\mathbf{E}(L_T) = 1$, then we can define a probability $\tilde{\mathbf{P}}$ by

$$\tilde{\mathbf{P}}(A) = \mathbf{E}(L_T \mathbf{1}_A) \quad (65)$$

Under this probability $\tilde{\mathbf{P}}$

$$\tilde{\mathbf{E}}(L_T^{-1} Z) = \mathbf{E}(Z) \text{ and } \tilde{\text{Var}}(L_T^{-1} Z) = 0 \quad (66)$$

Remark 5 Under $\tilde{\mathbf{P}}$, the random variable $L_T^{-1} Z$ has zero variance, and thus is almost surely constant. So if we are able to sample $L_T^{-1} Z$ under $\tilde{\mathbf{P}}$, we obtain a zero-variance estimator for $\mathbf{E}(Z)$.

Of course, an effective computation of h_t is almost always impossible. However, heuristic approximation methods can also be derived. We refer to [14] for an overview of some of these methods.

Proof The representation theorem for Brownian martingales proves the existence of a process $(H_t, t \leq T)$ such that, for $t \leq T$

$$\mathbf{E}(Z | \mathcal{F}_t) = \mathbf{E}(Z) + \int_0^t H_s dW_s \quad (67)$$

Let $\phi_t = \mathbf{E}(Z | \mathcal{F}_t) / \mathbf{E}(Z)$. Equality (67) becomes

$$\phi_t = 1 + \int_0^t \frac{H_s}{\mathbf{E}(Z)} dW_s = 1 - \int_0^t \phi_s h_s dW_s \quad (68)$$

This is a linear equation for ϕ , which can be solved to obtain

$$\begin{aligned} \phi_T &= \exp \left(- \int_0^T h_s dW_s - \frac{1}{2} \int_0^T |h_s|^2 ds \right) \\ &= L_T \end{aligned} \quad (69)$$

However, as Z is an $\mathcal{F}_T$ -measurable random variable, $L_T = \phi_T = Z/\mathbf{E}(Z)$. Thus, $\mathbf{E}(L_T) = 1$ and $L_T^{-1}Z = \mathbf{E}(Z)$ almost surely under $\mathbf{P}$ and $\tilde{\mathbf{P}}$.

The previous theorem can be used in a simulation context for diffusion processes. Let $(X_t, t \geq 0)$ be the unique solution of

$$\begin{aligned} dX_t &= b(X_t) dt + \sigma(X_t) dW_t, \\ X(0) &= x \end{aligned} \quad (70)$$

where b and σ are Lipschitz functions and $(W_t, t \geq 0)$ is a Brownian motion. If $(h_t, t \geq 0)$ is a process such that $\mathbf{E}(L_T) = 1$, then X is also a solution of

$$\begin{cases} dX_t = (b(X_t) - \sigma(X_t)h_t) dt + \sigma(X_t) d\tilde{W}_t \\ X(0) = x \end{cases} \quad (71)$$

where $(\tilde{W}_t = W_t + \int_0^t h_s ds, 0 \leq t \leq T)$ is a Brownian motion under the probability $\tilde{\mathbf{P}}$. Hence, we can, in principle, sample the process X under the new probability $\tilde{\mathbf{P}}$. This is easy when h_t can be written as $v(t, X_t)$, v being a function of t and x . Indeed, in this case, X satisfies the stochastic differential equation

$$dX_t = \tilde{b}(t, X_t) dt + \sigma(X_t) d\tilde{W}_t, \quad X(0) = x \quad (72)$$

with $\tilde{b}(t, x) = b(x) - \sigma(x)v(t, x)$. Since $\tilde{W}$ is a Brownian motion under $\tilde{\mathbf{P}}$, we can simulate X under $\tilde{\mathbf{P}}$ using a standard discretization scheme for this stochastic differential equation.

Now we give a more explicit formula for h_t when the random variable is given by

$$Z = e^{-\int_0^T r(X_s) ds} f(X_T) \quad (73)$$

when f and r are bounded positive functions and $(X_t, t \geq 0)$ is the $\mathbf{R}^d$ -diffusion process solution of equation (70). Note that such a Z has a form suitable for the computation of option or zero-coupon prices.

Assume that there exists a $C^{1,2}([0, T] \times \mathbf{R}^d)$ solution to

$$\begin{cases} u(T, x) = f(x) & \text{for } x \in \mathbf{R}^n \\ \left(\frac{\partial u}{\partial t} + Au - ru \right)(t, x) = 0 \\ \text{for } (t, x) \in [0, T] \times \mathbf{R}^n \end{cases} \quad (74)$$

where A is the infinitesimal generator of the diffusion $(X_t, t \geq 0)$ given by equation (30). Then

$$\begin{aligned} e^{-\int_0^t r(X_s) ds} u(t, X_t) &= u(0, x) \\ &+ \int_0^t e^{-\int_0^s r(X_u) du} \frac{\partial u}{\partial x}(s, X_s) \sigma(X_s) dW_s \end{aligned} \quad (75)$$

Assuming that the stochastic integral is a martingale, we have

$$\begin{aligned} &\mathbf{E} \left(e^{-\int_0^T r(X_s) ds} f(X_T) | \mathcal{F}_t \right) \\ &= \mathbf{E} \left(e^{-\int_0^T r(X_s) ds} u(T, X_T) | \mathcal{F}_t \right) \\ &= e^{-\int_0^t r(X_s) ds} u(t, X_t) \end{aligned} \quad (76)$$

Thus, we have $Z = \mathbf{E}(Z) + \int_0^T H_t dW_t$ with

$$H_t = e^{-\int_0^t r(X_s) ds} \frac{\partial u}{\partial x}(t, X_t) \sigma(X_t)$$

and

$$\begin{aligned} &\mathbf{E} \left(e^{-\int_0^T r(X_s) ds} f(X_T) | \mathcal{F}_t \right) \\ &= e^{-\int_0^t r(X_s) ds} u(t, X_t) \end{aligned} \quad (77)$$

Therefore, using the proof of Proposition 1, we can see that the process h_t defined by

$$\begin{aligned} h_t &= - \frac{e^{-\int_0^t r(X_s) ds} \frac{\partial u}{\partial x}(t, X_t) \sigma(X_t)}{e^{-\int_0^t r(X_s) ds} u(t, X_t)} \\ &= - \frac{\frac{\partial u}{\partial x}(t, X_t) \sigma(X_t)}{u(t, X_t)} \end{aligned} \quad (78)$$

allows to cancel the variance of the Monte Carlo method.

Remark 6 Note that, as h_t is a function of t and X_t , simulation is always possible in principle.

In practical terms, when we know (even rough) approximations $\bar{u}(t, x)$ of $u(t, x)$, it is natural to try to reduce the variance by substituting $\bar{u}$ for u in the previous formula. The reader is referred to [17] to see how large deviation theory can give a good approximation $\bar{u}$ and lead to effective variance reductions.

Antithetic Variables

Antithetic variables are widely used in Monte Carlo simulations because of generality and ease of implementation. Note, though, that it seldom leads to significant effects on the variance and, if not used with care, it can even lead to an increase of the variance.

First, let us consider a simple example, $I = \mathbf{E}(g(U))$ where U is uniformly distributed on the interval $[0, 1]$. As $1 - U$ has the same law as U

$$I = \mathbf{E}\left(\frac{1}{2}(g(U) + g(1 - U))\right) \quad (79)$$

Therefore, one can draw $2n$ independent random variables $U_1, \dots, U_{2n}$ following a uniform law on $[0, 1]$, and approximate I either by using

$$I_{2n}^0 = \frac{1}{2n}(g(U_1) + g(U_2) + \dots + g(U_{2n-1}) + g(U_{2n}))$$

or

$$I_{2n} = \frac{1}{2n}(g(U_1) + g(1 - U_1) + \dots + g(U_n) + g(1 - U_n)) \quad (80)$$

We can now compare the variances of I_{2n} and I_{2n}^0 . Observe that, in doing this, we assume that most of the numerical work relies on the evaluation of g and the time devoted to the simulation of the random variables is negligible; this is often a realistic assumption.

The variance of the two estimators are given, respectively, by

$$\begin{aligned} \text{Var}(I_{2n}^0) &= \frac{1}{2n} \text{Var}(g(U_1)) \\ \text{Var}(I_{2n}) &= \frac{1}{2n} (\text{Var}(g(U_1)) + \text{Cov}(g(U_1), g(1 - U_1))) \end{aligned} \quad (81)$$

Obviously, $\text{Var}(I_{2n}) \leq \text{Var}(I_{2n}^0)$ if and only if $\text{Cov}(g(U_1), g(1 - U_1)) \leq 0$. When g is either an increasing or a decreasing function, this can be shown to be true and thus the Monte Carlo method using antithetic variables (I_{2n}) is better than the standard one (I_{2n}^0).

This simple idea can be greatly generalized. If X is a random variable taking values in $\mathbf{R}^d$ (or even an infinite dimension space) and if T operates on $\mathbf{R}^d$ in

such a way that the law of X is preserved by T , we can construct a generalized antithetic method based on the equality

$$\mathbf{E}(g(X)) = 1/2 \mathbf{E}(g(X) + g(T(X))) \quad (82)$$

In other words, if $(X_1, \dots, X_{2n})$ are sampled along the law of X , we can consider the estimator

$$I_{2n} = \frac{1}{2n}(g(X_1) + g(T(X_1)) + \dots + g(X_n) + g(T(X_n))) \quad (83)$$

and compare it to

$$I_{2n}^0 = \frac{1}{2n}(g(X_1) + g(X_2) + \dots + g(X_{2n-1}) + g(X_{2n})) \quad (84)$$

The same computations as before prove that the estimator I_{2n} is better than the crude one I_{2n}^0 if and only if $\text{Cov}(g(X), g(T(X))) \leq 0$.

A Generic Example

Let T be the transformation from $\mathbf{R}^d$ to $\mathbf{R}^d$ defined by

$$(u^1, \dots, u^d) \rightarrow (1 - u^1, \dots, 1 - u^d) \quad (85)$$

Obviously, if $U = (U^1, \dots, U^d)$ is a vector of independent random variables, then the law of $T(U)$ is identical to the law of U . Hence, we can construct an antithetic estimator I_{2n} . It can be shown that this estimator improves upon the standard one when $f(u_1, \dots, u_d)$ is monotonic in each of its coordinates.

A Toy Financial Example

Let G be a standard Gaussian random variable. Clearly, the law of $-G$ and G are identical and we can consider the transformation $T(x) = -x$ to construct an antithetic method.

A very simple illustration is given by the call option in the Black–Scholes model where the payoff can be written as

$$\mathbf{E}\left((\lambda e^{\sigma G} - K)_+\right) \quad (86)$$

λ , σ , and K being positive real numbers. As this payoff is increasing as a function of G , the antithetic

estimator

$$I_{2n} = \frac{1}{2n} (g(G_1) + g(-G_1) + \dots + g(G_n) + g(-G_n)) \quad (87)$$

with $g(x) = (\lambda e^{\sigma x} - K)_+$, certainly reduces the variance.

This example can be easily extended to the (more useful) multidimensional Gaussian case, when $G = (G^1, \dots, G^d)$ are independent standard Gaussian random variables.

Antithetic Variables for Path-dependent Options

The antithetic variables method can also be applied to path-dependent options. For this, consider a path-dependent payoff $\psi(S_s, s \leq T)$, where $(S_t, t \geq 0)$ follows the Black–Scholes model

$$S_t = x \exp((r - 1/2\sigma^2)t + \sigma W_t) \quad (88)$$

$(W_t, t \geq 0)$ is a Brownian motion, r and σ positive real numbers. As the law of $(-W_t, t \geq 0)$ is identical to the law of $(W_t, t \geq 0)$

$$\begin{aligned} \mathbf{E} \left(\psi \left(x e^{(r-1/2\sigma^2)s + \sigma W_s}, s \leq T \right) \right) \\ = \mathbf{E} \left(\psi \left(x e^{(r-(1/2)\sigma^2)s - \sigma W_s}, s \leq T \right) \right) \end{aligned} \quad (89)$$

An antithetic method can be constructed using this equality.

Stratified Sampling

Stratified sampling aims at decomposing the computation of an expectation into specific subsets (called strata). Suppose we want to compute $I = \mathbf{E}(g(X))$, where X is an $\mathbf{R}^d$ -valued random variable and g a bounded measurable function from $\mathbf{R}^d$ to $\mathbf{R}$.

Let $(D_i, 1 \leq i \leq m)$ be a partition of $\mathbf{R}^d$. I can be expressed as

$$I = \sum_{i=1}^m \mathbf{E}(g(X)|X \in D_i) \mathbf{P}(X \in D_i) \quad (90)$$

where

$$\mathbf{E}(g(X)|X \in D_i) = \frac{\mathbf{E}(\mathbf{1}_{\{X \in D_i\}} g(X))}{\mathbf{P}(X \in D_i)} \quad (91)$$

Note that $\mathbf{E}(g(X)|X \in D_i)$ can be interpreted as $\mathbf{E}(g(X^i))$, where X^i is a random variable whose law is the law of X conditioned to belong to D_i . When X has a density given by $f(x)$, this conditional law also has a density given by

$$\frac{1}{\int_{D_i} f(y) dy} \mathbf{1}_{\{x \in D_i\}} f(x) dx$$

When we further assume that the numbers $p_i = \mathbf{P}(X \in D_i)$ can be explicitly computed, one can use a Monte Carlo method to approximate each conditional expectation $I_i = \mathbf{E}(g(X)|X \in D_i)$ by

$$\tilde{I}_i = \frac{1}{n_i} (g(X_1^i) + \dots + g(X_{n_i}^i)) \quad (92)$$

where $(X_1^i, \dots, X_{n_i}^i)$ are independent copies of X^i . An estimator $\tilde{I}$ of I is then given by

$$\tilde{I} = \sum_{i=1}^m p_i \tilde{I}_i \quad (93)$$

Of course, the samples used to compute $\tilde{I}_i$ are supposed to be independent and so the variance of $\tilde{I}$ is $\sum_{i=1}^m p_i^2 (\sigma_i^2 / n_i)$, where σ_i^2 be the variance of $g(X^i)$.

By fixing the total number of simulations $\sum_{i=1}^m n_i = n$ and minimizing the above variance, we obtain an *optimal allocation of points* in the strata

$$n_i = n \frac{p_i \sigma_i}{\sum_{i=1}^m p_i \sigma_i} \quad (94)$$

For these values of n_i , the variance of $\tilde{I}$, given in this case by $1/n (\sum_{i=1}^m p_i \sigma_i)^2$, is always smaller than the one obtained without stratification.

Remark 7 The optimal stratification involves the σ_i s which are almost never explicitly known. So one needs to estimate these σ_i s by some preliminary Monte Carlo simulations.

Moreover, let us underline that a bad repartition of n_i may *increase* the variance of the estimator.

A common way to circumvent these difficulties is to choose a *proportional repartition*: $n_i = n p_i$. The corresponding variance $1/n \sum_{i=1}^m p_i \sigma_i^2$, is still smaller than the original one, but not optimal. This choice is often made especially when the probabilities p_i are explicit.

For more considerations on the choice of the n_i and for hints on suitable choices of the sets D_i , see [3].

A Toy Financial Example

In the standard Black–Scholes, model the price of a call option is given by

$$\mathbf{E} \left((\lambda e^{\sigma G} - K)_+ \right) \quad (95)$$

It is natural to use the following strata for G : $\{G \leq d\}$ or $\{G > d\}$, where $d = \log(K/\lambda)/\sigma$. Of course the variance of the stratum $G \leq d$ is equal to 0. So, in the optimal allocation, we have to simulate only one point in this stratum: all other points have to be drawn in the stratum $G \geq d$. This can be done by using the (numerical) inverse of the cumulative distribution function of a Gaussian random variable.

Index Options

A European call or put index option in the multi-dimensional Black–Scholes model can be expressed as $\mathbf{E}(h(G))$, for $G = (G_1, \dots, G_n)$ a vector of independent standard Gaussian random variables and for some complicated but explicit function h from $\mathbf{R}^n$ to $\mathbf{R}$.

Now, choose a vector $u \in \mathbf{R}^n$ such that $|u| = 1$ (so $\langle u, G \rangle = \sum_{i=1}^n u_i G_i$ is also a standard Gaussian random variable) and a partition $(B_i, 1 \leq i \leq n)$ of $\mathbf{R}$ such that

$$\mathbf{P}(\langle u, G \rangle \in B_i) = \mathbf{P}(G_1 \in B_i) = 1/n \quad (96)$$

This can be done by setting $B_i =]N^{-1}((i-1)/n), N^{-1}(i/n)]$, where N is the cumulative distribution function of a standard Gaussian random variable and N^{-1} is its inverse. Then define the $\mathbf{R}^n$ -strata by $D_i = \{u \in \mathbf{R}^n, \langle u, x \rangle \in B_i\}$.

In order to implement a stratification method based on these strata, we need first to sample the Gaussian random variable $\langle u, G \rangle$ given that $\langle u, G \rangle$ belongs to B_i , then to sample the vector G when knowing already the value $\langle u, G \rangle$.

The first point is easy since, if U is uniformly distributed on $[0, 1]$, then the law of $N^{-1}((i-1)/n) + (U/N)$ is precisely the law of a standard Gaussian random variable conditioned to be in B_i .

To solve the second point, observe that $G - \langle u, G \rangle u$ is a Gaussian vector independent of

$\langle u, G \rangle$. So if Y is a copy of the vector G , $G = \langle u, G \rangle u + G - \langle u, G \rangle u$ and $\langle u, G \rangle u + Y - \langle u, Y \rangle u$ have the same distribution. This leads to a very simple simulation method for G given $\langle u, G \rangle = \lambda$ and to an effective way to implement the suggested stratification method.

Note that this method can be made efficient by choosing a good vector u . An almost optimal way to choose the vector u can be found in [6].

Conditioning

This method uses the well-known fact that conditioning reduces the variance. Indeed, for any square integrable random variable Z , we have

$$\mathbf{E}(Z) = \mathbf{E}(\mathbf{E}(Z|\mathcal{B})) \quad (97)$$

where $\mathcal{B}$ is any σ -algebra defined on the same probability space as Z . When, in addition, Z is square integrable, the conditional expectation is an L^2 projection so

$$\mathbf{E}(\mathbf{E}(Z|\mathcal{B})^2) \leq \mathbf{E}(Z^2) \quad (98)$$

and thus $\text{Var}(\mathbf{E}(Z|\mathcal{B})) \leq \text{Var}(Z)$.

When Y is a random variable defined on the same probability space as X and $\mathcal{B} = \sigma(Y)$, it is well known that $\mathbf{E}(Z|\sigma(Y))$ can be written as

$$\mathbf{E}(Z|Y) := \mathbf{E}(Z|\sigma(Y)) = \phi(Y) \quad (99)$$

for some measurable function ϕ and the practical efficiency of simulating $\phi(Y)$ instead of Z heavily relies on getting an explicit formula for the function ϕ . This can be achieved when $Z = f(X, Y)$, where X and Y are independent random variables. In this case, we have

$$\mathbf{E}(f(X, Y)|Y) = \phi(Y) \quad (100)$$

where $\phi(y) = \mathbf{E}(f(X, y))$.

A Basic Example

Suppose that we want to compute $\mathbf{P}(X \leq Y)$, where X and Y are independent random variables. This occurs in finance, in a slightly more complex setting, when computing the hedge of an exchange option (or the price of a digital exchange option). We have

$$\mathbf{P}(X \leq Y) = \mathbf{E}(F(Y)) \quad (101)$$

where $F(x) = \mathbf{P}(X \leq x)$ is the distribution function of X . This can be used to obtain a variance reduction which can be significant, especially when the probability $\mathbf{P}(X \leq Y)$ is small.

A Financial Example: A Stochastic Volatility Model

Let $(W_t, t \geq 0)$ be a Brownian motion and r be a real number. Assume that $(S_t, t \geq 0)$ follows a Black–Scholes model with stochastic volatility, which is solution of

$$dS_t = S_t (r dt + \sigma_t dW_t), \quad S_0 = x \quad (102)$$

where $(\sigma_t, t \geq 0)$ is a given continuous stochastic process independent of the Brownian motion $(W_t, t \geq 0)$. We want to compute the option price

$$\mathbf{E} (e^{-rT} f(S_T)) \quad (103)$$

where f is a bounded measurable function. Clearly S_T can be expressed as

$$S_T = x \exp \left(rT - \frac{1}{2} \int_0^T \sigma_t^2 dt + \int_0^T \sigma_t dW_t \right) \quad (104)$$

As the processes $(\sigma_t, t \geq 0)$ and $(W_t, t \geq 0)$ are independent, $\left(\int_0^T \sigma_t^2 dt, \int_0^T \sigma_t dW_t \right)$ has the same law as

$$\left(\int_0^T \sigma_t^2 dt, \sqrt{\frac{1}{T} \int_0^T \sigma_t^2 dt} \times W_T \right) \quad (105)$$

Conditioning with respect to the process $(\sigma_t, 0 \leq t \leq T)$, we obtain

$$\begin{aligned} \mathbf{E} (e^{-rT} f(S_T)) &= \mathbf{E} (\mathbf{E} [e^{-rT} f(S_T) | \sigma_t, 0 \leq t \leq T]) \\ &= \mathbf{E} (\psi(\sigma_t, 0 \leq t \leq T)) \end{aligned} \quad (106)$$

where, for a fixed volatility path $(v_t, 0 \leq t \leq T)$,

$\psi(v_t, 0 \leq t \leq T)$

$$\begin{aligned} &= \mathbf{E} \left(e^{-rT} f \left(x e^{rT - \int_0^T v_t^2/2 dt + W_T \sqrt{1/T \int_0^T v_t^2 dt}} \right) \right) \\ &= \phi \left(\sqrt{\frac{1}{T} \int_0^T v_t^2 dt} \right) \end{aligned} \quad (107)$$

$\phi(\sigma)$ being the price of the option in the standard Black–Scholes model with volatility σ , that is

$$\phi(\sigma) = \mathbf{E} \left(e^{-rT} f \left(x e^{(r - (\sigma^2/2))T + \sigma W_T} \right) \right) \quad (108)$$

Hence, we only need to sample $\int_0^T \sigma_t^2 dt$ in order to use a Monte Carlo method using the random variable $\psi(\sigma_t, 0 \leq t \leq T)$.

Additional References

For complements, we refer the reader to classical books devoted to Monte Carlo methods ([7, 9, 16, 18]). For a more specific discussion of Monte Carlo methods in finance see [5, 8].

References

- [1] Arouna, B. (2003/2004). Robbins-Monro algorithms and variance reduction in finance, *The Journal of Computational Finance* 7(2), 35–61.
- [2] Arouna, B. (2004). Adaptative Monte-Carlo method, a variance reduction technique, *Monte Carlo Methods Application* 10(1), 1–24.
- [3] Cochran, W.G. (1977). *Sampling Techniques, Series in Probabilities and Mathematical Statistics*, Wiley.
- [4] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes, CRC Financial Mathematics Series*, Chapman & Hall.
- [5] Glasserman, P. (2004). Monte-Carlo methods in financial engineering, *Applications of Mathematics (New York), Stochastic Modelling and Applied Probability*, Springer-Verlag, New York, Vol. 53.
- [6] Glasserman, P., Heidelberger, P. & Shahabuddin, P. (1999). Asymptotically optimal importance sampling and stratification for pricing path dependent options, *Mathematical Finance* 9(2), 117–152.
- [7] Hammersley, J. & Handscomb, D. (1979). *Monte-Carlo Methods*, Chapman & Hall, London.
- [8] Jäkel, P. (2002). *Monte-Carlo Methods in Finance*, Wiley.
- [9] Kalos, M.H. & Whitlock, P.A. (1986). *Monte Carlo Methods*, John Wiley & Sons.
- [10] Karatzas, I. & Shreve, S.E. (1991). *Brownian Motion and Stochastic Calculus*, 2nd Edition, Springer-Verlag, New York.
- [11] Kemna, A.G.Z. & Vorst, A.C.F. (1990). A pricing method for options based on average asset values, *Journal of Banking Finance* 14, 113–129.
- [12] Kim, S. & Anderson, S.G. (2004). Winter Simulation Conference, *Proceedings of the 36th conference on Winter simulation*, Washington, D.C.
- [13] Kloeden, P.E. & Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations, Applications of*

- Mathematics (New York)*, 3rd Edition, Springer-Verlag, Berlin, Vol. 23.
- [14] Newton, N.J. (1994). Variance reduction for simulated diffusions, *SIAM Journal on Applied Mathematics* **54**(6), 1780–1805.
- [15] Revuz, D. & Yor, M. (1991). *Continuous Martingales and Brownian Motion*, Springer-Verlag, Berlin.
- [16] Ripley, B.D. (1987). *Stochastic Simulation*, Wiley.
- [17] Fournié, E., Lasry, J.M. & Touzi, N. (1997). Monte Carlo methods for stochastic volatility models, in *Numerical methods in finance*, Publ. Newton Inst., Rogers, L.C.G. *et al.*, ed., Cambridge University Press, Cambridge 146–164.
- [18] Rubinstein, R.Y. (1981). *Simulation and the Monte-Carlo Method, Series in Probabilities and Mathematical Statistics*, Wiley.

Related Articles

Monte Carlo Greeks; Monte Carlo Simulation for Stochastic Differential Equations; Option Pricing: General Principles; Rare-event Simulation; Simulation of Square-root Processes.

BERNARD LAPEYRE

Weighted Monte Carlo

Weighted Monte Carlo (WMC) is the name given to an algorithm used to build and calibrate asset-pricing models for financial derivatives. The algorithm combines two fixtures of the toolbox of quantitative modeling. One is Monte Carlo simulation to generate “paths” for rates and market prices on which derivatives are written [7, 9]. The other is the maximum entropy (ME) criterion, used to calculate *a posteriori* statistical weights for the paths. ME is one of the main tools in science for calculating *a posteriori* probabilities in the presence of known constraints associated with the probability measure (see [8] for classical econometric applications of ME).

The essence of the method is as follows [3]: let $X_t(\omega)$, $0 \leq t \leq T$, $\omega \in \Omega$ represent a model for the evolution of market variables or factors of interest. One of the most common applications is the case when X_t is a multivariate diffusion or jump diffusion process, *for example*,

$$dX_{\alpha t} = \sum_j \sigma_{\alpha j} dW_{jt} + \mu_{\alpha} dt, \quad 1 \leq \alpha \leq n, \quad 1 \leq j \leq m \quad (1)$$

This process represents an *a priori* model for the joint forward evolution of the market. The parameters of the model, σ, μ typically correspond to econometrically estimated factors and expected returns. We note that, since the model is used for pricing derivatives, some of the parameters can also be implied from the prices of at-the-money options and forward prices. In the language of financial economics, the measure induced by X_t is either the “physical measure” or a hybrid of the physical measure and a risk-neutral measure with respect to select observable forwards and implied volatilities.

A Monte Carlo simulation of the ensemble with N paths is generated numerically, where the paths are denoted by ω_k , that is, they can be viewed as a sampling of the probability space Ω . The WMC algorithm calibrates the Monte Carlo model so that it fits the current market prices of M *benchmarks* or reference European-style derivatives, with discounted payoffs $g_1(\omega), g_2(\omega), \dots, g_N(\omega)$ and prices $c_1, c_2, \dots, c_M$. We denote the discounted payoffs along the simulated

paths by

$$G_{ik} = g_i(\omega_k), \quad i = 1, \dots, M, \quad k = 1, \dots, N \quad (2)$$

WMC associates a probability $p_k, k = 1, \dots, N$ to each path, in such a way that the pricing equations

$$c_i = \sum_{k=1}^N G_{ik} p_k \quad (3)$$

or $\mathbf{c} = \mathbf{G}\mathbf{p}$ in vector notation, hold for all indices i . Clearly, equation (3) states that the model reprices correctly the M reference instruments. In general, we assume that the number of simulation paths is much larger than the number of benchmarks (options, forwards), which is what happens in practical situations. The choice of the probabilities is done by applying the criterion of ME, that is, by maximizing

$$H(p_1, \dots, p_N) = - \sum_{k=1}^N p_k \log p_k \quad (4)$$

subject to the M constraints in equation (3). A *least-squares* version of the algorithm least squares weighted Monte Carlo (LSWMC) proposes to solve the problem

$$\min_{\mathbf{p}} \left\{ \sum_{i=1}^M \left(\sum_{k=1}^N G_{ik} p_k - c_i \right)^2 - 2\epsilon H(\mathbf{p}) \right\} \quad (5)$$

Here, $\epsilon > 0$ is a tolerance parameter that must be adjusted by the user. If $\epsilon \ll 1$, LSWMC corresponds to the classical WMC. For finite, relatively small, values of ϵ the algorithm returns, an approximate solution of equation (3). In practice, the implementation (5) is recommended, since a solution will exist for arbitrary data $\{G_{ik}, c_i\}$.

Dual Formulation

The WMC (LSWMC) algorithm is usually solved in its *dual form*. Define the partition function

$$Z(\lambda_1, \dots, \lambda_M) = \sum_{k=1}^N e^{\sum_{i=1}^M \lambda_i G_{ik}} \quad (6)$$

where $\lambda_1, \dots, \lambda_M$ are Lagrange multipliers. The dual problem is

$$\min_{\lambda} \left\{ \log Z(\lambda) - \sum_{i=1}^M c_i \lambda_i + \frac{\epsilon}{2} \sum_{i=1}^M \lambda_i^2 \right\} \quad (7)$$

The advantage of solving the dual problem is that the number of variables is M , hence much less than the number of simulated paths. It is well known that the latter problem is convex in λ and always admits a solution if $\epsilon > 0$. Furthermore, the probabilities are given explicitly in terms of the multipliers, which solve the dual problem, namely,

$$p_k = \frac{1}{Z} e^{\sum_{i=1}^M \lambda_i G_{ik}} \quad k = 1, 2, \dots, N \quad (8)$$

In practical implementations, the dual problem can be solved with a gradient-based convex optimization routine such as L-BFGS.

Connection with Kullback–Leibler Relative Entropy

We can view WMC as an algorithm that minimizes, in a discrete setting, the relative entropy, or Kullback–Leibler distance between the prior probability measure induced by the paths (3) (call it P_0) and the posterior measure induced by the probability vector $\mathbf{p}$ (call it P), in the sense that it provides a solution of

$$\min_P \left\{ 2\epsilon D(P||P_0) + \sum_{i=1}^M [E^P(g_i(\omega)) - c_i]^2 \right\} \quad (9)$$

with

$$D(P||P_0) = E^P \left(\log \frac{dP}{dP_0} \right) \quad (10)$$

where $\frac{dP}{dP_0}$ is the Radon–Nikodym derivative of P with respect to P_0 . The latter interpretation, however, should be taken with a grain of salt since the implementation is always done in the discrete setting, ensuring that the relative entropy between two measures defined on the paths of the MC simulation is always well defined (unlike in the continuous limit, where absolute continuity in Wiener space is often a stringent condition).

Connection with Utility Maximization

It can be shown, *via* an analysis of the dual problem, that the WMC algorithm gives a pricing measure, which corresponds to optimal investment by a representative investor in the reference instruments when this investor has an exponential utility [6].

Main Known Applications

Some of the most well-known applications of this method have been in the context of multiasset equity derivatives. In this case, the *a priori* measure corresponds to a multidimensional diffusion for stock prices, generated using a factor model (or model for the correlation matrix). The *a posteriori* measure is generated by calibrating to traded options on several underlying assets. For instance, the underlying stocks can be the components of the Nasdaq 100 index and the reference instruments all listed options on the underlying stocks. In the latter case, some care must be taken with the fact that listed options are American-style, but this difficulty can be overcome by generating prices of European options using the implied volatilities of the traded options. This yields a calibrated multiasset pricer for derivatives defined on the components of the Nasdaq 100. As a general rule, it is recommended to calibrate to forward prices (zero-strike calls) in addition to options, to ensure put–call parity in the *a posteriori* measure. The value $\epsilon = 0.25$ seems to give results that are within the bid–ask spread of listed options contracts [1, 2].

Another application of WMC is to the calibration of volatility surfaces for foreign-exchange (FX) options, to obtain a volatility surface that matches forward prices, at-the-money options, strangles, and risk-reversals on all available maturities. Owing to the nature of quotes in FX, the recommended value for the tolerance parameter should be of the order of 10^{-4} in this case [1, 2].

Applications of WMC have been also proposed in the context of credit derivatives, most notably for calibrating so-called top-down models [5].

Dispersion Trading

Dispersion trading corresponds to buying and selling index options and hedging with options on the component stocks. WMC gives a method for obtaining a model price for index options based on a model,

which incorporates a view of the correlation between stocks (expressed in the *a priori* probability for X_t) and is calibrated to all the options on the components of the index. Comparing the model price (or implied volatility) with the implied volatility of index options quoted in the market provides a rational setting for comparing the prices of index options with the prices of options on the components of the index. One of the important features of WMC is that it allows the user to incorporate views on the volatility skew/smile of the components in the valuation process [1, 2].

Connection to Control Variates

The WMC framework can be generalized to any concave function $H(\mathbf{p})$. Avellaneda and Gamba [4] suggest, as one practical approach,

$$H(\mathbf{p}) = -\sum_{k=1}^N \Psi(p_k) \quad (11)$$

with $\Psi(\cdot)$ being any convex function. Obvious choices are

$$\Psi^{(Q)}(p) = \left(p - \frac{1}{N}\right)^2 \quad (\text{Quadratic}) \quad (12)$$

$$\Psi^{(S)}(p) = p \log p \quad (\text{Shannon}) \quad (13)$$

The problem of minimization of $-H(\mathbf{p})$ subject to the constraints (probability normalization and calibration)

$$1 = \mathbf{p}^\top \mathbf{n} \quad \text{and} \quad \mathbf{c} = G\mathbf{p} \quad (14)$$

with the diagonal vector $\mathbf{n} := (1, \dots, 1)^\top \in \mathbb{R}^N$ (to simplify summation notation), leads to the Lagrange function

$$\mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \mu) = -H(\mathbf{p}) - \boldsymbol{\lambda}^\top (G\mathbf{p} - \mathbf{c}) - \mu (\mathbf{p}^\top \mathbf{n} - 1) \quad (15)$$

Assuming the existence of an extremum, solving

$$\nabla_{\mathbf{p}} \mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \mu) = 0 \quad (16)$$

gives

$$p_k = \psi^{-1}(\boldsymbol{\lambda}^\top G\mathbf{e}_k + \mu) \quad (17)$$

with $\psi(p) = \frac{d\Psi(p)}{dp}$ and $\mathbf{e}_k$ being the unit vector along the k -th axis in $\mathbb{R}^N$. For the specific choices (12) and (13), this means

$$p_k^{(Q)} = \frac{1}{N} + \frac{1}{2} \boldsymbol{\lambda}^\top G(\mathbf{e}_k - \frac{1}{N} \mathbf{n}) \quad (\text{Quadratic}) \quad (18)$$

$$p_k^{(S)} = \frac{e^{\boldsymbol{\lambda}^\top G\mathbf{e}_k}}{Z(\boldsymbol{\lambda})} \quad (\text{Shannon}) \quad (19)$$

with $Z(\boldsymbol{\lambda}) = \sum_{k=1}^N e^{\boldsymbol{\lambda}^\top G\mathbf{e}_k}$, where we have eliminated μ using the probability normalization condition $1 = \mathbf{p}^\top \mathbf{n}$. Substituting equation (18) and, respectively, equation (19), back into equation (15) leads to the Lagrange dual functions

$$\hat{\mathcal{L}}^{(Q)}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^\top (\mathbf{c} - G\frac{\mathbf{n}}{N}) + \frac{N}{2} \boldsymbol{\lambda}^\top G \left(\frac{\mathbf{n}\mathbf{n}^\top}{N^2} - \frac{1}{N} \right) G^\top \boldsymbol{\lambda} \quad (20)$$

$$\hat{\mathcal{L}}^{(S)}(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^\top \mathbf{c} - \log Z(\boldsymbol{\lambda}) \quad (21)$$

The dual formulation of the original problem is to find

$$\arg \max_{\boldsymbol{\lambda}} \hat{\mathcal{L}}(\boldsymbol{\lambda}) \quad (22)$$

For the quadratic case, this is guaranteed to have at least one solution given by the linear system

$$\frac{N}{2} G \left(\frac{1}{N} - \frac{\mathbf{n}\mathbf{n}^\top}{N^2} \right) G^\top \boldsymbol{\lambda}^{(Q)} = \mathbf{c} - G\frac{\mathbf{n}}{N} \quad (23)$$

Note that

$$G \left(\frac{1}{N} - \frac{\mathbf{n}\mathbf{n}^\top}{N^2} \right) G^\top = \langle \mathbf{g}\mathbf{g}^\top \rangle_N^{P_0} - \langle \mathbf{g} \rangle_N^{P_0} \langle \mathbf{g} \rangle_N^{P_0\top} \quad (24)$$

$$= \langle \mathbf{g}, \mathbf{g} \rangle_N^{P_0} \quad (25)$$

where $\langle \cdot \rangle_N^{P_0}$ stands for the Monte Carlo estimator of the expectation under the original measure P_0 computed as the plain average over the N simulated paths, and $\langle \cdot, \cdot \rangle_N^{P_0}$ for the according covariance (defined such that $\langle \mathbf{a}, \mathbf{b} \rangle = \langle \mathbf{b}, \mathbf{a} \rangle^\top$). In other words,

$$\boldsymbol{\lambda}^{(Q)} = \frac{2}{N} \langle \mathbf{g}, \mathbf{g} \rangle_N^{P_0^{-1}} \cdot \left(\mathbf{c} - \langle \mathbf{g} \rangle_N^{P_0} \right) \quad (26)$$

and

$$\mathbf{p}^{(Q)} = \frac{\mathbf{n}}{N} + \left(\frac{\mathbf{n}}{N} - \frac{\mathbf{n}\mathbf{n}^\top}{N^2} \right) G^\top \cdot \langle \mathbf{g}, \mathbf{g} \rangle_N^{P_0^{-1}} \cdot \left(\mathbf{c} - \langle \mathbf{g} \rangle_N^{P_0} \right) \quad (27)$$

Note that the inverse of the autocovariance matrix of the calibration instruments is to be understood in a Moore–Penrose sense to safeguard against the singular case.

When using these probabilities for the valuation of a payoff v , with $v_k := v(\omega_k)$, we arrive at

$$\langle v \rangle_N^{P^{(Q)}} = \mathbf{v}^\top \mathbf{p}^{(Q)} \quad (28)$$

$$= \langle v \rangle_N^{P_0} + \langle v, \mathbf{g} \rangle_N^{P_0} \langle \mathbf{g}, \mathbf{g} \rangle_N^{P_0^{-1}} \cdot \left(\mathbf{c} - \langle \mathbf{g} \rangle_N^{P_0} \right) \quad (29)$$

which is identical to the classic control variate rule [7, 9].

For $\hat{\mathcal{L}}^{(S)}(\boldsymbol{\lambda})$, a second-order expansion in $\boldsymbol{\lambda}$ around zero gives

$$\begin{aligned} \hat{\mathcal{L}}^{(S)} &= -\log N + \boldsymbol{\lambda}^\top \left(\mathbf{c} - \langle \mathbf{g} \rangle_N^{P_0} \right) \\ &\quad - \frac{1}{2} \boldsymbol{\lambda}^\top \langle \mathbf{g}, \mathbf{g} \rangle_N^{P_0} \boldsymbol{\lambda} + \mathcal{O}(\boldsymbol{\lambda}^3) \end{aligned} \quad (30)$$

and hence we obtain the analytical initial guess

$$\boldsymbol{\lambda}^{(S),1} := \frac{N}{2} \boldsymbol{\lambda}^{(Q)} \quad (31)$$

for any iterative procedure to solve (22). A simple algorithm can be based on a second-order expansion of $\hat{\mathcal{L}}^{(S)}(\boldsymbol{\lambda})$ around the previous iteration's estimate for $\boldsymbol{\lambda}^{(S)}$. This gives

$$\begin{aligned} \boldsymbol{\lambda}^{(S),i+1} &= \boldsymbol{\lambda}^{(S),i} \\ &\quad + \left(G \Pi^{(S),i} G^\top - G \mathbf{p}^{(S),i} \mathbf{p}^{(S),i\top} G^\top \right)^{-1} \\ &\quad \times \left(\mathbf{c} - G \mathbf{p}^{(S),i} \right) \end{aligned} \quad (32)$$

with

$$\Pi^{(S),i} := \text{diag} \left(p_1^{(S),i}, \dots, p_N^{(S),i} \right) \in \mathbb{R}^{N \times N} \quad (33)$$

and

$$p_k^{(S),i} = p_k^{(S)}(\boldsymbol{\lambda}^{(S),i}) \quad (34)$$

as defined in equation (19). Interestingly, the term

$$G \mathbf{p}^{(S),i}$$

in equation (32) is the vector of expectations for the M calibration instruments under the (numerical) measure $P^{(S),i}$ defined by the numerically computed vector of probabilities $\mathbf{p}^{(S),i}$, and

$$\left(G \Pi^{(S),i} G^\top - G \mathbf{p}^{(S),i} \mathbf{p}^{(S),i\top} G^\top \right) \quad (35)$$

is the associated (numerical) covariance matrix of the calibration instruments. The simple algorithm is thus, in a formal notation, to start with $\boldsymbol{\lambda}^{(S),0} = 0$ (in all entries of the vector), to compute

$$\mathbf{p}^{(S),i} = \mathbf{p}^{(S)}(\boldsymbol{\lambda}^{(S),i}) \quad (36)$$

using equation (19), to proceed to

$$\begin{aligned} \boldsymbol{\lambda}^{(S),i+1} &= \boldsymbol{\lambda}^{(S),i} + \langle \mathbf{g}, \mathbf{g} \rangle_N^{P^{(S),i-1}} \\ &\quad \cdot \left(\mathbf{c} - \langle \mathbf{g} \rangle_N^{P^{(S),i}} \right) \end{aligned} \quad (37)$$

and to $\mathbf{p}^{(S),i+1}$, and so on.

It is, in general, possible that a solution to equation (22) may not exist for $\hat{\mathcal{L}}^{(S)}$ if the model's initial calibration implies prices for the calibration instruments that are too far away from $\mathbf{c}$. When this happens, any iterative procedure will experience that $\hat{\mathcal{L}}^{(S)}(\boldsymbol{\lambda})$ grows at an ever decreasing rate in some direction in $\mathbb{R}^M$, and, eventually, the solver will terminate when it hits an internal minimum-progress criterion. A numerical approximation for $\boldsymbol{\lambda}^{(S)}$ computed in this way will represent an ME best-possible fit, and is still usable in a vein similar to that obtained by the least squares approach mentioned in the beginning. An inexpensive warning indication for this situation is given when any of the $p_k^{(Q)}$ are negative. Note that this then also signals that the classic control variate method implicitly uses a (numerical) measure that is not equivalent to the original model's measure, which in turn may result in arbitrageable prices.

Hedge Ratios

The fact that the fine tuning of the pricing measure P is achieved by varying the probabilities of the paths such that hedge instruments are correctly repriced allows for the calculation of hedge ratios *without recalibration* of the original model, and *without resimulation*. This can be seen as follows. We seek to compute the sensitivity of $\langle v \rangle_N^P$ with respect to the calibration prices $\mathbf{c}$. Since the probability vector $\mathbf{p}(\boldsymbol{\lambda})$ is computed as an analytical function of the Lagrange multipliers, which in turn are computed numerically from $\mathbf{c}$, we have

$$\nabla_{\mathbf{c}} \langle v \rangle_N^P = \nabla_{\mathbf{c}} (\mathbf{p}^\top \mathbf{v}) = J \cdot \nabla_{\boldsymbol{\lambda}} \mathbf{p}^\top \mathbf{v} \quad (38)$$

with the elements of the Jacobian matrix J given by

$$J_{lm} = \partial_{c_l} \lambda_m \quad (39)$$

Given any Ψ , which, together with the calibration constraints, ultimately *defines* our desired pricing measure P , we can combine equation (17) and the probability normalization condition $1 = \mathbf{p}^\top \mathbf{n}$ to arrive at

$$\nabla_c \langle v \rangle_N^P = s^{P_H} \cdot J \cdot \langle \mathbf{g}, v \rangle_N^{P_H} \quad (40)$$

where we have defined the hedge measure P_H in terms of the (numerical) probabilities $\mathbf{p}^{P_H}$ whose elements are given by

$$s^{P_H} := \sum_{k=1}^N 1/\psi'(p_k^P) \quad \text{with} \quad p_k^{P_H} := \frac{1/\psi'(p_k^P)}{s^{P_H}} \quad (41)$$

What remains to be calculated is the Jacobian J . This can be done in one of three ways, depending on the choice of Ψ :

1. Analytically (explicitly). For instance, for $\Psi^{(Q)}$, we obtain $s^{P_H^{(Q)}} = 1$, $p_k^{P_H^{(Q)}} = \frac{1}{N}$, $P_H^{(Q)} = P_0$, and therefore

$$\nabla_c \langle v \rangle_N^{P^{(Q)}} = \langle \mathbf{g}, \mathbf{g} \rangle_N^{P_0} \cdot \langle \mathbf{g}, v \rangle_N^{P_0} \quad (42)$$

2. Numerically. If λ^P is computed by an iterative procedure that starts with no information other than the simulated paths and $\mathbf{c}$ itself as indicated in equation (32), the chain rule propagation can be derived and implemented as part of the iterative procedure. This approach may have to be chosen if no solution to equation (22) exists. An alternative would be to precalibrate the original model better such that a Monte Carlo weighting scheme can be found that reprices the calibration instruments exactly.
3. Analytically (implicitly). As long as a solution to equation (22) exists, that is, as long as the WMC scheme reprices the calibration instruments correctly, we can use the fact that $\nabla_c \langle \mathbf{g} \rangle_N^P$ must be the $M \times M$ identity matrix. This gives the generic result

$$\nabla_c \langle v \rangle_N^P = \langle \mathbf{g}, \mathbf{g} \rangle_N^{P_H} \cdot \langle \mathbf{g}, v \rangle_N^{P_H} \quad (43)$$

which means that for any P_0 and P , that is, Ψ , that permit perfect repricing of the hedges (calibration instruments) under P , the hedge ratios for any payoff v can be seen as a regression of the covariances between v and the hedge instruments against the autocovariances of the hedge instruments under the calibration-adjusted measure P_H .

It is worth mentioning that for $\Psi^{(S)}(p) = p \log p$, we obtain $s^{P_H^{(S)}} = 1$, $p_k^{P_H^{(S)}} = p_k^{P^{(S)}}$, $P_H^{(S)} = P^{(S)}$, and

$$\nabla_c \langle v \rangle_N^{P^{(S)}} = \langle \mathbf{g}, \mathbf{g} \rangle_N^{P^{(S)}} \cdot \langle \mathbf{g}, v \rangle_N^{P^{(S)}} \quad (44)$$

In other words, the calibration-adjusted measure is the same as the pricing measure. This is a special property of the Shannon entropy pricing measure $P^{(S)}$.

As a final note on hedge ratio calculations with WMC, it should be noted that unlike most of the other sensitivity calculation schemes used with Monte Carlo methods, the above shown analysis results directly in *hedge ratios*, bypassing the otherwise common intermediate stage of *model parameter sensitivities*, which require remapping to hedge ratios for tradable instruments. This feature greatly reduces the noise often observed on risk figures that are computed by numerically fitting model parameters to market observable prices since the noise-compounding effects of recalibration and numerical calculation of sensitivities of hedge instrument prices to model parameters are avoided.

References

- [1] Avellaneda, M. (2002). *Empirical Aspects of Dispersion Trading in the US Equities Markets*. Powerpoint presentation, Courant Institute of Mathematical Sciences, New York University, November 2002. www.math.nyu.edu/faculty/avellane/ParisFirstTalkSlides.pdf
- [2] Avellaneda, M. (2002). *Weighted-Monte Carlo Methods for Equity Derivatives: Theory and Practice*. Powerpoint presentation, Courant Institute of Mathematical Sciences, New York University, November 2002. www.math.nyu.edu/faculty/avellane/ParisTalk2.pdf
- [3] Avellaneda, M., Buff, R., Friedman, C., Grandchamp, N., Kruk, L. & Newman, J. (2001). Weighted Monte Carlo: a new approach for calibrating asset-pricing models, *International Journal of Theoretical and Applied Finance*, 4(1), 91–119.
- [4] Avellaneda, M. & Gamba, R. (2000). Conquering the Greeks in Monte Carlo: efficient calculation of the market

- sensitivities and Hedge-Ratios of financial assets by direct numerical simulation, *Mathematical Finance–Bachelier Congress 2000*, Monte Carlo, pp. 93–109, 2000/2002. www.math.nyu.edu/faculty/avellane/ConqueringTheGreeks.pdf
- [5] Cont, R. & Minca, A. (2008). *Recovering Portfolio Default Intensities Implied by CDO Quotes*, Financial engineering report no. 2008-01, Columbia University Center for Financial Engineering, ssrn.com/abstract=1104855.
- [6] Delbaen, F., Grandits, P., Rheinlander, T., Samperi, D., Schweizer, M. & Stricker, C. (2002). Exponential hedging and entropic penalties, *Mathematical Finance* **12**, 99–123, ssrn.com/abstract=312802.
- [7] Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*, Springer.
- [8] Golan, A., Judge, G. & Miller, D. (1996). *Maximum Entropy Econometrics*, John Wiley & Sons.
- [9] Jäckel, P. (2002). *Monte Carlo Methods in Finance*, John Wiley & Sons.

MARCO AVELLANEDA & PETER JÄCKEL

Sensitivity Computations: Integration by Parts

Traditional Monte Carlo Sensitivity Estimators and Their difficulties

The calculation of price sensitivities is a central modeling and computational problem for derivative securities. The prices of derivative securities are observable in the market; however, price sensitivities, the important inputs in the hedging of derivative securities, are not. Models and computational tools are thus required to establish such information which the market does not provide directly.

Mathematically, price sensitivities or greeks are partial derivatives of financial derivative prices with respect to some specific parameters of the underlying market variables. For instance, “delta” means the sensitivity to changes in the price of the underlying asset. More formally, suppose that the underlying model dynamic under the risk neutral probability is given by a stochastic differential equation (SDE) on $[0, T]$,

$$dX_t = \mu(X_t) dt + \sigma(X_t) dW_t, X_0 = x \quad (1)$$

where W is a standard Brownian motion. By the no-arbitrage argument, the present value of a derivative should be

$$V(x) = E[\Phi(X_T)|X_0 = x] \quad (2)$$

where Φ is the (discounted) payoff function. For notation at simplicity, we restrict attention to scalar X and such Φ that depends only on X_T in the article. Then, the delta of such a model is defined as $dV(x)/dx$.

The simplest and crudest approach to the Monte Carlo estimation of greeks is *via* finite-difference approximation. In other words, we simulate the derivative prices at two or more values of the underlying parameter and then estimate greeks by taking difference quotients between these values. Finite-difference estimators are easy to implement, but are prone to large bias, large variance, and added computational requirements.

To overcome the shortages of the finite-difference method, traditionally there have been two categories

of methods for estimating sensitivities: methods that differentiate paths and methods that differentiate densities. The former one is known as the *pathwise derivative method* or the infinitesimal perturbation analysis in the literature and the latter is usually referred to as the *likelihood ratio method* (see **Monte Carlo Greeks**). Both of them yield unbiased estimators. But the former requires smooth conditions on the payoff function Φ . It fails to provide any sensible estimators for options with discontinuous payoff functions such as digital options. The estimator produced by the latter involves the transition density function of X_T , which is unavailable in most circumstances when the dynamics (1) is not trivial.

Method of Integration by Parts

Fournié *et al.* [7, 8] developed an approach to bypass both the difficulties the traditional methods encounter. It is based on the *integration-by-parts* formula, which lies at the heart of the theory of the Malliavin calculus. Here, we state several relevant conclusions only and leave readers with interest to find more on the detailed and rigorous treatment of the Malliavin calculus and the related financial applications in Nualart [11] and Malliavin and Thalmaier [10]. For notational simplicity, we use the scalar case only in the article to demonstrate the basic idea of the method and refer readers to the relevant literature for more general and rigorous treatments.

Let $\{W_t : 0 \leq t \leq T\}$ be a standard Brownian motion defined on a probability space $(\Omega, \mathcal{F}, P)$ and let $\{\mathcal{F}_t : 0 \leq t \leq T\}$ be the filtration generated by W . Consider a random variable F of the form

$$F = f \left(\int_0^T h_u dW_u \right) \quad (3)$$

where f is a real function with some proper smoothness and $\{h_u : 0 \leq t \leq T\}$ is an $L^2[0, T]$ -valued stochastic process on $(\Omega, \mathcal{F}, P)$. The *Malliavin derivative* of F is defined as a stochastic process $DF = \{D_t F : 0 \leq t \leq T\}$, where

$$D_t F = f' \left(\int_0^T h_u dW_u \right) \cdot h_t \quad (4)$$

Notice that $\int_0^T h_u dW_u$ is defined as the limiting summation $\sum_{u < T} h_u \cdot dW_u := \sum_{u < T} h_u \cdot (W_{u+du} - W_u)$.

2 Sensitivity Computations: Integration by Parts

So, one can view the Malliavin derivative as an ordinary derivative of the random variable F with respect to dW_t , the small increment of the Brownian motion over $[t, t + dt]$, heuristically.

The Malliavin derivative satisfies the chain rule as the ordinary derivative, that is, for any differentiable real function ϕ , $D\phi(F) = \phi'(F) \cdot DF$. Apply D_t on X_T defined by the SDE (1). Recall that

$$X_T = X_t + \int_t^T \mu(X_u) du + \int_t^T \sigma(X_u) dW_u \quad (5)$$

X_t depends only on the Brownian increments before t . Thus, $D_t X_t = 0$. By the chain rule, $D_t \left(\int_t^T \mu(X_u) du \right) = \int_t^T \mu'(X_u) \cdot D_t X_u du$. And

$$\begin{aligned} D_t \left(\int_t^T \sigma(X_u) dW_u \right) &= D_t(\sigma(X_t) dW_t) + \int_{t+dt}^T \sigma(X_u) dW_u \\ &= \sigma(X_t) + \int_{t+dt}^T \sigma'(X_u) \cdot D_t X_u dW_u \end{aligned} \quad (6)$$

So we have

$$\begin{aligned} D_t X_T &= \sigma(X_t) + \int_t^T \mu'(X_u) \cdot D_t X_u du \\ &\quad + \int_t^T \sigma'(X_u) \cdot D_t X_u dW_u \end{aligned} \quad (7)$$

On the other hand, if one introduces a new process Y such that it is the derivative of X with respect to its initial value, that is, $Y_t = dX_t/dx$, then

$$dY_t = \mu'(X_t)Y_t dt + \sigma'(X_t)Y_t dW_t, \quad Y_0 = 1 \quad (8)$$

Comparing equations (7) and (8), we can see the process $\{D_t X_u : t \leq u \leq T\}$ should follow the same SDE as Y but with different initial value at t . Thus,

$$D_t X_T = \frac{Y_T}{Y_t} \cdot \sigma(X_t) \quad (9)$$

One of the most important properties of the Malliavin derivative is the following duality property:

given a process $h = \{h_t : 0 \leq t \leq T\}$, there exists a random variable $D^*(h)$ such that

$$E \left[\int_0^T D_t \Phi(X_T) \cdot h_t dt \right] = E[\Phi(X_T) \cdot D^*(h)] \quad (10)$$

for all functions Φ with some proper smoothness conditions. Viewing D^* as a “derivative” in the weak sense—recall that the weak derivative in the PDE theory is defined in this way—one can see that equation (10) is exactly an analog to the integration-by-parts formula in the ordinary calculus. In the literature, D^* is called the *Skorohod integral*. It is easy to show that $D^*(h)$ should be equal to the Ito integral $\int_0^T h_u dW_u$ if h is adapted to the filtration $\mathcal{F}_t$.

Equation (10) is the cornerstone for the development of unbiased greeks estimators. Turn back to the derivation of unbiased estimators for the delta. Consider a smooth payoff function Φ first. Choose $h_t \equiv \frac{1}{T}(Y_t/\sigma(X_t))$, which is adapted, in equation (10). By the chain rule of D , the left-hand side of equation (10) is

$$\begin{aligned} E \left[\int_0^T D_t \Phi(X_T) \cdot h_t dt \right] &= E \left[\int_0^T \Phi'(X_T) \cdot D_t X_T \cdot \frac{1}{T} \frac{Y_t}{\sigma(X_t)} dt \right] \\ &= E[\Phi'(X_T) \cdot Y_T] \end{aligned} \quad (11)$$

where the last step uses equation (9). The right hand side of equation (10) equals

$$E \left[\Phi(X_T) \cdot \frac{1}{T} \int_0^T \frac{Y_t}{\sigma(X_t)} dW_t \right] \quad (12)$$

So,

$$E[\Phi'(X_T) \cdot Y_T] = E \left[\Phi(X_T) \cdot \frac{1}{T} \int_0^T \frac{Y_t}{\sigma(X_t)} dW_t \right] \quad (13)$$

Using the pathwise derivative method, we can easily derive that $\Phi'(X_T) \cdot Y_T$ is an unbiased estimator of the delta. Therefore we have another unbiased estimator

$$\frac{dV}{dx} = E \left[\Phi(X_T) \cdot \frac{1}{T} \int_0^T \frac{Y_t}{\sigma(X_t)} dW_t \right] \quad (14)$$

For any nonsmooth Φ such that $E[(\Phi(X_T))^2] < +\infty$, we can always find a sequence of differentiable $\Phi^{(n)}$ convergent to it in L^2 , that is, $E[\|\Phi^{(n)}(X_T) - \Phi(X_T)\|^2 | X_0 = x] \rightarrow 0$ as $n \rightarrow +\infty$ for all x . Let $V^{(n)}(x) = E[\Phi^{(n)}(X_T) | X_0 = x]$. Following the above arguments,

$$\frac{dV^{(n)}}{dx} = E \left[\Phi^{(n)}(X_T) \cdot \frac{1}{T} \int_0^T \frac{Y_t}{\sigma(X_t)} dW_t \right] \quad (15)$$

Using the Cauchy–Schwartz inequality, we can easily show that the right-hand side of equation (15) converges to $E \left[\Phi(X_T) \cdot \frac{1}{T} \int_0^T \frac{Y_t}{\sigma(X_t)} dW_t \right]$. Thus, equation (14) should hold for such Φ too. The advantage of such an estimator is that it does not involve the density functions of X_T nor does it require Φ to be smooth.

Implementation and Some Extensions

To implement equation (14), we discretize $[0, T]$ into grids: $0 = t_0 < t_1 < \dots < t_N = T$, where $t_i = iT/N$. Simulate the underlying process X and the derivative process Y simultaneously by

$$\begin{aligned} \hat{X}_{t_i} &= \hat{X}_{t_{i-1}} + \mu(\hat{X}_{t_{i-1}})\Delta t + \sigma(\hat{X}_{t_{i-1}})\Delta W_i, \\ \hat{X}_0 &= x \end{aligned} \quad (16)$$

$$\begin{aligned} \hat{Y}_{t_i} &= \hat{Y}_{t_{i-1}} + \mu'(\hat{X}_{t_{i-1}})\hat{Y}_{t_{i-1}}\Delta t + \sigma'(\hat{X}_{t_{i-1}})\hat{Y}_{t_{i-1}}\Delta W_i, \\ \hat{Y}_0 &= 1 \end{aligned} \quad (17)$$

where $\Delta t = T/N$ and $\Delta W_i = W_{t_i} - W_{t_{i-1}} \sim N(0, \Delta t)$. Then, approximately we have a delta estimator

$$\Phi(\hat{X}_T) \cdot \frac{1}{T} \sum_{i=1}^N \frac{\hat{Y}_{t_{i-1}}}{\sigma(\hat{X}_{t_{i-1}})} \Delta W_i \quad (18)$$

All the above derivations can be generalized to the cases in which X and W are both vectors, which has been shown by Fournié *et al.* [7]. They also established unbiased estimators for other greeks for European style options contingent on multidimensional underlying assets, such as the vega, the price sensitivity with respect to the underlying volatility; the rho, the sensitivity with respect to the riskless interest rate; the gamma, the second-order sensitivity with respect

to the underlying price. One crucial assumption they needed is that the underlying model is elliptic, that is, $\sigma(x)$ is a symmetric positive definite matrix function.

It is worth mentioning that the integration-by-parts method is still applicable for the market where the ellipticity assumption does not hold. For instance, the interest rate market has a high-dimensional state space constituted by the values of bonds at a large number of distinct maturities and a low-dimensionality variance driven by a few noise sources (Brownian motions). Under this setting, $\sigma(x)$ cannot be positive definite because it has more rows than columns. Malliavin and Thalmaier [10] provide the details on how to develop the corresponding unbiased estimators.

A lot of research has already been done so far in the literature to extend the seminal work of Fournié *et al.* [7]. Among others, Davis and Johansson [5], El-Khatib and Privault [6], and Bally *et al.* [1] considered greeks in a market driven by jump-diffusion processes; Gobet and Kohatsu-Higa [9], Bernis *et al.* [3] derived greek estimators for lookback and barrier options; Bally *et al.* [2] applied the Malliavin calculus method to pricing and hedging American options.

As shown here, Malliavin estimators have been derived directly for diffusion processes, but implementation typically requires simulation of a discrete-time approximation. This raises the question of whether one should discretize first and then differentiate, or differentiate first and then discretize. Chen and Glasserman [4] illustrated that both approaches will lead to the same estimators in several important cases, but the first approach uses only elementary techniques such as likelihood ratio and pathwise derivative methods.

Acknowledgments

The author thanks Arturo Kohatsu-Higa, Eckhard Platen, Peter Jaeckel, and David Yao for their suggestions and comments in the preparation of this article.

References

- [1] Bally, V., Bavouzet, M.-P. & Messaoud, M. (2007). Integration by parts formula for locally smooth laws and applications to sensitivity computations, *Annals of Applied Probability* **17**, 33–66.

- [2] Bally, V., Caramellino, L. & Zanette, A. (2005). Pricing and hedging American options by Monte Carlo methods using a Malliavin calculus approach, *Monte Carlo Methods and Applications* **11**, 97–133.
- [3] Bernis, G., Gobet, E. & Kohatsu-Higa, A. (2003). Monte Carlo evaluation of greeks for multidimensional barrier and lookback options, *Mathematical Finance* **13**, 99–113.
- [4] Chen, N. & Glasserman, P. (2007). Malliavin greeks without Malliavin calculus, *Stochastic Processes and their Applications* **117**, 1689–1723.
- [5] Davis, M.H.A. & Johansson, M.P. (2006). Malliavin Monte Carlo greeks for jump diffusions, *Stochastic Processes and their Applications* **116**, 101–129.
- [6] El-Khatib, Y. & Privault, N. (2004). Computations of greeks in a market with jumps via the Malliavin calculus, *Finance and Stochastics* **8**, 161–179.
- [7] Fournié, E., Lasry, J.-M., Lebuchoux, J., Lions, P.-L. & Touzi, N. (1999). Applications of Malliavin calculus to Monte Carlo methods in finance, *Finance and Stochastics* **3**, 391–412.
- [8] Fournié, E., Lasry, J.-M., Lebuchoux, J., Lions, P.-L. & Touzi, N. (2001). Applications of Malliavin calculus to Monte Carlo methods in finance. II, *Finance and Stochastics* **5**, 201–236.
- [9] Gobet, E. & Kohatsu-Higa, A. (2003). Computation of greeks for barrier and look-back options using Malliavin Calculus, *Electronic Communications in Probability* **8**, 51–62.
- [10] Malliavin, P. & Thalmaier, A. (2006). *Stochastic calculus of Variations in Mathematical Finance*, Springer.
- [11] Nualart, D. (2006). *The Malliavin Calculus and Related Topics*, 2nd Edition, Springer.

Related Articles

Monte Carlo Greeks; Stochastic Differential Equations; Scenario Simulation.

NAN CHEN

Stochastic Mesh Method

The stochastic mesh method has been proposed by Broadie and Glasserman [2] to price multivariate Bermudan options. These contracts include an early exercise feature, which makes their pricing a challenging computational problem. There are several methods available for options with payoff functions depending only on one asset. They include finite difference methods and approaches based on lattices (see **Finite Difference Methods for Early Exercise Options**). These methods, however, become typically ineffective for high-dimensional problems, like basket options.

The stochastic mesh approach belongs to simulation-based methods, whose attractiveness for complex computational problems stems from the fact that the convergence rate is typically independent of the dimension of the problem.

The method is based on the dynamic programming representation and in each period uses a finite collection of points from the state space. In these aspects, the method is similar to the binomial tree approach. Its distinctive feature is that at each time interval the nodes are selected randomly and their number is always the same.

To illustrate the method, we consider a Bermudan option whose payoff depends only on the current price of an underlying asset. We assume that changes of the asset's price can be described by a Markov chain: $\{S_{t_i}\}_{i=0,\dots,M}$, with values in $\mathcal{R}^b$. Since Bermudan options are often used to approximate prices of American options, $\{S_{t_i}\}_{i=0,\dots,M}$ may be obtained as a result of sampling a continuous process $\{S_t\}_{0 \leq t \leq T}$. The option can be exercised at $M + 1$ time points (including the initial time), which we shall denote as $t_0, t_1, \dots, t_M = T$. At the time of exercise τ , the options are equal to $G(\tau, S_\tau)$, where G is a given function.

An arbitrage-free price of the option can be obtained by using the risk-neutral representation:

$$P(t_0, S_0) := \max_{\tau} E[B(t_0, \tau)G(\tau, S_\tau)] \quad (1)$$

where τ is a stopping time that takes values in the set $\{t_0, t_1, \dots, T\}$. Here, the expectation is taken with respect to a given risk-neutral measure, Q , and $B(s, t)$ denotes the discount factor for the period (s, t) . The main difficulty in using equation (1) is

due to the fact that the optimal exercise strategy is unknown. The price of the option and the optimal strategy can be obtained, however, from the dynamic programming characterization of the problem. For this, we use the following backward recursion to calculate, for $i = M, \dots, 0$, functions $P(t_i, \cdot)$:

$$P(T, x) = G(T, x) \quad (2)$$

$$P(t_i, x) = \max\{G(t_i, x), C(t_i, x)\},$$

$$i = M - 1, \dots, 0 \quad (3)$$

where the continuation value, $C(t_i, x)$, is defined as

$$C(t_i, x) := B_{t_i} E[P(t_{i+1}, S_{t_{i+1}}) | S_{t_i} = x] \quad (4)$$

Then the price of the instrument at t_0 is given by $P(t_0, S_0)$. For simplicity of exposition, we assume that the discount factor $B_{t_i} \equiv B(t_i, t_{i+1})$ is deterministic, but in a more general formulation of the method it can be stochastic.

To use this method in practice, we must be able to efficiently calculate the conditional expectation

$$E[P(t_{i+1}, S_{t_{i+1}}) | S_{t_i} = x] \quad (5)$$

for $i = 0, \dots, M - 1$ and the selected set of points x from the state space. To accomplish this, several techniques have been proposed, including binomial trees and Monte Carlo simulations.

In the stochastic mesh method, the expectation (5) are approximated by arithmetic sums. For this, in the first phase, we generate at each exercise time t_i , $i = 0, \dots, M$, the same number of random points, $\{x_{ij}, j = 1, \dots, d\}$. Then, at each node of the mesh the option price is calculated using the backward recursion. At the terminal nodes, we take $\hat{P}(T, x_{Mj}) = G(T, x_{Mj})$, $j = 1, \dots, d$. At all other exercise times, we estimate the price of the option using the following mesh estimator

$$\hat{P}(t_i, x_{ij}) = \max\{G(t_i, x_{ij}), \hat{C}(t_i, x_{ij})\} \quad (6)$$

with the estimate $\hat{C}$ of the continuation value defined by

$$\hat{C}(t_i, x_{ij}) := B_{t_i} \frac{1}{d} \sum_{k=1}^d \hat{P}(t_{i+1}, x_{(i+1)k}) w_i(j, k) \quad (7)$$

where $w_i(j, k) \equiv w_i(x_{ij}, x_{(i+1)k})$ is a weight attached to the pair $(x_{ij}, x_{(i+1)k})$. Thus, for each fixed node at

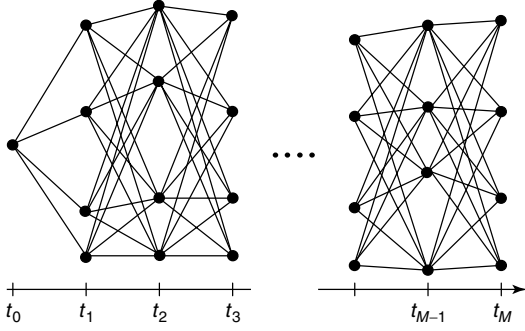


Figure 1 An example of a stochastic mesh with $d = 4$. Weight that is attached to each arc is used in a weighted-sum approximation to the corresponding continuation value

time t_i , all nodes at the next time are used to calculate the corresponding continuation value. This has been depicted in Figure 1 for $d = 4$. Selection of these weights depends on the mechanism we have used to generate the mesh points $\{x_{ij}\}$. The weighted sum (7) can be interpreted as an estimate of the conditional expectation (5) when the process is in state x_{ij} at time t_i . Since in calculations we always use all mesh points at the next time, regardless of the current state of the process, the main purpose of these weights is to estimate the continuation values without bias. We provide more details about different mechanisms of selecting mesh points and weights in the following section.

Selection of the Mesh Points and Weights

Proper choices of the mesh points and weights $w_i(j, k)$ in equation (7) are essential for the method to be successful. The approaches suggested in the literature are based on the assumption that the

Markov process $\{S_{t_i}\}_{i=0,\dots,M}$ admits transition densities and they are known or can be evaluated numerically. We therefore assume that conditional on $S_{t_i} = x$ the value of the process at the next time instance, $S_{t_{i+1}}$, has density $f(t_i, x, \cdot)$. Denote by $h(t_i, \cdot)$ the density of S_{t_i} conditional on the initial value S_0 . We refer to $h(t_i, \cdot)$ as the marginal density of S_{t_i} .

Broadie and Glasserman [2] have considered two constructions of mesh points, which we present in Figure 2. In both, the points $\{x_{ij}, j = 1, \dots, b\}$ at time t_i are generated as independent and identically distributed observations from some density $g(t_i, \cdot)$, called the *mesh density function*. In the first method, the mesh density is allowed to depend on time but not on the mesh points generated at previous time intervals. In the second method, the mesh is constructed in a forward procedure where at time t_i the points $\{x_{ij}, j = 1, \dots, b\}$ are generated as samples from a mesh density that depends on observations at the previous time t_{i-1} . Although other methods of generating mesh points have been proposed later (e.g., [4]), we focus only on these two.

A natural choice for the mesh density $g(t_i, \cdot)$ is to take it equal to the marginal density function $h(t_i, \cdot)$. In this case, the mesh points at time t_i are generated independently of points at previous times $t_1, \dots, t_{i-1}$. Using a simple change of measure, we can transform conditional expectation (5) to an expectation with respect to the density $h(t_{i+1}, \cdot)$:

$$\begin{aligned} E[P(t_{i+1}, S_{t_{i+1}}) | S_{t_i} = x_{ij}] &= \int P(t_{i+1}, z) f(t_i, x_{ij}, z) dz \\ &= \int P(t_{i+1}, z) \frac{f(t_i, x_{ij}, z)}{h(t_{i+1}, z)} h(t_{i+1}, z) dz \quad (8) \end{aligned}$$

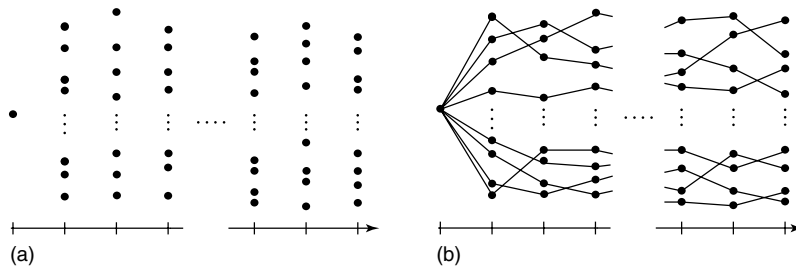


Figure 2 Two methods of generating mesh points. (a) The points are generated independently at each time interval. (b) The points are generated by independent paths

Since this result is valid for any integrable function P , it implies that weighted sums with weights equal to the likelihood ratio

$$w_i(j, k) := \frac{f(t_i, x_{ij}, x_{(i+1)k})}{g(t_{i+1}, x_{(i+1)k})} \quad (9)$$

provide unbiased estimates of conditional expectations with respect to the transition density $f(t_i, x_{ij}, \cdot)$, meaning that

$$E \left[\frac{B_{t_i}}{d} \sum_{k=1}^d P(t_{i+1}, X_k) w_i(x_{ij}, X_k) \right] = C(t_i, x_{ij}) \quad (10)$$

where $X_1, \dots, X_d$ are independent and identically distributed random variables with the common density equal to $g(t_{i+1}, \cdot)$.

A drawback of selecting the mesh density equal to the marginal density function is that it can lead to estimators whose variance grows exponentially with the number of exercise times. In the context of an European option, this fact has been demonstrated by Broadie and Glasserman [2]. The authors have also observed that this build up of variance can be avoided if we choose the mesh density equal to the average of transition density functions

$$g^A(t_{i+1}, x_i, x) = \frac{1}{d} \sum_{j=1}^d f(t_i, x_{ij}, x), \quad i = 1, \dots, M-1 \quad (11)$$

and $g^A(t_1, x) = f(t_0, S_0, x)$, where x_i denotes the vector of mesh points at time t_i . For this choice of the density, points at time t_{i+1} are generated from a distribution that depends on the position of the mesh points at the previous time t_i . The particular form of the mesh density (11) as a mixture of transition densities allows for two equivalent interpretations of the generation mechanism. In the first, we use a forward procedure where at each time we independently generate all points from the same distribution. To generate a point at time t_i from the density $g(t_i, x_{(i-1)\cdot}, \cdot)$, we first randomly choose a point from the set $\{x_{(i-1)1}, \dots, x_{(i-1)d}\}$, say $x_{(i-1)j^*}$, and then sample from the transition density $f(t_{i-1}, x_{(i-1)j^*}, \cdot)$. This procedure is repeated until all d points have been obtained, after which we proceed to the next time interval. Alternatively, generating a

mesh using equation (11) is equivalent to generating d independent paths of the process $\{S_0, S_{t_1}, \dots, S_{t_M}\}$, as presented in Figure 2(b), and then disconnecting the nodes on each path.

For this construction of a mesh, we can use the likelihood ratio weights (9) with the mesh density $g(t_i, \cdot)$ replaced by $g^A(t_i, x_i, \cdot)$. These weights preserve the property of the weighted sum being an unbiased estimator of conditional expectation. In this case, however, expectation (10) is conditional on the position of the mesh points at the previous time t_{i-1} .

Properties of the Method and Its Extensions

The stochastic mesh method has an important property of being asymptotically consistent, meaning that by increasing the number of mesh points it can be assured that the mesh estimates will converge to the true price of the option. When the distributions used to sample mesh points are set to marginal distributions of the process, this result has been established by Broadie and Glasserman [2]. Avramidis and Matzinger [1] have proved the asymptotic consistency under weaker assumptions on the generating mechanism, which, in particular, cover the construction based on independent paths of the process $\{S_{t_i}\}_{i=0, \dots, M}$.

Broadie and Glasserman [2] and Glasserman [4] have also shown that under some conditions on the mechanism of selecting the mesh points and weights in equation (7), the mesh estimator defined by equations (6) and (7) is biased high. This means that for a fixed number of mesh points and a fixed number of exercise opportunities, if we randomly generate sufficiently many meshes and calculate the corresponding mesh estimates, then the arithmetic average of all these estimates will be strictly greater than the true price of the option. The proof of this result uses Jensen's inequality and relies on a particular choice of weights that guarantees the unbiasedness property of the weighted sums (10). In particular, the likelihood ratios and the two constructions of the mesh points described in the previous section lead to mesh estimators that are bias high.

The bias high property of the method can be used to construct conservative confidence intervals for the option price. For this, we can combine the mesh estimator with any bias low estimator. The latter can be constructed quite easily, since an application of

any exercise strategy in equation (1) that is different from the optimal one will lead to a bias low estimate of the price. Such a suboptimal stopping rule can also be constructed within the stochastic mesh method [2]. In the process of finding the mesh estimate through the dynamic programming procedure (6) and (7), each node of the mesh can be classified as either belonging to the exercise region or not, depending whether the maximum in equation (6) is equal to the current value of the option or the continuation value. This classification can be extended to an arbitrary state of the process. Because the immediate exercise value $G(t_i, x)$ in equation (6) is defined for all x , we need to estimate a continuation value $\hat{C}(t_i, x)$ at an arbitrary state x . To do this, it suffices to extend the weights $w_i(j, k)$ to all points in the state space, which for the likelihood ratio weights (9) can be easily accomplished by simply replacing x_{ij} with x .

A bias low estimator, called *path estimator*, is based on the exercise region corresponding to these continuation values. Its value is obtained by randomly generating a certain number of trajectories of the underlying process and then stopping each trajectory if it reaches this exercise region. Thus, for the j th simulated path, S^j , the exercise time $\hat{\tau}$ is defined as

$$\hat{\tau}(S^j) = \min\{t_i : G(t_i, S^j_{t_i}) \geq \hat{C}(t_i, S^j_{t_i})\} \quad (12)$$

in the case when the trajectory reaches the exercise region, and $\hat{\tau} = T$, otherwise. Then, conditional on this mesh, the path estimate based on N -simulated trajectories is simply the average of the discounted payoffs at the exercise time

$$\frac{1}{N} \sum_{j=1}^N B(t_0, \hat{\tau}(S^j)) G(\hat{\tau}(S^j), S^j_{\hat{\tau}(S^j)}) \quad (13)$$

If we repeat this procedure for a number of independently generated meshes, then the sample mean of the estimates (13) obtained for each copy of the mesh will give a bias low estimate of the price of the option. Under some conditions, this estimator is asymptotically unbiased as the number of mesh points d increases to infinity. Therefore, the confidence interval that combines the bias high and the bias low estimators will shrink to the true price of the option as $d \rightarrow \infty$ and the number of generated trajectories N increases to infinity.

Several numerical examples illustrating applications of the stochastic mesh method have been presented by Broadie and Glasserman [2]. Numerical

tests conducted by the authors suggest that the efficiency of the method can be significantly improved when it is combined with some standard variance reduction techniques, in particular, those based on control variates (*see Variance Reduction*). Another way of enhancing the method is proposed in Boyle *et al.*, where mesh points are generated using low-discrepancy points (*see Quasi-Monte Carlo Methods*). A serious limitation of the method is the need for a transition density. To address this issue, in [3] the authors extend the stochastic mesh method by proposing to choose mesh weights through a constrained optimization problem.

References

- [1] Avramidis, A.N. & Matzinger, H. (2004). Convergence of the stochastic mesh estimator for pricing Bermudan options, *The Journal of Computational Finance* **7**, 73–91.
- [2] Broadie, M. & Glasserman, P. (2004). A stochastic mesh method for pricing high-dimensional American options, *The Journal of Computational Finance* **7**, 35–72.
- [3] Broadie, M., Glasserman, P. & Ha, Z. (2000). Pricing American options by simulation using a stochastic mesh with optimized weights, in *Probabilistic Constrained Optimization: Methodology and Applications*, S. Uryasev, ed, Kluwer, Norwell, pp. 32–50.
- [4] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*, Springer, New York.

Further Reading

- Boyle, P.P., Kolkiewicz, A. & Tan, K.S. (2001). Valuation of the reset options embedded in some equity-linked insurance products, *North American Actuarial Journal* **5**, 1–18.

Related Articles

American Options; Bermudan Options; Bermudan Swaptions and Callable Libor Exotics; Exercise Boundary Optimization Methods; Early Exercise Options: Upper Bounds; Finite Difference Methods for Early Exercise Options; Integral Equation Methods for Free Boundaries; Monte Carlo Simulation for Stochastic Differential Equations; Sparse Grids; Tree Methods; Weighted Monte Carlo.

ADAM KOLKIEWICZ

LIBOR Market Models: Simulation

The Libor market (LM) model (*see LIBOR Market Model*) involves a large number of Markov state variables—namely, the full number of Libor forward rates on the yield curve plus any additional variables such as stochastic volatility—so finite difference methods are rarely applicable, and securities pricing will nearly always require Monte Carlo methods. The specific form of the stochastic differential equations (SDEs) governing the dynamics of forward Libor rates gives rise to some specialized approaches, as well as to particular challenges, that are covered in this article. The notation used throughout is that of **LIBOR Market Model**.

Euler-type Schemes

Basic Stepping Method

Assume that we stand at time t , and Libor forwards (L) maturing at all dates in the tenor structure are known. We wish to devise a scheme to advance time to $t + \Delta$ and construct a sample^a of $L_{q(t+\Delta)}(t + \Delta), \dots, L_{N-1}(t + \Delta)$. Notice that $q(t + \Delta)$ may or may not exceed $q(t)$; if it does, some of the front-end forward rates expire and “drop off” the curve in the move to $t + \Delta$.

Under the spot measure $\mathbf{P}^B$, general LM model dynamics are of the form (*see LIBOR Market Model*)

$$dL_n(t) = \sigma_n(t)^\top (\mu_n(t) dt + dW_B(t)),$$

$$\mu_n(t) = \sum_{j=q(t)}^n \frac{\tau_j \sigma_j(t)}{1 + \tau_j L_j(t)} \quad (1)$$

where $\sigma_n(t)$ are adapted vector-valued volatility functions and W_B is an m -dimensional Brownian motion in measure $\mathbf{P}^B$. The simplest way of drawing an approximate sample $\hat{L}_n(t + \Delta)$ for $L_n(t + \Delta)$ would be to apply a first-order Euler-type scheme,^b

$$\begin{aligned} \text{Euler : } \hat{L}_n(t + \Delta) \\ = \hat{L}_n(t) + \sigma_n(t)^\top (\mu_n(t) \Delta + \sqrt{\Delta} z) \end{aligned} \quad (2)$$

$$\text{log-Euler : } \hat{L}_n(t + \Delta)$$

$$= \hat{L}_n(t) \exp \left\{ \frac{\sigma_n(t)^\top}{\hat{L}_n(t)} \left(\mu_n(t) \Delta - \frac{1}{2} \frac{\sigma_n(t)}{\hat{L}_n(t)} \Delta + \sqrt{\Delta} z \right) \right\} \quad (3)$$

where z is a vector of m independent $\mathcal{N}(0, 1)$ Gaussian draws. For specifications of $\sigma_n(t)^\top$ that are close to proportional in $L_n(t)$ (e.g., the lognormal LM model), the log-Euler scheme (3) can be expected to produce lower biases than the Euler scheme (2). The log-Euler scheme will keep forward rates positive, whereas the Euler scheme will not.

As shown, both schemes (2) and (3) advance time only by a single time step, but creation of a full path of forward curve evolution through time is merely a matter of repeated application of the single-period stepping schemes on a (possibly nonequidistant) time line $t_0, t_1, \dots$.

While Euler-type schemes such as (2) and (3) are not very sophisticated and result in rather slow convergence of the discretization bias ($O(\Delta)$), these schemes are appealing in their straightforwardness and universal applicability.

Iterative Drift Calculations

In the straight Euler scheme (2), the computational effort involved in advancing L_n is dominated by the computation of $\mu_n(t)$ which, in a direct implementation of equation (1), involves $m \cdot (n - q(t) + 1) = O(mn)$ operations for a given value of n . To advance all $N - q(t + \Delta)$ forwards, it follows that the computational effort is $O(mN^2)$ for a single time step. Generation of a full path of forward curve scenarios from time 0 to time T_{N-1} will thus require a total computational effort of $O(mN^3)$.

A simple observation allows one to substantially reduce the computational effort. Invoking the recursive relationship

$$\mu_n(t) = \mu_{n-1}(t) + \frac{\tau_n \sigma_n(t)}{1 + \tau_n \hat{L}_n(t)} \quad (4)$$

allows one to compute all $\mu_n, n = q(t + \Delta), \dots, N - 1$, by an $O(N)$ -step iteration starting from $\mu_{q(t+\Delta)}(t) = \sum_{j=q(t)}^{q(t+\Delta)} \tau_j \sigma_j(t) / (1 + \tau_j \hat{L}_j(t))$. In total, the computational effort of advancing the full curve one-

time step will be $O(mN)$, and the cost of taking N such time steps will be $O(mN^2)$ —and not $O(mN^3)$.

This reduction in effort is discussed, in considerable level of detail, in [6]. It can be verified to hold for any of the probability measures used in LM modeling; the starting point of recursions would depend on which measure is used.

Long Time Steps

Most exotic interest rate derivatives involve revolving cash flows paid on a tightly spaced schedule (e.g., quarterly). Hence, the average time spacing used in path generation will thus normally, by necessity, be quite small. In certain cases, however, there may be large gaps between cash flow dates, for example, when a security is forward-starting or has an initial lockout period. One may want, then, to be able to take a “large” time step in the simulation; here “large” is defined as spanning more than one period in the tenor structure.

When taking large time steps in simulation, not all probability measures are equally useful. In particular, the usage of the spot measure $\mathbf{P}^B$ is inconvenient, as skipping over points in the tenor structure will preclude one from “rolling” $B(t)$ correctly. Circumventing this issue, however, is merely a matter of changing the numeraire from B to an asset that involves no rollover in the time step in question. For instance, we could use the discount bond $P(t, T_N)$, the choice of which corresponds to running simulated paths in the terminal measure. Model dynamics under this measure are given in **LIBOR Market Model**.

Other Simulation Schemes

When simulating on a reasonably tight time schedule, the performance of the Euler scheme is often sufficiently accurate to serve as the default scheme for many types of LM models. However, as discussed above, in some cases, it could be beneficial to use coarse time steps in some parts of the path-generation algorithm, requiring one to pay more attention to the discretization scheme.

Special-purpose Schemes with Drift Predictor–Corrector

In integrated form, the general LM dynamics in equation (1) become

$$\begin{aligned} L_n(t + \Delta) &= L_n(t) + \int_t^{t+\Delta} \sigma_n(u)^\top \mu_n(u) du \\ &\quad + \int_t^{t+\Delta} \sigma_n(u)^\top dW_B(u) \\ &\equiv L_n(t) + D_n(t + \Delta) + M_n(t + \Delta) \end{aligned} \quad (5)$$

where M_n and D_n is a martingale and a predictable process, respectively, on the interval $[t, t + \Delta]$. In many cases of practical interest, high-performance special-purpose schemes exist for simulation of the martingale process M_n . A simple approach is to use Euler stepping for the predictable part,

$$\hat{L}_n(t + \Delta) = \hat{L}_n(t) + \sigma_n(t)^\top \mu_n(t) \Delta + \hat{M}_n(t + \Delta) \quad (6)$$

where $\hat{M}_n(t + \Delta)$ is generated by a special-purpose scheme.

The drift adjustments in equation (6) are explicit in nature, as they are based only on the forward curve at time t . To incorporate information from time $t + \Delta$, one can use the *predictor–corrector* scheme, see [8] and **Monte Carlo Simulation for Stochastic Differential Equations**, which for equation (6) takes the two-step form

$$\begin{aligned} \bar{L}_n(t + \Delta) &= \hat{L}_n(t) + \sigma_n(t; \{\hat{L}_i(t)\})^\top \mu_n(t; \{\hat{L}_i(t)\}) \Delta \\ &\quad + \hat{M}_n(t + \Delta) \\ \hat{L}_n(t + \Delta) &= \hat{L}_n(t) + \theta \sigma_n(t; \{\hat{L}_i(t)\})^\top \mu_n(t; \{\hat{L}_i(t)\}) \Delta \\ &\quad + (1 - \theta) \sigma_n(t + \Delta; \{\bar{L}_i(t + \Delta)\})^\top \\ &\quad \times \mu_n(t + \Delta; \{\bar{L}_i(t + \Delta)\}) \Delta + \hat{M}_n(t + \Delta) \end{aligned} \quad (7)$$

where θ is a parameter in $[0, 1]$ that determines the degree of implicitness in the scheme ($\theta = 1$: fully explicit; $\theta = 0$: fully implicit). In practice, one would

nearly always make the balanced choice of $\theta = 1/2$. In equations (7) and (8) the short-hand notation

$$\mu_n \left(t; \left\{ \hat{L}_i(t) \right\} \right)$$

is used to indicate that μ_n (and σ_n) may depend on the state of the entire forward curve at time t .

Clearly, the same idea could be developed with the log-Euler scheme (3) as a starting point.

While the weak convergence order of simulation schemes may not be affected by predictor–corrector schemes, experiments show that equations (7) and (8) often will reduce the bias significantly relative to a fully explicit Euler scheme. Some results (and further refinements) can be found in [4, 9] and [10]. As the computational effort of computing the predictor–step is not insignificant, the speed–accuracy trade-off must be evaluated on a case-by-case basis.

Further Refinements of Drift Estimation

For large time steps, it may be useful to explicitly integrate the time-dependent parts of the drift, rather than rely on pure Euler-type approximations. Focusing on, say, equation (6), assume that one can write

$$\sigma_n(u)^\top \mu_n(u) \approx f(u, \{L_i(t)\}), \quad u \geq t \quad (9)$$

for a function f that depends on time as well as the state of the forward curve frozen at time t . Then,

$$\begin{aligned} D_n(t + \Delta) &= \int_t^{t+\Delta} \sigma_n(u)^\top \mu_n(u) du \\ &\approx \int_t^{t+\Delta} f(u, \{L_i(t)\}) du \end{aligned} \quad (10)$$

As f evolves deterministically for $u > t$, the integral on the right-hand side can be evaluated either analytically (if f is simple enough) or by numerical quadrature.

The approach in equation (10) easily combines with predictor–corrector logic, that is,

$$\begin{aligned} D_n(t + \Delta) &\approx \theta \int_t^{t+\Delta} f(u, \{L_i(t)\}) du \\ &\quad + (1 - \theta) \int_t^{t+\Delta} f(u, \{\bar{L}_i(t + \Delta)\}) du \end{aligned} \quad (11)$$

where $\bar{L}_i(t + \Delta)$ has been found in a predictor step using equation (10) in equation (6).

Brownian Bridge Schemes and Other Ideas

As a variation on the predictor–corrector scheme, a further refinement to take into account variance of the Libor curve between the sampling dates t and $t + \Delta$ could be made. Schemes attempting to do so by application of *Brownian bridge* techniques were proposed in [9], among others. While performance of these schemes vary—tests in [7] show rather mixed results in comparison to simpler predictor–corrector schemes—the basic idea is sufficiently simple and instructive to merit a brief mention. In a nutshell, the Brownian bridge approach aims to replace in equation (10) all forward rates $\{L_i(t)\}$ with their expected values at time u , *conditional* upon the forward rates ending up at $\{\bar{L}_i(t + \Delta)\}$, with $\{\bar{L}_i(t + \Delta)\}$ generated in a predictor step. Under simplifying assumptions on the dynamics of $L_n(t)$, a closed-form expression is possible for this expectation. For example, if we assume that

$$dL_n(t) \approx \sigma_n(t)^\top dW(t) \quad (12)$$

where σ_n is *deterministic* and $W(t)$ is an m -dimensional Brownian motion in some probability measure $\mathbf{P}$, then, for $u \in [t, t + \Delta]$,

$$\begin{aligned} E(L_n(u) | L_n(t), L_n(t + \Delta)) \\ = L_n(t) + \frac{\Sigma_1}{\Sigma_2} (L_n(t + \Delta) - L_n(t)) \end{aligned} \quad (13)$$

where

$$\Sigma_1 = \int_t^u |\sigma_n(s)|^2 ds, \quad \Sigma_2 = \int_t^{t+\Delta} |\sigma_n(s)|^2 ds \quad (14)$$

If it is more appropriate to assume that L_n is roughly lognormal, that is,

$$dL_n(t)/L_n(t) \approx \sigma_n(t)^\top dW(t) \quad (15)$$

for a deterministic volatility σ_n , then, for $u \in [t, t + \Delta]$,

$$\begin{aligned}
& E(L_n(u)|L_n(t), L_n(t + \Delta)) \\
&= L_n(t) \left(\frac{L_n(t + \Delta)}{L_n(t)} \right)^{\Sigma_1 \Sigma_2^{-1}} \\
&\quad \times \exp \left(\frac{1}{2} \Sigma_2^{-1} \Sigma_1 (\Sigma_2 - \Sigma_1) \right) \quad (16)
\end{aligned}$$

The article [7] investigates a number of other possible discretization schemes for the drift term in the LM model, including those that attempt to incorporate information about the correlation between various forward rates. In general, many of these schemes will result in some improvement of the discretization error, but at the cost of more computational complexity and effort—effort that in some instances might be better spent on just using a finer time discretization in a simpler discretization scheme.

High-order Schemes

Higher order schemes (such as the Milstein scheme and similar Taylor-based approaches, *see Monte Carlo Simulation for Stochastic Differential Equations* and *Stochastic Taylor Expansions*) are cumbersome to use for LM modeling due to the high dimensionality of the model. Consequently, there are very few empirical results on their relative performance. Andersen and Andreasen [1] list some results for Richardson extrapolation (see [8]), the effect of which seems to be rather modest.

Martingale Discretization

It is possible to select a measure such that a particular zero-coupon bond and a particular forward rate agreement (FRA) will be priced bias-free^c by Monte Carlo simulation, even when using a simple Euler scheme. While one is rarely interested in pricing zero-coupon bonds by Monte Carlo methods, this observation can, nevertheless, occasionally help guide the choice of simulation measure, particularly if, say, a security can be argued to depend primarily on a single forward rate (e.g., caplet-like securities). In practice, matters are rarely this clear-cut, and one wonders whether perhaps simulation schemes exist that will simultaneously price all zero-coupon bonds $P(\cdot, T_1)$, $P(\cdot, T_2)$, $\dots$, $P(\cdot, T_N)$ bias-free. It should be obvious that this cannot be accomplished by a simple measure

shift, but will require a more fundamental change in simulation strategy.

Deflated Bond Price Discretization

To develop a simulation scheme, which, by construction, will ensure that all numeraire-deflated bond prices are martingales, one can follow the suggestion offered in [3]: instead of discretizing the dynamics for Libor rates directly, simply discretize the deflated bond prices themselves. Consider the spot measure, and define

$$U(t, T_{n+1}) = \frac{P(t, T_{n+1})}{B(t)} \quad (17)$$

The dynamics for deflated zero-coupon bond prices (17) are given by

$$\begin{aligned}
\frac{dU(t, T_{n+1})}{U(t, T_{n+1})} &= - \sum_{j=q(t)}^n \tau_j \frac{U(t, T_{j+1})}{U(t, T_j)} \sigma_j(t)^\top dW_B(t), \\
n &= q(t), \dots, N-1 \quad (18)
\end{aligned}$$

Discretization schemes for equation (18), which preserve the martingale property, are easy to construct. For instance, the log-Euler scheme

$$\begin{aligned}
\hat{U}(t + \Delta, T_{n+1}) &= \hat{U}(t, T_{n+1}) \\
&\times \exp \left(-\frac{1}{2} |\eta_{n+1}(t)|^2 \Delta + \eta_{n+1}(t)^\top z \sqrt{\Delta} \right) \quad (19)
\end{aligned}$$

has this property where, as before, z is an m -dimensional standard Gaussian variable, and

$$\eta_{n+1}(t) \triangleq - \sum_{j=q(t)}^n \tau_j \frac{\hat{U}(t, T_{j+1})}{\hat{U}(t, T_j)} \sigma_j(t) \quad (20)$$

One can replace equation (19) with another discretization of equation (18), or one can try to discretize a quantity other than the deflated bonds $U(t, T_n)$. The latter idea is pursued in [3], where several suggestions for discretization variables are considered. For instance, one can consider the differences

$$U(t, T_n) - U(t, T_{n+1}) \quad (21)$$

which are martingales, since the U s are. Discretizing $U(t, T_n) - U(t, T_{n+1})$ is, in a sense, close to discretizing L_n itself, which may be advantageous. [7]

contains some tests of discretization schemes based on equation (21), in a lognormal model setting.

End Notes

^aWe recall from **LIBOR Market Model** that $q(t)$ is an index function, indicating which bucket of a tenor structure $\{T_n\}_{n=0}^N$ time t belongs to.

^bIn the interest of brevity and conciseness, the simulation schemes discussed in this article omit a number of real-life details, such as proper rate fixing conventions, interpolation techniques and introduction of stochasticity in the front stub, that is, in the bucket $[t, T_{q(t)}]$. Interpolation and propagation of “nonstandard” rates, that is, rates with dates not aligned with the tenor structure $\{T_n\}$, are covered in, for example, [2], and a simple representative technique for stochastic front stubs can be found in [5].

^cBut not error-free, obviously—there will still be a statistical mean-zero error on the simulation results.

References

- [1] Andersen, L.B.G. & Andreasen, J. (2000). Volatility skews and extensions of the Libor market model, *Applied Mathematical Finance* **7**, 1–32.
- [2] Brigo, D. & Mercurio, F. (2001). *Interest-Rate Models—Theory and Practice*, Springer Verlag.
- [3] Glasserman, P. & Zhao, X. (2000). Arbitrage-free discretization of lognormal forward Libor and swap rate models, *Finance and Stochastics* **4**, 35–68.
- [4] Hunter, C.J., Jäckel, P. & Joshi, M.S. (2001). Getting the drift, *Risk* **14**(7), 81–84.
- [5] Jäckel, P. (2005). *The Practicalities of Libor Market Models*. Training course notes, <http://www.jaeckel.org/ThePracticalitiesOfLiborMarketModels.pdf>.
- [6] Joshi, M.S. (2003). *Rapid Drift Computations in the Libor Market Model*, Wilmott.
- [7] Joshi, M.S. & Stacey, A.M. (2008). New and robust drift approximations for the Libor market model, *Quantitative Finance* **8**, 427–434.
- [8] Kloeden, P.E. & Platen E. (2000). *Numerical Solution of Stochastic Differential Equations, Stochastic Modelling and Applied Probability*, Springer.
- [9] Pietersz, R., Pelsser, A. & van Regenmortel, M. (2004). Fast drift approximated pricing in the BGM model, *Journal of Computational Finance* **8**, 93–124.
- [10] Rebonato, R. (2002). *Modern Pricing of Interest Rate Derivatives: The Libor Market Model and Beyond*, Princeton University Press.

Related Articles

LIBOR Market Model; Monte Carlo Simulation for Stochastic Differential Equations; Stochastic Taylor Expansions.

LEIF B.G. ANDERSEN & VLADIMIR V.
PITERBARG

Quasi-Monte Carlo Methods

Many typical problems of modern computational finance can be rephrased mathematically as problems of calculating integrals with high-dimensional integration domains (see the section Applications to Computational Finance for several examples). In fact, the dimensions may very well be in the hundreds or even in the thousands. Very often in such finance problems the integrand will be quite complicated, so that the integral cannot be evaluated analytically and precisely. In such cases, one has to resort to *numerical integration*, that is, to a numerical scheme for the approximation of integrals.

High-dimensional numerical integration is a challenging problem. Classical methods for multidimensional numerical integration, namely Cartesian products of one-dimensional integration rules such as the trapezoidal rule and Simpson's rule (see [15]), work well only for dimensions up to three or four, or if the given high-dimensional integral can be reduced to a low-dimensional integral by analytic tricks. For most high-dimensional integrals arising in finance, the classical methods fail.

A more powerful approach to multidimensional numerical integration employs Monte Carlo methods. In a nutshell, a *Monte Carlo method* is a numerical method based on random sampling. A comprehensive treatment of Monte Carlo methods can be found in [20].

Monte Carlo methods for numerical integration can be explained in a straightforward manner. In many cases, by using suitable transformations, we can assume that the integration domain is an s -dimensional unit cube $I^s := [0, 1]^s$, so this is the situation on which we focus. We assume also that the integrand f is square integrable over I^s . Then the *Monte Carlo approximation* for the integral is

$$\int_{I^s} f(\mathbf{u}) \, d\mathbf{u} \approx \frac{1}{N} \sum_{n=1}^N f(\mathbf{x}_n) \quad (1)$$

where $\mathbf{x}_1, \dots, \mathbf{x}_N$ are independent random samples drawn from the uniform distribution on I^s . In statistics, we approximate the expected value of a random variable by sample means. The law of large numbers

guarantees that with probability 1 (i.e., for “almost all” sequences of sample points) we have

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N f(\mathbf{x}_n) = \int_{I^s} f(\mathbf{u}) \, d\mathbf{u} \quad (2)$$

and so the Monte Carlo method for numerical integration converges almost surely.

We can, in fact, be more precise about the error committed in the Monte Carlo approximation (1). It can be verified quite easily that the square of the error in equation (1) is, on the average over all samples of size N , equal to $\sigma^2(f)N^{-1}$, where $\sigma^2(f)$ is the variance of f . Thus, with overwhelming probability we have

$$\int_{I^s} f(\mathbf{u}) \, d\mathbf{u} - \frac{1}{N} \sum_{n=1}^N f(\mathbf{x}_n) = O(N^{-1/2}) \quad (3)$$

This means, very roughly, that if we want to compute the given integral with an error tolerance of the order 10^{-2} , say, then we need about 10^4 sample points, and to reduce the error tolerance by one order of magnitude, we need to increase the sample size by a factor of about 100. An important fact here is that the convergence rate in equation (3) is independent of the dimension s , and this makes Monte Carlo methods attractive for high-dimensional problems.

Despite the initial appeal of Monte Carlo methods for numerical integration, there are several drawbacks of these methods: (i) it is difficult to generate truly random samples; (ii) Monte Carlo methods for numerical integration provide only probabilistic error bounds; and (iii) in many applications the convergence rate in equation (3) is considered too slow. Quasi-Monte Carlo (QMC) methods were introduced to address these concerns.

General Background on QMC Methods

A *QMC method* is a deterministic version of a Monte Carlo method, in the sense that the random samples used in the implementation of a Monte Carlo method are replaced by *quasi-random points*, which are judiciously chosen deterministic points with good distribution properties. The general idea is that the Monte Carlo error bound (3) describes the average performance of integration points $\mathbf{x}_1, \dots, \mathbf{x}_N$, and there should exist points that perform better than

average. These are the quasi-random points we are seeking.

We again consider these methods in the context of numerical integration over an s -dimensional unit cube $I^s = [0, 1]^s$. The approximation scheme is the same as for the Monte Carlo method, namely

$$\int_{I^s} f(\mathbf{u}) \, d\mathbf{u} \approx \frac{1}{N} \sum_{n=1}^N f(\mathbf{x}_n) \quad (4)$$

but now $\mathbf{x}_1, \dots, \mathbf{x}_N$ are deterministic points in I^s . For such a deterministic numerical integration scheme we expect a deterministic error bound, and this is indeed provided by the Koksma–Hlawka inequality. It depends on the star discrepancy, a measure for the irregularity of distribution of a point set P consisting of $\mathbf{x}_1, \dots, \mathbf{x}_N \in I^s$. For any Borel set $M \subseteq I^s$, let $A(M; P)$ be the number of integers n , with $1 \leq n \leq N$ such that $\mathbf{x}_n \in M$. We put

$$R(M; P) = \frac{A(M; P)}{N} - \lambda_s(M) \quad (5)$$

which is the difference between the relative frequency of the points of P in M and the s -dimensional Lebesgue measure $\lambda_s(M)$ of M . If the points of P have a very uniform distribution over I^s , then the values of $R(M; P)$ will be close to 0 for a reasonable collection of Borel sets, such as for all subintervals of I^s .

Definition 1 *The star discrepancy of the point set P is given by*

$$D_N^* = D_N^*(P) = \sup_J |R(J; P)| \quad (6)$$

where the supremum is extended over all intervals $J = \prod_{i=1}^s [0, u_i]$ with $0 < u_i \leq 1$ for $1 \leq i \leq s$.

Koksma–Hlawka Inequality

For any function f of bounded variation $V(f)$ on I^s in the sense of Hardy and Krause and any points $\mathbf{x}_1, \dots, \mathbf{x}_N \in [0, 1]^s$, we have

$$\left| \int_{I^s} f(\mathbf{u}) \, d\mathbf{u} - \frac{1}{N} \sum_{n=1}^N f(\mathbf{x}_n) \right| \leq V(f) D_N^* \quad (7)$$

where D_N^* is the star discrepancy of $\mathbf{x}_1, \dots, \mathbf{x}_N$.

Note that $V(f)$ is a measure for the oscillation of the function f . The precise definition of the variation

$V(f)$ can be found in [49, p. 19]. For $f(\mathbf{u}) = f(u_1, \dots, u_s)$, a sufficient condition for $V(f) < \infty$ is that the partial derivative $\partial^s f / \partial u_1 \cdots \partial u_s$ be continuous on I^s . A detailed proof of the Koksma–Hlawka inequality is by Kuipers and Niederreiter [37, Section 2.5]. There are all types of variants of this inequality; see [46, 49, Section 2.2, 24].

A different kind of error bound for QMC integration was shown by Niederreiter [50]. It relies on the following concept.

Definition 2 *Let $\mathcal{M}$ be a nonempty collection of Borel sets in I^s . Then a point set P of elements of I^s is called $(\mathcal{M}, \lambda_s)$ -uniform if $R(M; P) = 0$ for all $M \in \mathcal{M}$.*

Now let $\mathcal{M} = \{M_1, \dots, M_k\}$ be a partition of I^s into nonempty Borel subsets of I^s . For a bounded Lebesgue-integrable function f on I^s and for $1 \leq j \leq k$, we put

$$G_j(f) = \sup_{\mathbf{u} \in M_j} f(\mathbf{u}), \quad g_j(f) = \inf_{\mathbf{u} \in M_j} f(\mathbf{u}) \quad (8)$$

Then for any $(\mathcal{M}, \lambda_s)$ -uniform point set consisting of $\mathbf{x}_1, \dots, \mathbf{x}_N \in I^s$ we have the error bound

$$\begin{aligned} & \left| \int_{I^s} f(\mathbf{u}) \, d\mathbf{u} - \frac{1}{N} \sum_{n=1}^N f(\mathbf{x}_n) \right| \\ & \leq \sum_{j=1}^k \lambda_s(M_j) (G_j(f) - g_j(f)) \end{aligned} \quad (9)$$

An analog of the bound (9) holds, in fact, for numerical integration over any probability space (see [50]).

A family of QMC methods that is particularly suited for periodic integrands is formed by lattice rules. For a given dimension $s \geq 1$, consider the factor group $\mathbb{R}^s / \mathbb{Z}^s$, which is an abelian group under addition of residue classes. Let $L / \mathbb{Z}^s$ be an arbitrary finite subgroup of $\mathbb{R}^s / \mathbb{Z}^s$ and let $\mathbf{x}_n + \mathbb{Z}^s$ with $\mathbf{x}_n \in [0, 1]^s$ for $1 \leq n \leq N$ be the distinct residue classes making up the group $L / \mathbb{Z}^s$. The points $\mathbf{x}_1, \dots, \mathbf{x}_N$ form the integration points of an N -point lattice rule. This terminology stems from the fact that the subset $L = \bigcup_{n=1}^N (\mathbf{x}_n + \mathbb{Z}^s)$ of $\mathbb{R}^s$ is an s -dimensional lattice. The dual lattice $L^\perp$ of L is defined by

$$L^\perp = \{\mathbf{h} \in \mathbb{Z}^s : \mathbf{h} \cdot \mathbf{x} \in \mathbb{Z} \text{ for all } \mathbf{x} \in L\} \quad (10)$$

where $\mathbf{h} \cdot \mathbf{x}$ denotes the standard inner product of $\mathbf{h}$ and $\mathbf{x}$. For real numbers, $\alpha > 1$ and $C > 0$, let $\mathcal{E}_\alpha^s(C)$ be the class of all continuous periodic functions f on $\mathbb{R}^s$ with period interval I^s and with Fourier coefficients $\hat{f}(\mathbf{h})$ satisfying

$$|\hat{f}(\mathbf{h})| \leq Cr(\mathbf{h})^{-\alpha} \quad \text{for all nonzero } \mathbf{h} \in \mathbb{Z}^s \quad (11)$$

where for $\mathbf{h} = (h_1, \dots, h_s) \in \mathbb{Z}^s$, we put

$$r(\mathbf{h}) = \prod_{i=1}^s \max(1, |h_i|) \quad (12)$$

Then it can be shown that with the notation above,

$$\begin{aligned} \max_{f \in \mathcal{E}_\alpha^s(C)} \left| \int_{I^s} f(\mathbf{u}) d\mathbf{u} - \frac{1}{N} \sum_{n=1}^N f(\mathbf{x}_n) \right| \\ = C \sum_{\mathbf{h} \in \mathbb{Z}^s \setminus \{0\}} r(\mathbf{h})^{-\alpha} \end{aligned} \quad (13)$$

Further analysis leads to the result that for any $s \geq 2$ and $N \geq 2$ there exists an s -dimensional N -point lattice rule with an error bound of order $O(N^{-\alpha}(\log N)^{c(\alpha,s)})$ for all $f \in \mathcal{E}_\alpha^s(C)$, where the exponent $c(\alpha, s) > 0$ depends only on α and s . Expository accounts of the theory of lattice rules are given in [49, Chapters 5, 71]. A more recent detailed discussion of lattice rules can be found in [25]. Algorithms for the construction of efficient lattice rules are presented, for example, in [16, 58, 72].

This article can present only a rough overview of QMC methods. For a full treatment of QMC methods, we refer to [49]. Developments from the invention of QMC methods in the early 1950s up to 1978 are covered in detail in the survey article [46].

Low-discrepancy Sequences

The Koksma–Hlawka inequality leads to the conclusion that point sets with small star discrepancy guarantee small errors in QMC integration over I^s . This raises the question of how small we can make the star discrepancy of N points in I^s for fixed N and s . For any $N \geq 2$ and $s \geq 1$, the least order of magnitude that can be achieved at present is

$$D_N^*(P) = O(N^{-1}(\log N)^{s-1}) \quad (14)$$

where the implied constant is independent of N . (Strictly speaking, one has to consider infinitely many values of N , that is, an infinite collection of point sets of increasing size, for this O -bound to make sense in a rigorous fashion, but this technicality is often ignored.) A point set P achieving equation (14) is called a *low-discrepancy point set*. The points in a low-discrepancy point set are an ideal form of quasi-random points. It is conjectured that the order of magnitude in equation (14) is best possible, that is, the star discrepancy of any $N \geq 2$ points in I^s is at least of the order of magnitude $N^{-1}(\log N)^{s-1}$. This conjecture is proved for $s = 1$ and $s = 2$ (see [37, Sections 2.1 and 2.2]).

A very useful concept is that of a *low-discrepancy sequence*, which is an infinite sequence S of points in I^s such that for all $N \geq 2$ the star discrepancy $D_N^*(S)$ of the first N terms of S satisfies

$$D_N^*(S) = O(N^{-1}(\log N)^s) \quad (15)$$

with an implied constant independent of N . It is conjectured that the order of magnitude in equation (15) is best possible, but in this case the conjecture has been verified only for $s = 1$ (see [37, Section 2.2]).

Low-discrepancy sequences have several practical advantages. In the first place, if $\mathbf{x}_1, \mathbf{x}_2, \dots \in I^s$ is a low-discrepancy sequence and $N \geq 2$ is an integer, then it is easily seen that the points

$$\mathbf{y}_n = \left(\frac{n-1}{N}, \mathbf{x}_n \right) \in I^{s+1}, \quad n = 1, \dots, N \quad (16)$$

form a low-discrepancy point set. Thus, if a low-discrepancy sequence has been constructed, then we immediately obtain arbitrarily large low-discrepancy point sets. Hence, in the following, we concentrate on the construction of low-discrepancy sequences. Furthermore, given a low-discrepancy sequence S and a budget of N integration points, we can simply use the first N terms of the sequence S to get a good QMC method. If later on we want to increase N to achieve a higher accuracy, we can do so while retaining the results of the earlier computation. This is an advantage of low-discrepancy sequences over low-discrepancy point sets.

It is clear from the Koksma–Hlawka inequality and equation (15) that if we apply QMC integration with an integrand f of bounded variation on I^s in the sense of Hardy and Krause and with the first N terms

$\mathbf{x}_1, \dots, \mathbf{x}_N \in [0, 1]^s$ of a low-discrepancy sequence, then

$$\int_{I^s} f(\mathbf{u}) \, d\mathbf{u} - \frac{1}{N} \sum_{n=1}^N f(\mathbf{x}_n) = O(N^{-1}(\log N)^s) \quad (17)$$

This yields a significantly faster convergence rate than the convergence rate $O(N^{-1/2})$ in equation (3). Thus, for many types of integrals, the QMC method will outperform the Monte Carlo method.

Over the years, various constructions of low-discrepancy sequences have been obtained. Historically, the first one was designed by Halton [23]. For integers $b \geq 2$ and $n \geq 1$, let

$$n-1 = \sum_{j=0}^{\infty} a_j(n) b^j, \quad a_j(n) \in \{0, 1, \dots, b-1\} \quad (18)$$

be the digit expansion of $n-1$ in base b . Then put

$$\phi_b(n) = \sum_{j=0}^{\infty} a_j(n) b^{-j-1} \quad (19)$$

Now let $p_1 = 2, p_2 = 3, \dots, p_s$ be the first s prime numbers. Then

$$\mathbf{x}_n = (\phi_{p_1}(n), \dots, \phi_{p_s}(n)) \in I^s, \quad n = 1, 2, \dots \quad (20)$$

is the *Halton sequence* in the bases $p_1, \dots, p_s$. This sequence S satisfies

$$D_N^*(S) = O(N^{-1}(\log N)^s) \quad (21)$$

for all $N \geq 2$, with an implied constant depending only on s . The standard software implementation of Halton sequences is that of Fox [21]. More sophisticated constructions of better low-discrepancy sequences are described in the following section.

Nets and $(\mathbf{T}, s)$ -Sequences

Current methods of constructing low-discrepancy sequences rely on the following definition, which is a special case of Definition 2.

Definition 3 Let $s \geq 1$, $b \geq 2$, and $0 \leq t \leq m$ be integers and let $\mathcal{M}_{b,m,t}^{(s)}$ be the collection of all subintervals J of I^s of the form

$$J = \prod_{i=1}^s [a_i b^{-d_i}, (a_i + 1) b^{-d_i}) \quad (22)$$

with integers $d_i \geq 0$ and $0 \leq a_i < b^{d_i}$ for $1 \leq i \leq s$ and with $\lambda_s(J) = b^{t-m}$. Then an $(\mathcal{M}_{b,m,t}^{(s)}, \lambda_s)$ -uniform point set consisting of b^m points in I^s is called a (t, m, s) -net in base b .

It is important to note that the smaller the value of t for given b, m , and s , the larger the collection $\mathcal{M}_{b,m,t}^{(s)}$ of intervals in Definition 3, and so the stronger the uniform point set property in Definition 2. Thus, the primary interest is in (t, m, s) -nets in base b with a small value of t .

There is an important sequence analog of Definition 3. Given a real number $x \in [0, 1]$, let

$$x = \sum_{j=1}^{\infty} z_j b^{-j}, \quad z_j \in \{0, 1, \dots, b-1\} \quad (23)$$

be a b -adic expansion of x , where the case $z_j = b-1$ for all but finitely many j is allowed. For an integer $m \geq 1$, we define the truncation

$$[x]_{b,m} = \sum_{j=1}^m z_j b^{-j} \quad (24)$$

If $\mathbf{x} = (x^{(1)}, \dots, x^{(s)}) \in I^s$ and the $x^{(i)}$, $1 \leq i \leq s$, are given by prescribed b -adic expansions, then we define

$$[\mathbf{x}]_{b,m} = ([x^{(1)}]_{b,m}, \dots, [x^{(s)}]_{b,m}) \quad (25)$$

We write $\mathbb{N}$ for the set of positive integers and $\mathbb{N}_0$ for the set of nonnegative integers.

Definition 4 Let $s \geq 1$ and $b \geq 2$ be integers and let $\mathbf{T} : \mathbb{N} \rightarrow \mathbb{N}_0$ be a function with $\mathbf{T}(m) \leq m$ for all $m \in \mathbb{N}$. Then a sequence $\mathbf{x}_1, \mathbf{x}_2, \dots$ of points in I^s is a $(\mathbf{T}, s)$ -sequence in base b if for all $k \in \mathbb{N}_0$ and $m \in \mathbb{N}$, the points $[\mathbf{x}_n]_{b,m}$ with $kb^m < n \leq (k+1)b^m$ form a $(\mathbf{T}(m), m, s)$ -net in base b . If for some integer $t \geq 0$, we have $\mathbf{T}(m) = m$ for $m \leq t$ and $\mathbf{T}(m) = t$ for $m > t$, then we speak of a (t, s) -sequence in base b .

A general theory of (t, m, s) -nets and (t, s) -sequences was developed by Niederreiter [47]. The

concept of a $(\mathbf{T}, s)$ -sequence was introduced by Larcher and Niederreiter [40], with the variant in Definition 4 being due to Niederreiter and Özbudak [53]. Recent surveys of this topic are presented in [51, 52]. For a (t, s) -sequence in base b we have

$$D_N^*(S) = O(b^t N^{-1} (\log N)^s) \quad (26)$$

for all $N \geq 2$, where the implied constant depends only on b and s . Thus, any (t, s) -sequence is a low-discrepancy sequence.

The standard technique of constructing $(\mathbf{T}, s)$ -sequences is the *digital method*. Fix a dimension $s \geq 1$ and a base $b \geq 2$. Let R be a finite commutative ring with identity and of order b . Set up a map $\rho : R^\infty \rightarrow [0, 1]$ by selecting a bijection $\eta : R \rightarrow \mathbb{Z}_b := \{0, 1, \dots, b-1\}$ and putting

$$\rho(r_1, r_2, \dots) = \sum_{j=1}^{\infty} \eta(r_j) b^{-j} \quad \text{for } (r_1, r_2, \dots) \in R^\infty \quad (27)$$

Furthermore, choose $\infty \times \infty$ matrices $C^{(1)}, \dots, C^{(s)}$ over R which are called *generating matrices*. For $n = 1, 2, \dots$ let

$$n-1 = \sum_{j=0}^{\infty} a_j(n) b^j, \quad a_j(n) \in \mathbb{Z}_b \quad (28)$$

be the digit expansion of $n-1$ in base b . Choose a bijection $\psi : \mathbb{Z}_b \rightarrow R$ with $\psi(0) = 0$ and associate with n the sequence

$$\mathbf{n} = (\psi(a_0(n)), \psi(a_1(n)), \dots) \in R^\infty \quad (29)$$

Then the sequence $\mathbf{x}_1, \mathbf{x}_2, \dots$ of points in I^s is defined by

$$\mathbf{x}_n = (\rho(\mathbf{n}C^{(1)}), \dots, \rho(\mathbf{n}C^{(s)})) \quad \text{for } n = 1, 2, \dots \quad (30)$$

Note that the products $\mathbf{n}C^{(i)}$ are well defined since $\mathbf{n}$ contains only finitely many nonzero terms. In practice, the ring R is usually chosen to be a finite field $\mathbb{F}_q$ of order q , where q is a prime power. The success of the digital method depends on a careful choice of the generating matrices $C^{(1)}, \dots, C^{(s)}$.

The first application of the digital method occurred in the construction of *Sobol' sequences* in [76]. This construction uses primitive polynomials over $\mathbb{F}_2$ to set up the generating matrices $C^{(1)}, \dots, C^{(s)}$ and leads

to (t, s) -sequences in base 2. The wider family of irreducible polynomials was used in the construction of *Niederreiter sequences* in [48], and this construction works for arbitrary prime-power bases q . Let $p_1, \dots, p_s$ be the first s monic irreducible polynomials over $\mathbb{F}_q$, ordered according to nondecreasing degrees and put,

$$T_q(s) = \sum_{i=1}^s (\deg(p_i) - 1) \quad (31)$$

The construction of Niederreiter sequences yields (t, s) -sequences in base q with $t = T_q(s)$. Let $U(s)$ denote the least value of t that can be achieved by Sobol' sequences for given s . Then $T_2(s) = U(s)$ for $1 \leq s \leq 7$ and $T_2(s) < U(s)$ for all $s \geq 8$. Thus, according to equation (26), for all dimensions $s \geq 8$ Niederreiter sequences in base 2 lead to a smaller upper bound on the star discrepancy than Sobol' sequences. Convenient software implementations of Sobol' and Niederreiter sequences are described in [8–10].

The potentially smallest, and thus best, t -value for any (t, s) -sequence in base b would be $t = 0$. However, according to [49, Corollary 4.24], a necessary condition for the existence of a $(0, s)$ -sequence in base b is $s \leq b$. For primes p , a construction of $(0, s)$ -sequences in base p for $s \leq p$ was given by Faure [18]. For prime powers q , a construction of $(0, s)$ -sequences in base q for $s \leq q$ was given by Niederreiter [47]. Since $T_q(s) = 0$ for $s \leq q$ by equation (31), the Niederreiter sequences in [48] also yield $(0, s)$ -sequences in base q for $s \leq q$.

An important advance in the construction of low-discrepancy sequences was made in the mid-1990s with the design of *Niederreiter–Xing sequences*, which utilizes powerful tools from algebraic geometry and the theory of algebraic function fields and generalizes the construction of Niederreiter sequences. The basic articles here are [54, 83], and expository accounts of this work and further results are given in [55, 56, 57, Chapter 8]. Niederreiter–Xing sequences are (t, s) -sequences in a prime-power base q with $t = V_q(s)$. Here, $V_q(s)$ is a number determined by algebraic curves (or equivalently algebraic function fields) over $\mathbb{F}_q$, and we have $V_q(s) \leq T_q(s)$ for all $s \geq 1$. In fact, much more is true. If we fix q and consider $V_q(s)$ and $T_q(s)$ as functions of s , then $V_q(s)$ is of the order of magnitude s , whereas $T_q(s)$ is of the order of magnitude $s \log s$.

This yields an enormous improvement on the bound for the star discrepancy in equation (7). It is known that for any (t, s) -sequences in base b the parameter t must grow at least linearly with s for fixed b (see [55, Section 10]), and so Niederreiter–Xing sequences yield t -values of the optimal order of magnitude as a function of s . A software implementation of Niederreiter–Xing sequences is described in [68] and available at <http://math.iit.edu/~mcqmc/Software.html> by following the appropriate links.

To illustrate the comparative quality of the above constructions of (t, s) -sequences, we consider the case of the most convenient base 2 and tabulate some values of $U(s)$ for Sobol’ sequences, of $T_2(s)$ for Niederreiter sequences, and of $V_2(s)$ for Niederreiter–Xing sequences in Table 1. Note that the values of $V_2(s)$ in Table 1 for $2 \leq s \leq 7$ are the least values of t for which a (t, s) -sequence in base 2 can exist.

The approach by algebraic function fields was followed up recently by Mayor and Niederreiter [45] who gave an alternative construction of Niederreiter–Xing sequences using differentials of global function fields. Niederreiter and Özbudak [53] obtained the first improvement on Niederreiter–Xing sequences for some special pairs (q, s) of prime-power bases q and dimensions s . For instance, consider the case where q is an arbitrary prime power and $s = q + 1$. Then $T_q(q + 1) = 1$ by equation (8) and this is the least possible t -value for a $(t, q + 1)$ -sequence in base q . However, the construction in [53] yields a $(\mathbf{T}, q + 1)$ -sequence in base q with $\mathbf{T}(m) = 0$ for even m and $\mathbf{T}(m) = 1$ for odd m , which is even better.

We remark that all constructions mentioned in this section are based on the digital method. We note also that the extensive database at <http://mint.sbg.ac.at> is devoted to (t, m, s) -nets and (t, s) -sequences. In summary, for a given prime-power base q , the currently best low-discrepancy sequences are as follows: (i) the Faure or Niederreiter sequences (depending on whether q is prime or not) for all dimensions

$s \leq q$ and (ii) the Niederreiter–Xing sequences for all dimensions $s > q$, except for some special values of $s > q$ where the Niederreiter–Özbudak sequences are better. We emphasize that the bound (7) on the star discrepancy of (t, s) -sequences is completely explicit; see [49, Section 4.1] and a recent improvement in [36]. For the best (t, s) -sequences, the coefficient of the leading term $N^{-1}(\log N)^s$ in the bound on the star discrepancy tends to 0 at a superexponential rate as $s \rightarrow \infty$.

Effective Dimension

In view of equation (6), the QMC method for numerical integration performs asymptotically better than the Monte Carlo method for any dimension s . However, in practical terms, the number N of integration points cannot be taken too large, and then already for moderate values of s the size of the factor $(\log N)^s$ may wipe out the advantage over the Monte Carlo method. On the other hand, numerical experiments with many types of integrands have shown that even for large dimensions s the QMC method will often lead to a convergence rate $O(N^{-1})$ rather than $O(N^{-1}(\log N)^s)$ as predicted by the theory, thus beating the Monte Carlo method by a wide margin. One reason may be that the Koksma–Hlawka inequality is in general overly pessimistic. Another explanation can sometimes be given by means of the nature of the integrand f . Even though f is a function of s variables, the influence of these variables could differ greatly. For numerical purposes, f may behave like a function of much fewer variables, so that the numerical integration problem is in essence a low-dimensional one with a faster convergence rate. This idea is captured by the notion of effective dimension.

We start with the ANOVA decomposition of a random variable $f(\mathbf{u}) = f(u_1, \dots, u_s)$ on I^s of finite variance. This decomposition amounts to writing f in the form

$$f(\mathbf{u}) = \sum_{K \subseteq \{1, \dots, s\}} f_K(\mathbf{u}) \quad (32)$$

where $f_\emptyset$ is the expected value of f and each $f_K(\mathbf{u})$ with $K \neq \emptyset$ depends only on the variables u_i with $i \in K$ and has expected value 0. Furthermore, f_{K_1} and f_{K_2} are orthogonal whenever $K_1 \neq K_2$. Then the

Table 1 Values of $U(s)$, $T_2(s)$, and $V_2(s)$

s	2	3	4	5	6	7	8	9	10	15	20
$U(s)$	0	1	3	5	8	11	15	19	23	45	71
$T_2(s)$	0	1	3	5	8	11	14	18	22	43	68
$V_2(s)$	0	1	1	2	3	4	5	6	8	15	21

variance $\sigma^2(f)$ of f decomposes as

$$\sigma^2(f) = \sum_{K \subseteq \{1, \dots, s\}} \sigma^2(f_K) \quad (33)$$

The following definition relates to this decomposition.

Definition 5 *Let d be an integer with $1 \leq d \leq s$ and r a real number with $0 < r < 1$. Then the function f has effective dimension d at the rate r in the superposition sense if*

$$\sum_{|K| \leq d} \sigma^2(f_K) \geq r \sigma^2(f) \quad (34)$$

The function f has effective dimension d at the rate r in the truncation sense if

$$\sum_{K \subseteq \{1, \dots, d\}} \sigma^2(f_K) \geq r \sigma^2(f) \quad (35)$$

Values of r of practical interest are $r = 0.95$ and $r = 0.99$, for instance. The formalization of the idea of effective dimension goes back to the articles of Caflisch *et al.* [13] and Hickernell [25]. There are many problems of high-dimensional numerical integration arising in computational finance for which the integrands have a relatively small effective dimension, one possible reason being discount factors, which render variables corresponding to distant time horizons essentially negligible. The classical example here is that of the valuation of mortgage-backed securities (see [13, 66]). For further interesting work on the ANOVA decomposition and effective dimension, with applications to the pricing of Asian and barrier options, we refer to [32, 42, 43, 81].

A natural way of capturing the relative importance of variables is to attach weights to them. More generally, one may attach weights to any collection of variables, thus measuring the relative importance of all projections—and not just of the one-dimensional projections—of the given integrand. This leads then to a weighted version of the theory of QMC methods, an approach that was pioneered by Sloan and Woźniakowski [75].

Given a dimension s we consider the set $\{1, \dots, s\}$ of coordinate indices. To any nonempty subset K of $\{1, \dots, s\}$ we attach a weight $\gamma_K \geq 0$. To avoid a trivial case, we assume that not all weights are 0. Let γ denote the collection of all these weights γ_K . Then we

introduce the *weighted star discrepancy* $D_{N,\gamma}^*$, which generalizes Definition 1. For $\mathbf{u} = (u_1, \dots, u_s) \in I^s$, we abbreviate the interval $\prod_{i=1}^s [0, u_i]$ by $[\mathbf{0}, \mathbf{u}]$. For any nonempty $K \subseteq \{1, \dots, s\}$, let $\mathbf{u}_K$ denote the point in I^s with all coordinates whose indices are not in K replaced by 1. Now for a point set P consisting of N points from I^s , we define

$$D_{N,\gamma}^* = \sup_{\mathbf{u} \in I^s} \max_K \gamma_K |R([\mathbf{0}, \mathbf{u}_K]; P)| \quad (36)$$

We recover the classical star discrepancy if we choose $\gamma_{\{1, \dots, s\}} = 1$ and $\gamma_K = 0$ for all nonempty proper subsets K of $\{1, \dots, s\}$. With this weighted star discrepancy, one can then prove a weighted analog of the Koksma–Hlawka inequality (see [75]).

There are some special kinds of weights that are of great practical interest. In the case of *product weights*, one attaches a weight η_i to each $i \in \{1, \dots, s\}$ and puts

$$\gamma_K = \prod_{i \in K} \eta_i \quad \text{for all } K \subseteq \{1, \dots, s\}, K \neq \emptyset \quad (37)$$

In the case of *finite-order weights*, one chooses a threshold k and puts $\gamma_K = 0$ for all K of cardinality larger than k .

The theoretical analysis of the performance of weighted QMC methods requires the introduction of weighted function spaces in which the integrands live. These can, for instance, be weighted Sobolev spaces or weighted Korobov spaces. In this context again, the weights reflect the relative importance of variables or collections of variables. The articles [38, 70, 75] are representative for this approach.

The analysis of the integration error utilizing weighted function spaces also leads to powerful results on tractability, a concept stemming from the theory of information-based complexity. The emphasis here is on the performance of multidimensional numerical integration schemes as a function not only of the number N of integration points but also of the dimension s as $s \rightarrow \infty$. Let $\mathcal{F}_s$ be a Banach space of integrands f on I^s with norm $\|f\|$. Write

$$L_s(f) = \int_{I^s} f(\mathbf{u}) \, d\mathbf{u} \quad \text{for } f \in \mathcal{F}_s \quad (38)$$

Consider numerical integration schemes of the form

$$\mathcal{A}(f) = \sum_{n=1}^N a_n f(\mathbf{x}_n) \quad (39)$$

with real numbers $a_1, \dots, a_N$ and points $\mathbf{x}_1, \dots, \mathbf{x}_N \in I^s$. The QMC method is of course a special case of such a scheme. For $\mathcal{A}$ as in equation (39), we write $\text{card}(\mathcal{A}) = N$. Furthermore, we put

$$\text{err}(\mathcal{A}, \mathcal{F}_s) = \sup_{\|f\| \leq 1} |L_s(f) - \mathcal{A}(f)| \quad (40)$$

For any $N \geq 1$ and $s \geq 1$, the N th minimal error of the s -dimensional numerical integration problem is defined by

$$\text{err}(N, \mathcal{F}_s) = \inf \{ \text{err}(\mathcal{A}, \mathcal{F}_s) : \mathcal{A} \text{ with } \text{card}(\mathcal{A}) = N \} \quad (41)$$

The numerical integration problem is called *tractable* if there exist constants $C \geq 0$, $e_1 \geq 0$, and $e_2 > 0$ such that

$$\begin{aligned} \text{err}(N, \mathcal{F}_s) &\leq C s^{e_1} N^{-e_2} \|L_s\|_{\text{op}} \\ \text{for all } N &\geq 1, s \geq 1 \end{aligned} \quad (42)$$

where $\|L_s\|_{\text{op}}$ is the operator norm of L_s . If, in addition, the exponent e_1 may be chosen to be 0, then the problem is said to be *strongly tractable*.

Tractability and strong tractability depend very much on the choice of the spaces $\mathcal{F}_s$. Weighted function spaces using product weights have proved particularly effective in this connection. Since the interest is in $s \rightarrow \infty$, product weights are set up by choosing a sequence $\eta_1, \eta_2, \dots$ of positive numbers and then, for fixed $s \geq 1$, defining appropriate weights γ_K by equation (37). If the η_i tend to 0 sufficiently quickly as $i \rightarrow \infty$, then in a Hilbert-space setting strong tractability can be achieved by QMC methods based on Halton, Sobol', or Niederreiter sequences (see [29, 79]). Further results on (strong) tractability as it relates to QMC methods can be found, for example, in [27, 28, 73, 74, 80, 82].

Randomized QMC

Conventional QMC methods are fully deterministic and thus do not allow statistical error estimation as in Monte Carlo methods. However, one may introduce an element of randomness into a QMC method by randomizing (or “scrambling”) the deterministic integration points used in the method. In this way,

one can combine the advantages of QMC methods, namely faster convergence rates, and those of Monte Carlo methods, namely the possibility of error estimation.

Historically, the first scrambling scheme is *Cranley–Patterson rotation*, which was introduced in [14]. This scheme can be applied to any point set in I^s . Let $\mathbf{x}_1, \dots, \mathbf{x}_N \in I^s$ be given and put

$$\mathbf{y}_n = \{\mathbf{x}_n + \mathbf{r}\} \quad \text{for } n = 1, \dots, N \quad (43)$$

where $\mathbf{r}$ is a random vector uniformly distributed over I^s and $\{\cdot\}$ denotes reduction modulo 1 in each coordinate of a point in $\mathbb{R}^s$. This scheme transforms low-discrepancy point sets into low-discrepancy point sets.

A sophisticated randomization of (t, m, s) -nets and (t, s) -sequences is provided by *Owen scrambling* (see [60]). This scrambling scheme works with mutually independent random permutations of the digits in the b -adic expansions of the coordinates of all points in a (t, m, s) -net in base b or a (t, s) -sequence in base b . The scheme is set up in such a way that the scrambled version of a (t, m, s) -net, respectively (t, s) -sequence, in base b is a (t, m, s) -net, respectively (t, s) -sequence, in base b with probability 1. Further investigations of this scheme, particularly regarding the resulting mean square discrepancy and variance, were carried out, for example, by Hickernell and Hong [26], Hickernell and Yue [30], and Owen [61–63].

Since Owen scrambling is quite time consuming, various faster special versions have been proposed, such as a method of Matoušek [44] and the method of *digital shifts* in which the permutations in Owen scrambling are additive shifts modulo b and the shift parameters may depend on the coordinate index $i \in \{1, \dots, s\}$ and on the position of the digit in the digit expansion of the coordinate. In the binary case $b = 2$, digital shifting amounts to choosing s infinite bit strings $\mathbf{B}_1, \dots, \mathbf{B}_s$ and then taking each point $\mathbf{x}_n$ of the given (t, m, s) -net or (t, s) -sequence in base 2 and bitwise XORing the binary expansion of the i th coordinate of $\mathbf{x}_n$ with $\mathbf{B}_i$ for $1 \leq i \leq s$. Digital shifts and their applications are discussed, for example, in [17, 41]. The latter article presents also a general survey of randomized QMC methods and stresses the interpretation of these methods as variance reduction techniques.

Convenient scrambling schemes are also obtained by operating on the generating matrices of (t, m, s) -nets and (t, s) -sequences constructed by the digital method. The idea is to multiply the generating matrices by suitable random matrices from the left or from the right in such a way that the value of the parameter t is preserved. We refer to [19, 64] for such scrambling schemes. Software implementations of randomized low-discrepancy sequences are described in [22, 31] and are integrated into the Java library SSJ available at <http://www.iro.umontreal.ca/~simardr/ssj>, which contains also many other simulation tools.

Applications to Computational Finance

The application of Monte Carlo methods to challenging problems in computational finance was pioneered by Boyle [3] in 1977. Although QMC methods were already known at that time, they were not applied to computational finance because it was thought that they would be inefficient for problems involving the high dimensions occurring in this area.

A breakthrough came in the early 1990s when Paskov and Traub applied QMC integration to the problem of pricing a 30-year collateralized mortgage obligation provided by Goldman Sachs; see [67] for a report on this work. This problem required the computation of 10 integrals of dimension 360 each, and the results were astounding. For the hardest of the 10 integrals, the QMC method achieved accuracy 10^{-2} with just 170 points, whereas the Monte Carlo method needed 2700 points for the same accuracy. When higher accuracy is desired, the QMC method can be about 1000 times faster than the Monte Carlo method. For further work on the pricing of mortgage-backed securities, we refer to [13, 66, 78].

Applications of QMC methods to option pricing were first considered in the technical report of Birge [2] and the article of Joy *et al.* [35]. These works concentrated on European and Asian options. In the case of path-dependent options, if the security's terminal value depends only on the prices at s intermediate times, then after discretization the expected discounted payoff under the risk-neutral measure can be converted into an integral over the s -dimensional unit cube I^s . For instance, in [35] an Asian option with 53 time steps is studied numerically.

A related problem in which an s -dimensional integral arises is the pricing of a multiasset option

with s assets; see [1] in which numerical experiments comparing Monte Carlo and QMC methods are reported for dimensions up to $s = 100$. This article also discusses Brownian bridge constructions for option pricing. Related work on the pricing of multiasset European-style options using QMC and randomized QMC methods was carried out in [39, 69, 77], and comparative numerical experiments for Asian options can be found in [4, 59]. Jiang [33] gave a detailed error analysis of the pricing of European-style options using QMC methods, which is based on a variant of the bound (3) and requires only minimal smoothness assumptions.

Owing to its inherent difficulty, it took much longer for Monte Carlo and QMC methods to be applied to the problem of pricing American options. An excellent survey of early work on Monte Carlo methods for pricing American options is presented in [4]. The first important idea in this context was the *bundling algorithm* in which paths in state space for which the stock prices behave in a similar way are grouped together in the simulation. Initially, the bundling algorithm was applicable only to single-asset American options. Jin *et al.* [34] recently extended the bundling algorithm in order to price high-dimensional American-style options, and they also showed that computing representative states by a QMC method improves the performance of the algorithm.

Another approach to pricing American options by simulation is the *stochastic mesh method*. The choice of mesh density functions at each discrete time step is crucial for the success of this method. The standard mesh density functions are mixture densities, and so in a Monte Carlo approach one can use known techniques for generating random samples from mixture densities. In a QMC approach, these random samples are replaced by deterministic points whose empirical distribution function is close to the target distribution function. Work on the latter approach was carried out by Boyle, Kolkiewicz, and Tan [5–7] and Broadie *et al.* [11]. Another application of QMC methods to the pricing of American options occurs in *regression-based methods*, which are typically least-squares Monte Carlo methods. Here Caffisch and Chaudhary [12] have shown that QMC versions improve the performance of such methods.

We conclude by mentioning two more applications of QMC methods to computational finance, namely by Papageorgiou and Paskov [65] to value-at-risk

computations and by Jiang [33] to the pricing of interest-rate derivatives in a LIBOR market model.

References

- [1] Acworth, P., Broadie, M. & Glasserman, P. (1998). A comparison of some Monte Carlo and quasi Monte Carlo techniques for option pricing, in *Monte Carlo and Quasi-Monte Carlo Methods 1996*, H. Niederreiter, P. Hellekalek, G. Larcher & P. Zinterhof, eds, Springer, New York, pp. 1–18.
- [2] Birge, J.R. (1994). Quasi-Monte Carlo approaches to option pricing, Technical report 94–19, Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI.
- [3] Boyle, P.P. (1977). Options: a Monte Carlo approach, *Journal of Financial Economics* **4**, 323–338.
- [4] Boyle, P., Broadie, M. & Glasserman, P. (1997). Monte Carlo methods for security pricing, *Journal of Economic Dynamics and Control* **21**, 1267–1321.
- [5] Boyle, P.P., Kolkiewicz, A.W. & Tan, K.S. (2001). Valuation of the reset options embedded in some equity-linked insurance products, *North American Actuarial Journal* **5**(3), 1–18.
- [6] Boyle, P.P., Kolkiewicz, A.W. & Tan, K.S. (2002). Pricing American derivatives using simulation: a biased low approach, in *Monte Carlo and Quasi-Monte Carlo Methods 2000*, K.T. Fang, F.J. Hickernell & H. Niederreiter, eds, Springer, Berlin, pp. 181–200.
- [7] Boyle, P.P., Kolkiewicz, A.W. & Tan, K.S. (2003). An improved simulation method for pricing high-dimensional American derivatives, *Mathematics and Computers in Simulation* **62**, 315–322.
- [8] Bratley, P. & Fox, B.L. (1988). Algorithm 659: implementing Sobol's quasirandom sequence generator, *ACM Transactions on Mathematical Software* **14**, 88–100.
- [9] Bratley, P., Fox, B.L. & Niederreiter, H. (1992). Implementation and tests of low-discrepancy sequences, *ACM Transactions on Modeling and Computer Simulation* **2**, 195–213.
- [10] Bratley, P., Fox, B.L. & Niederreiter, H. (1994). Algorithm 738: programs to generate Niederreiter's low-discrepancy sequences, *ACM Transactions on Mathematical Software* **20**, 494–495.
- [11] Broadie, M., Glasserman, P. & Ha, Z. (2000). Pricing American options by simulation using a stochastic mesh with optimized weights, in *Probabilistic Constrained Optimization: Methodology and Applications*, S.P. Uryasev, ed, Kluwer Academic Publishers, Dordrecht, pp. 26–44.
- [12] Caflisch, R.E. & Chaudhary, S. (2004). Monte Carlo simulation for American options, in *A Celebration of Mathematical Modeling*, D. Givoli, M.J. Grote & G.C. Papanicolaou, eds, Kluwer Academic Publishers, Dordrecht, pp. 1–16.
- [13] Caflisch, R.E., Morokoff, M. & Owen, A. (1997). Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension, *The Journal of Computational Finance* **1**, 27–46.
- [14] Cranley, R. & Patterson, T.N.L. (1976). Randomization of number theoretic methods for multiple integration, *SIAM Journal on Numerical Analysis* **13**, 904–914.
- [15] Davis, P.J. & Rabinowitz, P. (1984). *Methods of Numerical Integration*, 2nd Edition, Academic Press, New York.
- [16] Dick, J. & Kuo, F.Y. (2004). Constructing good lattice rules with millions of points, in *Monte Carlo and Quasi-Monte Carlo Methods 2002*, H. Niederreiter, ed, Springer, Berlin, pp. 181–197.
- [17] Dick, J. & Pillichshammer, F. (2005). Multivariate integration in weighted Hilbert spaces based on Walsh functions and weighted Sobolev spaces, *Journal of Complexity* **21**, 149–195.
- [18] Faure, H. (1982). Discrepance de suites associées à un système de numération (en dimension s), *Acta Arithmetica* **41**, 337–351.
- [19] Faure, H. & Tezuka, S. (2002). Another random scrambling of digital (t, s) -sequences, in *Monte Carlo and Quasi-Monte Carlo Methods 2000*, K.T. Fang, F.J. Hickernell & H. Niederreiter, eds, Springer, Berlin, pp. 242–256.
- [20] Fishman, G.S. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*, Springer, New York.
- [21] Fox, B.L. (1986). Algorithm 647: implementation and relative efficiency of quasirandom sequence generators, *ACM Transactions on Mathematical Software* **12**, 362–376.
- [22] Friedel, I. & Keller, A. (2002). Fast generation of randomized low-discrepancy point sets, in *Monte Carlo and Quasi-Monte Carlo Methods 2000*, K.T. Fang, F.J. Hickernell & H. Niederreiter, eds, Springer, Berlin, pp. 257–273.
- [23] Halton, J.H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals, *Numerische Mathematik* **2**, 84–90, 196.
- [24] Hickernell, F.J. (1998). A generalized discrepancy and quadrature error bound, *Mathematics of Computation* **67**, 299–322.
- [25] Hickernell, F.J. (1998). Lattice rules: how well do they measure up? in *Random and Quasi-Random Point Sets*, P. Hellekalek & G. Larcher, eds, Springer, New York, pp. 109–166.
- [26] Hickernell, F.J. & Hong, H.S. (1999). The asymptotic efficiency of randomized nets for quadrature, *Mathematics of Computation* **68**, 767–791.
- [27] Hickernell, F.J., Sloan, I.H. & Wasilkowski, G.W. (2004). On tractability of weighted integration for certain Banach spaces of functions, in *Monte Carlo and Quasi-Monte Carlo Methods 2002*, H. Niederreiter, ed, Springer, Berlin, pp. 51–71.
- [28] Hickernell, F.J., Sloan, I.H. & Wasilkowski, G.W. (2004). The strong tractability of multivariate integration using lattice rules, in *Monte Carlo and Quasi-Monte Carlo Methods 2002*, H. Niederreiter, ed, Springer, Berlin, pp. 259–273.

-
- [29] Hickernell, F.J. & Wang, X.Q. (2002). The error bounds and tractability of quasi-Monte Carlo algorithms in infinite dimension, *Mathematics of Computation* **71**, 1641–1661.
 - [30] Hickernell, F.J. & Yue, R.-X. (2001). The mean square discrepancy of scrambled (t, s) -sequences, *SIAM Journal on Numerical Analysis* **38**, 1089–1112.
 - [31] Hong, H.S. & Hickernell, F.J. (2003). Algorithm 823: implementing scrambled digital sequences, *ACM Transactions on Mathematical Software* **29**, 95–109.
 - [32] Imai, J. & Tan, K.S. (2004). Minimizing effective dimension using linear transformation, in *Monte Carlo and Quasi-Monte Carlo Methods 2002*, H. Niederreiter, ed, Springer, Berlin, pp. 275–292.
 - [33] Jiang, X.F. (2007). *Quasi-Monte Carlo methods in finance*. Ph.D. dissertation, Northwestern University, Evanston, IL.
 - [34] Jin, X., Tan, H.H. & Sun, J.H. (2007). A state-space partitioning method for pricing high-dimensional American-style options, *Mathematical Finance* **17**, 399–426.
 - [35] Joy, C., Boyle, P.P. & Tan, K.S. (1996). Quasi-Monte Carlo methods in numerical finance, *Management Science* **42**, 926–938.
 - [36] Kritzer, P. (2006). Improved upper bounds on the star discrepancy of (t, m, s) -nets and (t, s) -sequences, *Journal of Complexity* **22**, 336–347.
 - [37] Kuipers, L. & Niederreiter, H. (1974). *Uniform Distribution of Sequences*, Wiley, New York. Reprint by Dover Publications, Mineola, NY, 2006.
 - [38] Kuo, F.Y. (2003). Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces, *Journal of Complexity* **19**, 301–320.
 - [39] Lai, Y.Z. & Spanier, J. (2000). Applications of Monte Carlo/quasi-Monte Carlo methods in finance: option pricing, in *Monte Carlo and Quasi-Monte Carlo Methods 1998*, H. Niederreiter & J. Spanier, eds, Springer, Berlin, pp. 284–295.
 - [40] Larcher, G. & Niederreiter, H. (1995). Generalized (t, s) -sequences, Kronecker-type sequences, and diophantine approximations of formal Laurent series, *Transactions of the American Mathematical Society* **347**, 2051–2073.
 - [41] L'Ecuyer, P. & Lemieux, C. (2002). Recent advances in randomized quasi-Monte Carlo methods, in *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, M. Dror, P. L'Ecuyer & F. Szidarovszky, eds, Kluwer Academic Publishers, Boston, pp. 419–474.
 - [42] Lemieux, C. & Owen, A.B. (2002). Quasi-regression and the relative importance of the ANOVA components of a function, in *Monte Carlo and Quasi-Monte Carlo Methods 2000*, K.T. Fang, F.J. Hickernell & H. Niederreiter, eds, Springer, Berlin, pp. 331–344.
 - [43] Liu, R.X. & Owen, A.B. (2006). Estimating mean dimensionality of analysis of variance decompositions, *Journal of the American Statistical Association* **101**, 712–721.
 - [44] Matoušek, J. (1998). On the L_2 -discrepancy for anchored boxes, *Journal of Complexity* **14**, 527–556.
 - [45] Mayor, D.J.S. & Niederreiter, H. (2007). A new construction of (t, s) -sequences and some improved bounds on their quality parameter, *Acta Arithmetica* **128**, 177–191.
 - [46] Niederreiter, H. (1978). Quasi-Monte Carlo methods and pseudo-random numbers, *Bulletin of the American Mathematical Society* **84**, 957–1041.
 - [47] Niederreiter, H. (1987). Point sets and sequences with small discrepancy, *Monatshefte für Mathematik* **104**, 273–337.
 - [48] Niederreiter, H. (1988). Low-discrepancy and low-dispersion sequences, *Journal of Number Theory* **30**, 51–70.
 - [49] Niederreiter, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*, SIAM, Philadelphia.
 - [50] Niederreiter, H. (2003). Error bounds for quasi-Monte Carlo integration with uniform point sets, *Journal of Computational and Applied Mathematics* **150**, 283–292.
 - [51] Niederreiter, H. (2005). Constructions of (t, m, s) -nets and (t, s) -sequences, *Finite Fields and Their Applications* **11**, 578–600.
 - [52] Niederreiter, H. (2008). Nets, (t, s) -sequences, and codes, in *Monte Carlo and Quasi-Monte Carlo Methods 2006*, A. Keller, S. Heinrich & H. Niederreiter, eds, Springer, Berlin, pp. 83–100.
 - [53] Niederreiter, H. & Özbudak, F. (2007). Low-discrepancy sequences using duality and global function fields, *Acta Arithmetica* **130**, 79–97.
 - [54] Niederreiter, H. & Xing, C.P. (1996). Low-discrepancy sequences and global function fields with many rational places, *Finite Fields and Their Applications* **2**, 241–273.
 - [55] Niederreiter, H. & Xing, C.P. (1996). Quasirandom points and global function fields, in *Finite Fields and Applications*, S. Cohen & H. Niederreiter, eds, Cambridge University Press, Cambridge, pp. 269–296.
 - [56] Niederreiter, H. & Xing, C.P. (1998). Nets, (t, s) -sequences, and algebraic geometry, in *Random and Quasi-Random Point Sets*, P. Hellekalek & G. Larcher, eds, Springer, New York, pp. 267–302.
 - [57] Niederreiter, H. & Xing, C.P. (2001). *Rational Points on Curves over Finite Fields: Theory and Applications*, Cambridge University Press, Cambridge.
 - [58] Nuyens, D. & Cools, R. (2006). Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces, *Mathematics of Computation* **75**, 903–920.
 - [59] Ökten, G. & Eastman, W. (2004). Randomized quasi-Monte Carlo methods in pricing securities, *Journal of Economic Dynamics and Control* **28**, 2399–2426.
 - [60] Owen, A.B. (1995). Randomly permuted (t, m, s) -nets and (t, s) -sequences, in *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, H. Niederreiter & P.J.-S. Shiue, eds, Springer, New York, pp. 299–317.

- [61] Owen, A.B. (1997). Monte Carlo variance of scrambled net quadrature, *SIAM Journal on Numerical Analysis* **34**, 1884–1910.
- [62] Owen, A.B. (1997). Scrambled net variance for integrals of smooth functions, *The Annals of Statistics* **25**, 1541–1562.
- [63] Owen, A.B. (1998). Scrambling Sobol’ and Niederreiter-Xing points, *Journal of Complexity* **14**, 466–489.
- [64] Owen, A.B. (2003). Variance with alternative scramblings of digital nets, *ACM Transactions on Modeling and Computer Simulation* **13**, 363–378.
- [65] Papageorgiou, A. & Paskov, S. (1999). Deterministic simulation for risk management, *Journal of Portfolio Management* **25**(5), 122–127.
- [66] Paskov, S.H. (1997). New methodologies for valuing derivatives, in *Mathematics of Derivative Securities*, M.A.H. Dempster & S.R. Pliska, eds, Cambridge University Press, Cambridge, pp. 545–582.
- [67] Paskov, S.H. & Traub, J.F. (1995). Faster valuation of financial derivatives, *Journal of Portfolio Management* **22**(1), 113–120.
- [68] Pirsic, G. (2002). A software implementation of Niederreiter-Xing sequences, in *Monte Carlo and Quasi-Monte Carlo Methods 2000*, K.T. Fang, F.J. Hickernell & H. Niederreiter, eds, Springer, Berlin, pp. 434–445.
- [69] Ross, R. (1998). Good point methods for computing prices and sensitivities of multi-asset European style options, *Applied Mathematical Finance* **5**, 83–106.
- [70] Sloan, I.H. (2002). QMC integration—beating intractability by weighting the coordinate directions, in *Monte Carlo and Quasi-Monte Carlo Methods 2000*, K.T. Fang, F.J. Hickernell & H. Niederreiter, eds, Springer, Berlin, pp. 103–123.
- [71] Sloan, I.H. & Joe, S. (1994). *Lattice Methods for Multiple Integration*, Oxford University Press, Oxford.
- [72] Sloan, I.H., Kuo, F.Y. & Joe, S. (2002). Constructing randomly shifted lattice rules in weighted Sobolev spaces, *SIAM Journal on Numerical Analysis* **40**, 1650–1665.
- [73] Sloan, I.H., Kuo, F.Y. & Joe, S. (2002). On the step-by-step construction of quasi-Monte Carlo integration rules that achieve strong tractability error bounds in weighted Sobolev spaces, *Mathematics of Computation* **71**, 1609–1640.
- [74] Sloan, I.H., Wang, X.Q. & Woźniakowski, H. (2004). Finite-order weights imply tractability of multivariate integration, *Journal of Complexity* **20**, 46–74.
- [75] Sloan, I.H. & Woźniakowski, H. (1998). When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? *Journal of Complexity* **14**, 1–33.
- [76] Sobol’, I.M. (1967). Distribution of points in a cube and approximate evaluation of integrals, *USSR Computational Mathematics and Mathematical Physics* **7**(4), 86–112.
- [77] Tan, K.S. & Boyle, P.P. (2000). Applications of randomized low discrepancy sequences to the valuation of complex securities, *Journal of Economic Dynamics and Control* **24**, 1747–1782.
- [78] Tezuka, S. (1998). Financial applications of Monte Carlo and quasi-Monte Carlo methods, in *Random and Quasi-Random Point Sets*, P. Hellekalek & G. Larcher, eds, Springer, New York, pp. 303–332.
- [79] Wang, X.Q. (2002). A constructive approach to strong tractability using quasi-Monte Carlo algorithms, *Journal of Complexity* **18**, 683–701.
- [80] Wang, X.Q. (2003). Strong tractability of multivariate integration using quasi-Monte Carlo algorithms, *Mathematics of Computation* **72**, 823–838.
- [81] Wang, X.Q. & Sloan, I.H. (2005). Why are high-dimensional finance problems often of low effective dimension? *SIAM Journal on Scientific Computing* **27**, 159–183.
- [82] Woźniakowski, H. (2000). Efficiency of quasi-Monte Carlo algorithms for high dimensional integrals, in *Monte Carlo and Quasi-Monte Carlo Methods 1998*, H. Niederreiter & J. Spanier, eds, Springer, Berlin, pp. 114–136.
- [83] Xing, C.P. & Niederreiter, H. (1995). A construction of low-discrepancy sequences using global function fields, *Acta Arithmetica* **73**, 87–102.

HARALD NIEDERREITER

Rare-event Simulation

In finance, the need for evaluation of the probability $\mathbb{P}(A)$ that an event is rare (in the sense that $\mathbb{P}(A)$ is small) arises, for example, in pricing out-of-the-money options and **Value-at-Risk** (VaR) calculations in **credit risk** and **operational risk**. However, similar problems arise in a large variety of application areas: insurance risk (ruin probabilities and probabilities of large accumulated claims); communications engineering (probabilities of blocking, bit errors, packet loss, etc.); reliability (probabilities of unavailability in steady state or in a given time interval); and so on.

In all of these examples, explicit evaluation of $\mathbb{P}(A)$ is seldom possible unless unrealistically simple model assumptions have been made. **Monte Carlo simulation** is therefore one of the main tools in this area. However, the fact that $\mathbb{P}(A)$ is small raises some specific problems.

By an estimator Z for $z = \mathbb{P}(A)$, we refer to a random variable (rv) Z that can be generated by simulation and satisfies $\mathbb{E}Z = z$. In practice, one simulates R (say) independent copies of Z and gives the estimate of z as the average $\bar{Z}_R$, supplemented with a confidence interval, say, an equitailed 95% confidence interval $\bar{Z}_R \pm w_R$ where s^2 is the empirical variance of the R simulated values of Z (the natural estimator of $\sigma_Z^2 = \text{Var} Z$) and $w_R = 1.96s/R^{1/2}$ is the half-width. In most other Monte Carlo contexts, a small w_R is the relevant criterion for an estimator to be efficient. However, in a rare-event setting, one is rather interested in the relative half-width $w_R/\mathbb{P}(A)$. For example, the confidence interval $10^{-5} \pm 10^{-4}$ may look narrow, but it does not help to tell whether z is of the magnitude 10^{-4} , 10^{-5} , or even much smaller. Another way to illustrate the problem is in terms of the sample size R needed to acquire a given relative precision, say 10%, in terms of the half-width of the 95% confidence interval. This leads to the equation $1.96 \sigma_Z / (z\sqrt{R}) = 0.1$, that is,

$$R = \frac{100 \cdot 1.96^2 z(1-z)}{z^2} \sim \frac{100 \cdot 1.96^2}{z} \quad (1)$$

which increases like z^{-1} as $z \downarrow 0$. Thus, if z is small, large sample sizes are required.

The standard formal setup for discussing such efficiency concepts is to consider a family $\{A(x)\}$

where $x \in (0, \infty)$ or $x \in \mathbb{N}$, assume that $z(x) \stackrel{\text{def}}{=} \mathbb{P}(A(x)) \rightarrow 0$ as $x \rightarrow \infty$, and for each x let $Z(x)$ be an unbiased estimator for $z(x)$, that is, $\mathbb{E}Z(x) = z(x)$. An algorithm is defined as a family $\{Z(x)\}$ of such rv's.

The best performance that has been observed in realistic rare-event settings is *bounded relative error* as $x \rightarrow \infty$, meaning

$$\limsup_{x \rightarrow \infty} \frac{\text{Var} Z(x)}{z(x)^2} < \infty \quad (2)$$

In particular, such an algorithm will have the feature that R as computed in equation (1), with $z(1-z)$ replaced by $\text{Var} Z(x)$, remains bounded as $x \rightarrow \infty$.

An efficiency concept slightly weaker than equation (2) is *logarithmic efficiency*: $\text{Var}(Z(x)) \rightarrow 0$ so quickly that

$$\limsup_{x \rightarrow \infty} \frac{\text{Var} Z(x)}{z(x)^{2-\varepsilon}} = 0 \quad (3)$$

for all $\varepsilon > 0$, or, equivalently, that

$$\liminf_{x \rightarrow \infty} \frac{|\log \text{Var} Z(x)|}{|\log z(x)^2|} \geq 1 \quad (4)$$

The most established method for producing estimators with such efficiency properties is *importance sampling*, where one simulates not from the physical measure $\mathbb{P}$ but another probability measure $\tilde{\mathbb{P}}$ with changed parameters. Then $Z = I(A)W$ where $W = d\tilde{\mathbb{P}}/d\mathbb{P}$ is the likelihood ratio. Here are two standard examples:

Example 1 $A = I(S_n > nm)$ where $S_n = X_1 + \dots + X_n$ with $X_1, X_2, \dots$ independent and identically distributed (i.i.d.) with density f and $m > \mathbb{E}X$. Here one defines $\tilde{f}(y) = e^{\theta y} f(y) / \mathbb{E}e^{\theta X}$ with θ chosen such that $\int y \tilde{f}(y) dy = m$. The estimator becomes

$$Z(n) = I(S_n > nm) \prod_{k=1}^n \frac{f(X_k)}{\tilde{f}(X_k)} \quad (5)$$

and is logarithmically efficient as $n \rightarrow \infty$.

Example 2 $A = I(\max S_n > x)$ with $X_1, X_2, \dots$ as in (1) and $\mathbb{E}X < 0$. Here one defines $\tilde{f}(y) = e^{\theta y} f(y)$ with $\theta > 0$ chosen such that $\int \tilde{f}(y) dy = 1$

(this θ is always unique and will exist for most standard light-tailed distributions). The estimator becomes $Z(x) = \exp\{-S_{\tau(x)}\}$ where $\tau(x) = \inf\{n : S_n > x\}$ and has bounded relative error as $x \rightarrow \infty$.

The definition of bounded relative error implies that in Example 2, the variance, compared to crude Monte Carlo, is reduced by a factor of order $z(x) = \mathbb{P}(A(x))$. Similar remarks apply to Example 1, though the variance reduction may be slightly smaller since there is only logarithmical efficiency there.

The principle behind the choice of $\tilde{\mathbb{P}}$ in these two examples (exponential change of measure) is to take $\tilde{\mathbb{P}}$ as an asymptotic conditional distribution given the rare event A . This is motivated from the exact conditional distribution giving zero variance. Often, the asymptotics of the conditional distribution is found by using the theory of large deviations, as illustrated by the following example:

Example 3 For a digital **barrier option**, the problem that arises (after some rewriting) is of estimating

$$z = \mathbb{P}(\underline{W}(T) \leq -a, W(T) \geq b),$$

$$\text{where } \underline{W}(T) \stackrel{\text{def}}{=} \inf_{t \leq T} W(t) \quad (6)$$

and W is a Brownian motion with drift μ and volatility σ , denoted as $\text{BM}(\mu, \sigma^2)$. If a, b are not too close to 0, this is a rare-event problem: if $\mu < 0$, W is unlikely to go from $-a$ to b , and if $\mu > 0$, W is unlikely to ever hit $-a$. It seems reasonable that on a fluid scale, the most likely path φ^* should be piecewise linear as in Figure 1.

The theory of large deviations suggests that finding φ^* is equivalent to a one-dimensional minimization problem, where one looks for the infimum over t of

$$2 \int_0^T I(\varphi'(t)) dt = \int_0^t (-a/t - \mu)^2 ds$$

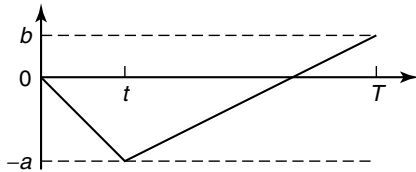


Figure 1 The most likely path for the barrier option

$$+ \int_t^T (c/(T-t) - \mu)^2 ds$$

$$= t(a/t + \mu)^2 + (T-t)$$

$$\times (c/(T-t) - \mu)^2 \quad (7)$$

where $c = b + a$, we have taken $\sigma^2 = 1$, and $I(y) = (y - \mu)^2/2$ is the so-called rate function for $\text{BM}(\mu, 1)$ (see **Large Deviations**). Elementary calculus shows that the minimum is attained at $t = aT/(a + c)$. This means that the most likely path is linear with slope $-\mu^*$ on $[0, t]$ and slope μ^* on $(t, T]$ where $\mu^* = (a + 2b)/T$, so that the above principle of approximating the conditional distribution given the rare event suggests that importance sampling may be done with the Brownian drift changed as stated.

This suggests simulating with drift $-\mu^*$ until the time τ where $-a$ is hit and then changing the drift to μ^* . Since the likelihood ratio for changing the drift from μ to $\tilde{\mu}$ in an interval $[0, t]$ is

$$Q = Q(t, \tilde{\mu}) = \exp\{-(\tilde{\mu} - \mu)W(t) + (\tilde{\mu} - \mu)^2 t\} \quad (8)$$

and this formula extends to stopping times, it follows that the importance sampling estimator is

$$Q(\tau, -\mu^*)Q((T - \tau), \mu^*)$$

$$\times I(\underline{W}(T) \leq -a, W(T) \geq b) \quad (9)$$

(modified to 0 if $\tau > T$). For a further discussion, see [4, pp. 264 ff].

Example 4 Glasserman *et al.* [5] considered a portfolio exposed to d (dependent) risk factors $X_1, \dots, X_d$ in a certain time horizon h , say, 1 day or 2 weeks. The initial value of the portfolio is denoted by v and the (discounted) terminal value by $V = V(h, X_1, \dots, X_d)$. The loss is $L = v - V$, and the VaR is a quantile of L (say the 99% one or the 99.97% one). It is assumed that $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$.

The basis of the algorithm is to invoke the *delta-gamma approximation*, which is based on the Taylor expansion

$$L \approx -\frac{\partial V}{\partial h} h - \sum_{i=1}^d \delta_i X_i - \frac{1}{2} \sum_{i,j=1}^d \gamma_{ij} X_i X_j \quad (10)$$

where $\delta_i \stackrel{\text{def}}{=} \partial V / \partial x_i$, $\gamma_{ij} \stackrel{\text{def}}{=} \partial^2 V / \partial x_i \partial x_j$. For brevity, we rewrite the right-hand side of equation (10) as $a_0 + Q$, where $Q \stackrel{\text{def}}{=} -\delta^T X - X^T \Gamma X / 2$, $a_0 \stackrel{\text{def}}{=} -h \partial V / \partial h$, and the proposal of Glasserman *et al.* [5] is to use the same exponential change of measure for the X_i as one would if the conditional distribution of Q , not L , was the target. Writing $X = CY$ where the components of $Y = (Y_1 \dots Y^d)$ are i.i.d. standard normal, one can choose C to satisfy

$$Q = \sum_{i=1}^d (b_i Y_i + \lambda_i Y_i^2) \quad (11)$$

and then it can be seen (e.g., [1] pp. 432–434) that under the conditional probability thus described, $Y_1, \dots, Y^d$ are still independent and Gaussian but have mean and variance parameters

$$\mu_i = \frac{\theta b_i}{1 - \theta \lambda_i}, \quad \sigma_i^2 = \frac{1}{1 - \theta \lambda_i} \quad (12)$$

where θ is determined by

$$x - a_0 = \tilde{\mathbb{E}} Q = \sum_{i=1}^d [b_i \mu_i + \lambda_i (\mu_i^2 + \sigma_i^2)] \quad (13)$$

Thus, the importance sampling estimator for estimating $\mathbb{P}(L > \ell)$ is

$$I(L > \ell) \prod_{i=1}^2 \frac{e^{-Y_i^2/2} / \sqrt{2\pi}}{e^{-(Y_i - \mu_i)^2/2\sigma_i^2} / \sqrt{2\pi\sigma_i^2}} \quad (14)$$

The empirical finding of Glasserman *et al.* [5] is that this procedure typically (i.e., in the case of a number of selected test portfolios) reduces the variance by a factor of 20–50.

The examples we have mentioned so far involve light tails. However, **heavy tails** are relevant particularly in areas such as credit risk and operational risk. The algorithms with heavy tails typically look completely different from those with light tails since exponential moments do not exist and hence exponential change of measure is impossible. We consider only the most important case of a tail $\bar{F}(x)$ that is regularly varying, that is, of the form $\bar{F}(x) = L(x)/x^{\alpha+1}$ with $L(\cdot)$ slowly varying (e.g., a Pareto tail). Statistical tests for distinguishing between light and heavy tails based on i.i.d. observations $X_1, \dots, X_n$ from F are discussed

in [6] (see also [1] VI.4). A popular tool is the mean excess plot, where the mean of the observations exceeding x is plotted as function of x . For a heavy-tailed distribution, and not for a light-tailed one, one expects to see a function going to infinity. The standard tool for estimating α in the regularly varying case is the so-called Hill estimator; see [1, 6].

Example 5 Let $X_1, \dots, X_n$ be i.i.d. with regularly varying distribution F and $S_N = X_1 + \dots + X_N$ where N is fixed or an independent rv. The problem of estimating $z(x) = \mathbb{P}(S_N > x)$ arises in a number of areas, for example, insurance risk, credit risk, and operational risk. The first efficient algorithms for this problem are remarkable in that they use conditional Monte Carlo, and not importance sampling. Currently, the most efficient of such algorithms uses the identity $z(x) = n \mathbb{P}(S_n > x, M_n = X_n)$ (keeping $N = n$ fixed and writing $M_k = \max(X_1, \dots, X_k)$); the intuition behind involving M_n is the fact that basically one X_i is large (and then equal to M_n) when $S_N > x$. One then simulates $X_1, \dots, X_{n-1}$ and returns the conditional expectation

$$\begin{aligned} Z(x) &= n \mathbb{P}(S > x, M_n = X_n \mid X_1, \dots, X_{n-1}) \\ &= n \bar{F}(\max(x - S_{n-1}, M_{n-1})) \end{aligned} \quad (15)$$

Example 6 A very active area is dynamic importance sampling, where the importance distribution varies with time and current state (e.g., in the barrier option example one would let the Brownian drift depend on both time t and $W(t)$). A principle for doing this is based upon Doob's h transform and requires that approximations for $z(x)$ are available; see [1] VI.7. Some of the important recent examples in the heavy-tailed area are found in [2] and [3], and the approach seems to carry greater hope for generality than conditional Monte Carlo.

We conclude by mentioning some ideas beyond standard importance sampling (as exemplified above) that are relevant in the general area of rare-event simulation. An interesting recent development is the cross-entropy method [7], which performs an automatic search for a good importance distribution within a given parametric class. Another development is that of splitting the methods (cf [1] V.6, VI.9), where the rare event is decomposed as the intersection of events, each of which is nonrare.

End Notes

^aNote that the parameter indexing the rare event is discrete in this example and hence denoted as n rather than x .

References

- [1] Asmussen, S. & Glynn, P.W. (2007). *Stochastic Simulation. Algorithms and Analysis*, Springer-Verlag.
- [2] Blanchet, J. & Glynn, P.W. (2008/09). Efficient rare-event simulation for the maximum of heavy-tailed random walks, *Annals of Applied Probability* **18**, 1351–1378.
- [3] Dupuis, P., Leder, K. & Wang, H. (2007). Importance sampling for sums of random variables with heavy tails, *ACM TOMACS* **17**, 1–21.
- [4] Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*, Springer-Verlag.
- [5] Glasserman, P., Heidelberger, P. & Shahabuddin, P. (2000). Variance reduction techniques for estimating value-at-risk, *Management Science* **46**, 1349–1364.
- [6] Resnick, S. (2007). *Heavy-Tailed Phenomena. Probabilistic and Statistical Modelling*, Springer-Verlag.
- [7] Rubinstein, R.Y. & Kroese, D.P. (2005). *The Cross-Entropy Method. A Unified Approach to Combinatorial Optimization, Simulation and Machine Learning*, Springer-Verlag.

Further Reading

- Asmussen, S. & Rubinstein, R.Y. (1995). Steady-state rare events simulation in queueing models and its complexity properties, in *Advances in Queueing: Models, Methods and Problems*, J. Dshalalow, ed., CRC Press, pp. 429–466.
- Heidelberger, P. (1995). Fast simulation of rare events in queueing and reliability models, *ACM TOMACS* **6**, 43–85.
- Juneja, S. & Shahabuddin, P. (2006). Rare event simulation techniques, in *Simulation*, S.G. Henderson & B.L. Nelson, eds, Handbooks in Operations Research and Management Science, Elsevier, pp. 291–350.

Related Articles

Barrier Options; Heavy Tails; Large Deviations; Monte Carlo Simulation; Operational Risk; Saddlepoint Approximation; Value-at-Risk; Variance Reduction.

SØREN ASMUSSEN

Stochastic Taylor Expansions

Deterministic Taylor Formula

As the concept of a stochastic Taylor expansion can be widely applied, we first review the *deterministic Taylor expansion*, using some terminology which will facilitate the presentation of its stochastic counterparts. Let us consider the solution $X = \{X_t, t \in [0, T]\}$ of the ordinary differential equation $dX_t = a(X_t) dt$, for $t \in [0, T]$ with initial value X_0 . We can write this equation in its integral form as

$$X_t = X_0 + \int_0^t a(X_s) ds \quad (1)$$

To justify the following calculations, we require the drift function a to be smooth and such that the solution of equation (1) does not explode. Then by using the deterministic chain rule, we can write

$$df(X_t) = a(X_t) \frac{\partial}{\partial x} f(X_t) dt \quad (2)$$

Using the operator $L = a \partial / \partial x$, we may express equation (2) as the integral equation

$$f(X_t) = f(X_0) + \int_0^t L f(X_s) ds \quad (3)$$

for all $t \in [0, T]$. Note for the special case $f(x) \equiv x$ we have $Lf = a$, $LLf = (L)^2 f = La, \dots$, and equation (3) reduces to equation (1). If we now apply relation (3) to the integrand $f = a$ in equation (1), then we obtain

$$\begin{aligned} X_t &= X_0 + \int_0^t \left(a(X_0) + \int_0^s L a(X_z) dz \right) ds \\ &= X_0 + a(X_0) \int_0^t ds + \int_0^t \int_0^s L a(X_z) dz ds \end{aligned} \quad (4)$$

which is the simplest nontrivial Taylor expansion for X_t . We can now apply equation (3) to the function

$f = La$ in the above double integral in equation (4). Consequently, we obtain

$$\begin{aligned} X_t &= X_0 + a(X_0) \int_0^t ds \\ &\quad + L a(X_0) \int_0^t \int_0^s dz ds + R_1 \end{aligned} \quad (5)$$

with remainder term

$$R_1 = \int_0^t \int_0^s \int_0^z (L)^2 a(X_u) du dz ds \quad (6)$$

for $t \in [0, T]$. Continuing this way then we get a version of the classical deterministic *Taylor formula*

$$\begin{aligned} f(X_t) &= f(X_0) + \sum_{l=1}^r \frac{t^l}{l!} (L)^l f(X_0) \\ &\quad + \int_0^t \cdots \int_0^{s_2} (L)^{r+1} f(X_{s_1}) ds_1 \cdots ds_{r+1} \end{aligned} \quad (7)$$

for $t \in [0, T]$ and $r \in \mathcal{N}$. In the sum on the right-hand side of equation (5), we find the expansion terms that are expanded at time zero or, more precisely, at the value X_0 . Furthermore, the last term, which is a multiple integral, represents the remainder term. Its integrand is, in general, not a constant. The deterministic Taylor formula allows the approximation of a sufficiently smooth function in a neighborhood of a given expansion point to any desired order of accuracy as long as f and a are sufficiently smooth.

Wagner–Platen Expansion

It is important to be able to approximate the increments of smooth functions of solutions of stochastic differential equations. For these tasks, a stochastic expansion, with analogous properties to the deterministic Taylor formula, is needed. The *Wagner–Platen expansion* (see [1, 2, 5, 8, 10, 11]) is such a stochastic Taylor expansion. We derive here one of its versions for a diffusion process $X = \{X_t, t \in [0, T]\}$ satisfying

$$X_t = X_0 + \int_0^t a(X_s) ds + \int_0^t b(X_s) dW_s \quad (8)$$

for $t \in [0, T]$, where W is a standard Wiener process. The coefficient functions a and b are assumed to be sufficiently smooth, real valued, and such that a unique solution of equation (6) exists. Then, for a sufficiently smooth function $f : \Re \rightarrow \Re$, the Itô formula provides the representation

$$\begin{aligned} f(X_t) &= f(X_0) + \int_0^t L^0 f(X_s) ds \\ &\quad + \int_0^t L^1 f(X_s) dW_s \end{aligned} \quad (9)$$

for $t \in [0, T]$, where we used the operators $L^0 = a(\partial/\partial x) + 1/2b^2(\partial^2/\partial x^2)$ and $L^1 = b(\partial/\partial x)$.

Obviously, for the special case $f(x) \equiv x$, we have $L^0 f = a$ and $L^1 f = b$, for which the representation (9) reduces to equation (8). Since the representation (9) involves integrands that are functions of processes, we can now apply the Itô formula (9) to the functions a and b in equation (8) to obtain

$$\begin{aligned} X_t &= X_0 + \int_0^t \left(a(X_0) + \int_0^s L^0 a(X_z) dz \right. \\ &\quad \left. + \int_0^s L^1 a(X_z) dW_z \right) ds + \int_0^t \left(b(X_0) \right. \\ &\quad \left. + \int_0^s L^0 b(X_z) dz + \int_0^s L^1 b(X_z) dW_z \right) dW_s \\ &= X_0 + a(X_0) \int_0^t ds + b(X_0) \int_0^t dW_s + R_2 \end{aligned} \quad (10)$$

with remainder term

$$\begin{aligned} R_2 &= \int_0^t \int_0^s L^0 a(X_z) dz ds \\ &\quad + \int_0^t \int_0^s L^1 a(X_z) dW_z ds \\ &\quad + \int_0^t \int_0^s L^0 b(X_z) dz dW_s \\ &\quad + \int_0^t \int_0^s L^1 b(X_z) dW_z dW_s \end{aligned} \quad (11)$$

This already represents a simple example of a Wagner–Platen expansion. We can extend the above expansion by applying the Itô formula (9), for

instance, to the function $L^1 b$ in the remainder. In this case, we obtain the expansion

$$\begin{aligned} X_t &= X_0 + a(X_0) \int_0^t ds + b(X_0) \int_0^t dW_s \\ &\quad + L^1 b(X_0) \int_0^t \int_0^s dW_z dW_s + R_3 \end{aligned} \quad (12)$$

with a new remainder R_3 . In equation (12), the leading terms are functions of the value of the diffusion at the expansion point, which are weighted by corresponding multiple stochastic integrals. In principle, one can derive such an expansion for a general multifactor diffusion process X , a general smooth function f , and an arbitrary high expansion level (see [6]). The main properties of this type of expansion are already apparent in the preceding example. The Wagner–Platen expansion can be interpreted as a generalization of both the Itô formula and the classical deterministic Taylor formula. It can be derived via an iterated application of the Itô formula.

The following version of a Wagner–Platen expansion, involving triple integrals in the expansion part, can be useful in various applications:

$$\begin{aligned} X_t &= X_0 + a I_{(0)} + b I_{(1)} + (a a' + 1/2 b^2 a'') I_{(0,0)} \\ &\quad + (a b' + 1/2 b^2 b'') I_{(0,1)} + b a' I_{(1,0)} + b b' I_{(1,1)} \\ &\quad + [a (a a'' + (a')^2 + b b' a'' + 1/2 b^2 a''') \\ &\quad + 1/2 b^2 (a a''' + 3 a' a'' + ((b')^2 + b b'') a'' \\ &\quad + 2 b b' a''') + 1/4 b^4 a^{(4)}] I_{(0,0,0)} \\ &\quad + [a (a' b' + a b'' + b b' b'' + 1/2 b^2 b''') \\ &\quad + 1/2 b^2 (a'' b' + 2 a' b'' + a b''' + ((b')^2 + b b'') b'' \\ &\quad + 2 b b' b''' + 1/2 b^2 b^{(4)})] I_{(0,0,1)} + [a (b' a' + b a'') \\ &\quad + 1/2 b^2 (b'' a' + 2 b' a'' + b a''')] I_{(0,1,0)} \\ &\quad + [a ((b')^2 + b b'') + 1/2 b^2 (b'' b' + 2 b b'' \\ &\quad + b b''')] I_{(0,1,1)} + b (a a'' + (a')^2 + b b' a'' \\ &\quad + 1/2 b^2 a''') I_{(1,0,0)} + b (a b'' + a' b' + b b' b'' \\ &\quad + 1/2 b^2 b''') I_{(1,0,1)} + b (a' b' + a'' b) I_{(1,1,0)} \\ &\quad + b ((b')^2 + b b'') I_{(1,1,1)} + R_6 \end{aligned} \quad (13)$$

Here, the coefficient functions a , b , and their derivatives a' , b' , a'' , b'' , a''' , b''' are valued at the

expansion point X_0 , which we suppress in our notation. The multiple stochastic integrals $I_{(j_1, j_2, j_3)} = \int_0^t \int_0^s \int_0^z dW_u^{j_1} dW_z^{j_2} dW_s^{j_3}$, where we set $dW_t^0 = dt$ and $dW_t^1 = dW_t$, are taken on $[0, t]$. Important applications of Wagner–Platen expansions arise in the construction of strong and weak discrete time approximations for scenario simulation (see **Stochastic Differential Equations: Scenario Simulation**) and Monte Carlo simulation (see **Monte Carlo Simulation for Stochastic Differential Equations**). Detailed results for higher level stochastic Taylor expansions and derivations of estimates for the remainder can be found in [6].

Generalized Wagner–Platen Expansions

By following the same ideas, one can expand the changes of a function value with respect to the underlying diffusion process X itself. For example, from an iterated application of the Itô formula it follows for a sufficiently smooth function $f : [0, T] \times \mathfrak{R} \rightarrow \mathfrak{R}$ an expansion of the form

$$\begin{aligned} f(t, X_t) = & f(0, X_0) + \frac{\partial}{\partial t} f(0, X_0) t \\ & + \frac{\partial}{\partial x} f(0, X_0) (X_t - X_0) \\ & + \frac{1}{2} \frac{\partial^2}{\partial x^2} f(0, X_0) \langle X \rangle_t \\ & + \frac{\partial^2}{\partial x \partial t} f(0, X_0) \int_0^t \int_0^s dX_z ds \\ & + \frac{1}{2} \frac{\partial^2}{\partial x^2} \frac{\partial}{\partial t} f(0, X_0) \int_0^t \int_0^s d\langle X \rangle_z ds \\ & + \frac{\partial^2}{\partial t \partial x} f(0, X_0) \int_0^t \int_0^s dz dX_s \\ & + \frac{\partial^2}{\partial x^2} f(0, X_0) \int_0^t \int_0^s dX_z dX_s \\ & + \frac{1}{2} \frac{\partial^3}{\partial x^3} f(0, X_0) \int_0^t \int_0^s d\langle X \rangle_z dX_s \\ & + \frac{1}{2} \frac{\partial^3}{\partial t \partial x^3} f(0, X_0) \int_0^t \int_0^s dz d\langle X \rangle_s \\ & + \frac{1}{2} \frac{\partial^3}{\partial x^3} f(0, X_0) \int_0^t \int_0^s dX_z d\langle X \rangle_s \end{aligned}$$

$$\begin{aligned} & + \frac{1}{4} \frac{\partial^4}{\partial x^4} f(0, X_0) \int_0^t \int_0^s d\langle X \rangle_z d\langle X \rangle_s \\ & + \bar{R}_f(0, t) \end{aligned} \quad (14)$$

for $t \in [0, T]$, where $\bar{R}_f(0, t)$ expresses the corresponding remainder term.

There exist multidimensional versions of Wagner–Platen expansions with respect to several driving processes. By using such expansions, one can, for instance, expand the increment of an option price in a multifactor setting. This provides a better understanding of the sensitivities with respect to given factor processes. Another application is the approximate evaluation of risk measures, for instance, **Value-at-Risk** (see [12]).

General stochastic Taylor expansions in a semimartingale setting have been derived in [8, 10]. Stochastic Taylor expansions based on multiple Stratonovich integrals are detailed in [5]. Wagner–Platen expansions for jump-diffusions and pure jump processes can be found in [3, 4, 9]. Expansions of functionals of Lévy processes *via* power processes have been described in [7].

References

- [1] Azencott, R. (1982). Stochastic Taylor formula and asymptotic expansion of Feynman integrals, *Séminaire de Probabilités XVI, Supplement, Lecture Notes in Mathematics*, Vol. 921, Springer, pp. 237–285.
- [2] BenArous, G. (1989). Flots et series de Taylor stochastiques, *Probability Theory Related Fields* **81**, 29–77.
- [3] Bruti-Liberati, N. & Platen, E. (2007). Strong approximations of stochastic differential equations with jumps, *Journal of Computational and Applied Mathematics* **205**(2), 982–1001.
- [4] Engel, D. (1982). The multiple stochastic integral, *Memiors of the American Mathematical Society* **38**, 265.
- [5] Kloeden, P.E. & Platen, E. (1991). Stratonovich and Itô stochastic Taylor expansions, *Mathematische Nachrichten* **151**, 33–50.
- [6] Kloeden, P.E. & Platen, E. (1999). *Numerical Solution of Stochastic Differential Equations*, Applied Mathematics, Vol. 23, Springer (Third Printing).
- [7] Nualart, D. & Schoutens, W. (2000). Chaotic and predictable representations for Lévy processes, *Stochastic Processes and Their Applications* **90**, 109–122.
- [8] Platen, E. (1981). A Taylor formula for semimartingales solving a stochastic differential equation, *Stochastic Differential Systems, Lecture Notes in Control and Information Sciences*, Vol. 36, Springer, pp. 157–164.

- [9] Platen, E. (1982). An approximation method for a class of Itô processes with jump component, *Lietuvos Matematikos Rinkiny* **22**(2), 124–136.
- [10] Platen, E. (1982). A generalized Taylor formula for solutions of stochastic differential equations, *Sankhya A* **44**(2), 163–172.
- [11] Platen, E. & Wagner, W. (1982). On a Taylor formula for a class of Itô processes, *Probability and Mathematical Statistics* **3**(1), 37–51.
- [12] Schoutens, W. & Studer, M. (2003). Short term risk management using stochastic Taylor expansions under Lévy models, *Insurance: Mathematics and Economics* **33**(1), 173–188.

Related Articles

Monte Carlo Simulation for Stochastic Differential Equations; Stochastic Differential Equations: Scenario Simulation; Stochastic Differential Equations with Jumps: Simulation.

ECKHARD PLATEN

Exercise Boundary Optimization Methods

One of the most important characteristics of the Monte Carlo simulation method is its intrinsically forward-looking nature. While this feature enables us to take into account inherently path-dependent payoff structures of any derivative contract with comparative ease, it makes it difficult to accommodate the inclusion of any early exercise rights into the contract. The two most commonly used methods to handle products with early exercise opportunities within a Monte Carlo simulation framework are *regression-based* techniques (which are covered in **Bermudan Options**) and *exercise boundary optimization* approaches that are discussed in this article.

Optimal Stopping Time as Exercise Boundary Optimization

For the sake of generality, we assume that we are dealing with the valuation of a financial product Π_0 , with embedded exercise optionality and discrete cash flows that span m time horizons $t_1 < \dots < t_m$, with the current time being $t \equiv t_0$. The product is considered to have two underlying contracts A and B of contingent cash flows, both of which, individually, contain no exercise optionality. The product Π_0 initially pays the same cash flows as product A , but, in addition, permits the exercise option holder to switch into B at one of the time horizons t_j . This formulation is fairly generic and encompasses practically all callable structures, including what is sometimes referred to as *options on options* or *higher order options* since they can be reformulated as a sequence of payable cash flows to continue until a final contingent payoff is attained, or to opt out at any of the intermediate exercise times.

Assuming a finite-dimensional Markovian representation of our usual probability space given a state vector $\mathbf{x}$, the contingency of the cash flows a_j in product A means that $a_j = a_j(\mathbf{x}(t_j))$, and likewise for product B . Risk-neutral valuation of the exercisable financial product Π_0 requires that we identify the optimal exercise strategy represented by the *optimal stopping time*. Valuation of Π_0 in the filtration

$\mathcal{F}_{t_0}$, denoted by $\mathcal{V}_{t_0}[\Pi_0]$, is given by

$$\mathcal{V}_{t_0}[\Pi_0] = N(\mathbf{x}(t_0)) \cdot \sup_{\tau} \mathbb{E}_{t_0}^{\mathcal{M}(N)} \left[\sum_{j=1}^{\tau-1} \frac{a_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} + \sum_{j=\tau}^m \frac{b_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} \right] \quad (1)$$

wherein the discrete random variable $\tau \in 1, \dots, m$ is the stopping time index, $N(\cdot)$ is the chosen numéraire, and $\mathbb{E}_{t_0}^{\mathcal{M}(N)}[f]$ is the expectation of any given f under the measure induced by the numéraire in the filtration $\mathcal{F}_{t_0}$. Define Π_k in complete analogy to the financial product Π_0 except that it can only be exercised on or after t_k . A crucial observation in the following is that if we had knowledge of the optimal stopping time process

$$\tau_k^* := \arg \sup_{\tau | \tau \geq k} \mathbb{E}_{t_0}^{\mathcal{M}(N)} \left[\sum_{j=1}^{\tau-1} \frac{a_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} + \sum_{j=\tau}^m \frac{b_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} \right] \quad (2)$$

we could value Π_k by means of the simple expectation

$$\mathcal{V}_{t_0}[\Pi_k] = N(\mathbf{x}(t_0)) \cdot \mathbb{E}_{t_0}^{\mathcal{M}(N)} \left[\sum_{j=1}^{k-1} \frac{a_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} + \sum_{j=k}^{\tau_k^*-1} \frac{a_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} + \sum_{j=\tau_k^*}^m \frac{b_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} \right] \quad (3)$$

By virtue of the assumed Markovian representation in the state vector $\mathbf{x}(t)$, it is possible to rephrase the valuation problem based on an indicator function $\mathcal{I}_k(\mathbf{x}(t_k))$, which takes the value 1 when exercise is optimal and 0 otherwise. This gives us the recursive formulation

$$\begin{aligned} \mathcal{V}_{t_0}[\Pi_k] &= N(\mathbf{x}(t_0)) \cdot \mathbb{E}_{t_0}^{\mathcal{M}(N)} \left[\mathcal{I}_k(\mathbf{x}(t_k)) \cdot \left(\sum_{j=1}^{k-1} \frac{a_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} + \sum_{j=k}^m \frac{b_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} \right) + \tilde{\mathcal{I}}_k(\mathbf{x}(t_k)) \cdot \frac{\mathcal{V}_{t_k}[\Pi_{k+1}]}{N(\mathbf{x}(t_k))} \right] \quad (4) \end{aligned}$$

2 Exercise Boundary Optimization Methods

wherein $\tilde{\mathcal{I}}_k = 1 - \mathcal{I}_k$. If we define the product $\hat{\Pi}_k$ as Π_k minus all cash flows occurring before t_k such that the first cash flow of $\hat{\Pi}_k$ is b_k or a_k at t_k , depending on whether exercise was invoked at t_k or not, we have

$$\begin{aligned} \mathcal{V}_{t_0}[\hat{\Pi}_k] &= N(\mathbf{x}(t_0)) \cdot \mathbb{E}_{t_0}^{\mathcal{M}(N)} \left[\tilde{\mathcal{I}}_k(\mathbf{x}(t_k)) \right. \\ &\quad \times \frac{a_k(\mathbf{x}(t_k)) + \mathcal{V}_{t_k}[\hat{\Pi}_{k+1}]}{N(\mathbf{x}(t_k))} \\ &\quad \left. + \mathcal{I}_k(\mathbf{x}(t_k)) \cdot \sum_{j=k}^m \frac{b_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} \right] \end{aligned} \quad (5)$$

In this form, it seems that we need to know the conditional value $\mathcal{V}_{t_k}[\hat{\Pi}_{k+1}]$ in order to value Π_0 (which is, by construction, equal to $\hat{\Pi}_0$, since no cash flows occur before t_1). However, since this conditional expectation appears *inside an expectation over its conditioning filtration*, by virtue of the *tower law*, we can replace it by the sequence of (numéraire-deflated) contingent cash flow values:

$$\begin{aligned} \mathcal{V}_{t_0}[\hat{\Pi}_k] &= N(\mathbf{x}(t_0)) \cdot \mathbb{E}_{t_0}^{\mathcal{M}(N)} \left[\tilde{\mathcal{I}}_k(\mathbf{x}(t_k)) \cdot \left(\sum_{j=k}^{\tau_{k+1}^* - 1} \frac{a_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} \right. \right. \\ &\quad \left. \left. + \sum_{j=\tau_{k+1}^*}^m \frac{b_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} \right) + \mathcal{I}_k(\mathbf{x}(t_k)) \cdot \sum_{j=k}^m \frac{b_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} \right] \end{aligned} \quad (6)$$

In this form, we can view

$$c_k = \sum_{j=k}^{\tau_{k+1}^* - 1} \frac{a_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} + \sum_{j=\tau_{k+1}^*}^m \frac{b_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} \quad (7)$$

as the numéraire-deflated *continuation value*, and

$$q_k = \sum_{j=k}^m \frac{b_j(\mathbf{x}(t_j))}{N(\mathbf{x}(t_j))} \quad (8)$$

as the numéraire-deflated *cancellation value*, to arrive at

$$\begin{aligned} \mathcal{V}_{t_0}[\hat{\Pi}_k] &= N(\mathbf{x}(t_0)) \cdot \mathbb{E}_{t_0}^{\mathcal{M}(N)} \left[\tilde{\mathcal{I}}_k(\mathbf{x}(t_k)) \cdot c_k \right. \\ &\quad \left. + \mathcal{I}_k(\mathbf{x}(t_k)) \cdot q_k \right] \end{aligned} \quad (9)$$

The only trouble now is that we do not know, *a priori*, the optimal exercise indicator function $\mathcal{I}_k(\cdot)$. We can confidently say, though, that given any trial exercise decision function $\mathcal{E}_k(\mathbf{x}; \lambda_k)$ (which we choose to indicate exercise if its value is positive), with parameter vector λ_k , we have

$$\begin{aligned} \mathcal{V}_{t_0}[\hat{\Pi}_k] &\geq N(\mathbf{x}(t_0)) \cdot \max_{\lambda_k} \mathbb{E}_{t_0}^{\mathcal{M}(N)} \left[\mathbf{1}_{\{\mathcal{E}_k(\mathbf{x}(t_k); \lambda_k) \leq 0\}} \cdot c_k \right. \\ &\quad \left. + \mathbf{1}_{\{\mathcal{E}_k(\mathbf{x}(t_k); \lambda_k) > 0\}} \cdot q_k \right] \end{aligned} \quad (10)$$

The key idea of *exercise boundary optimization methods* for the valuation of exercisable financial products in a Monte Carlo simulation framework is to use a suitably chosen exercise decision function $\mathcal{E}_k(\mathbf{x}; \lambda_k)$ and to find the value for λ_k that maximizes the Monte Carlo estimator objective function

$$F_k := \frac{1}{n} \sum_{i=1}^n (\mathbf{1}_{\{\mathcal{E}_k(\mathbf{x}(t_k); \lambda_k) \leq 0\}} \cdot c_{ik} + \mathbf{1}_{\{\mathcal{E}_k(\mathbf{x}(t_k); \lambda_k) > 0\}} \cdot q_{ik}) \quad (11)$$

with c_{ik} and q_{ik} representing the continuation and cancellation values of the i th path at time horizon t_k for a set of n -simulated discrete evolutions of the state vector in the chosen measure. Incidentally, it becomes apparent in this formulation that the exercise boundary optimization method allows readily for the switch product B to be, in principle, a different one at each exercise time horizon since the switch payments enter only in the form of the term q_{ik} as the sum of all numéraire-deflated cash flows that are payable if exercise is made at time t_k . Once the maximizing value λ_k has been established, the procedure continues at t_{k-1} , in analogy to a backward induction method. An important observation in this context is that throughout the entire backward iteration the same set of simulated paths and associated values can be used. This is a key point for the efficiency of this method. An initial simulation only needs to store the

required state variables $\mathbf{x}(t_k)$ and associated continuation and cancellation values for each exercise time horizon. All subsequent optimization calculations can then be done over this precomputed *training set*. For typical training set sizes in the range of 8191–65 535 paths, the evaluation of the objective function (11) is thus very fast indeed.

Choices of Exercise Boundary Specification

The choice of exercise decision function $\mathcal{E}(\mathbf{x}; \lambda)$ is crucial for the performance of exercise boundary optimization methods. This applies to both the actual choice of the functional form and, with it, the number of free parameters, as well as the effective financial variables that are considered the primary arguments of $\mathcal{E}(\cdot)$.

Assessment by Related Financial Contracts

An intuitive approach to the choice of a reduced set of state variables is to monitor related financial contracts. Andersen [1] and Piterbarg [3] describe how swap rate levels and European swaptions values, even if only attainable in an approximate analytical fashion in any one given model, can be used to capture most of the callability value for Bermudan swaptions. They also show that for this family of financial contracts, a one-parameter choice for $\mathcal{E}_k(\mathbf{x}; \lambda)$ is often sufficient. This particularly holds when the used model is itself driven by a single Brownian motion, even if there is no one-dimensional Markovian representation as is the case for Libor market model. In this case, the exercise decision function can be as simple as

$$SR(\mathbf{x}) - \lambda$$

with $SR(\mathbf{x})$ denoting the coterminal swap rate, for a payer's Bermudan swaption. More complicated decision rules for Bermudan swaptions can be found in [1].

Financial Coordinate Transformations

It is not always intuitive and easy to find a related financial contract that can be used as an exercise indicator and whose value is attainable (semi)analytically.

Jäckel [2] discusses the exercise boundary optimization method in more general terms and suggests, when necessary, the use of tree methods and non-linear transformations of $\mathbf{x}$ for the assessment of the suitability of any particular functional form and choice of variables prior to its use in an exercise boundary optimization context. The useful observation is made that a functional form that works well for a given financial contract's exercise domain delineation with one model, even if the contract is simplified to a significant extent, practically always also works very well with other models for the fully fledged product version. This makes it possible, for instance, to visualize exercise domains computed with a nonrecombining tree implementation of a two-factor Libor market model for a short contract life such as 6-noncall-2, and to apply the same functional form with a fully factorized model for very long-dated contracts.

A Useful Generic Functional Form

In practice, the functional form

$$\begin{aligned} \mathcal{E}_{\text{hyperbolic}}(x, y; a, b, c, d, g) \\ = a - y + c(x - b) + g\sqrt{(c(x - b))^2 + d^2} \end{aligned} \quad (12)$$

suits many practical applications where two financial variables are required to unlock most of the callability value.

The typical shapes of its zero-level contour line (which is the exercise boundary) are shown in Figure 1. As an example for the use of this functional form, consider the callability of a payer's Bermudan swaption in a multifactor model, and consider x to represent the front Libor and y as the coterminal swap rate. In the limit of both x and y very large, it is clearly beneficial to exercise as soon as possible, whence $\mathcal{E}$ ought to be large in this limit. In contrast, when both are very low, we should not exercise, and $\mathcal{E}$ must be negative in this limit. When x is small and y is moderate but larger than the fixed rate, exercise should not be done now but is likely to become optimal at a later stage. When y is very small, exercise should be avoided, even if x is large. This simple analysis already suggests that $-\mathcal{E}_{\text{hyperbolic}}$ with $a > 0$,

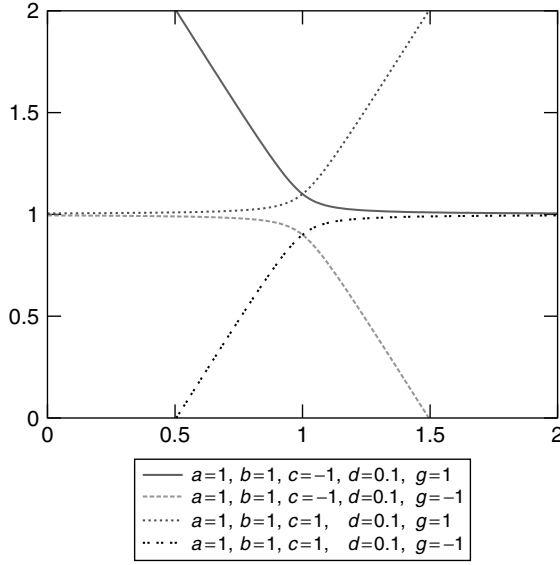


Figure 1 The hyperbolic exercise boundary given by equation (12)

$b > 0$, $c < 0$, $g > 0$ might be a good choice, and empirical tests show that this indeed works very well.

The Full Algorithm

The key stages of the exercise boundary optimization method are as follows:-

- Step 1.* Decide on a functional form for the exercise decision functions $\mathcal{E}$ for all exercise time horizons. Note that different functions may need to be chosen for different time horizons if the product exhibits strong inhomogeneity in its features over time. Also, note that the exercise domain may not be singly connected whence the implicit formulation of the exercise domain in the form of $\mathcal{E}(\mathbf{x}) > 0$ is generally preferable over explicit specifications of the boundary. A simple example for this is a multicallable best-of option paying $(\max(S_1, S_2) - K)_+$.
- Step 2.* Generation of the n -path training set. The only values that need to be stored are each path's continuation values, cancellation values, and exercise decision function argument values for each exercise decision horizon.

Note that for complex models, with contemporary computer's typical memory capacities, this reduction in storage requirements is typically necessary in order to be able to store all data in memory. Also note that the reduction of required memory typically leads to significant speedup since the cache memory access speed and main memory access speed differ considerably.

- Step 3.* In reverse chronological order, optimize the discretely sampled objective function (11) for each exercise time horizon in turn. Note that the objective function, at a high-resolution level, appears to be piecewise constant in its parameters whence an optimization method ought to be used that can cope with the fact that the function appears to change only at scales compatible with the granularity of the Monte Carlo sampling. One of the simplest methods that allows for a scale change during the optimization is the *Downhill-Simplex* algorithm [4]. For the case that λ is one-dimensional, golden section search [4] or outright sorting also works well.

- Step 4.* Using the exercise strategy now defined by the fully specified exercise decision functions established in stage 3, reevaluate the callable financial contract by an independent N -path Monte Carlo simulation with $N \gg n$. In practice, $N \sim 4 \cdot n$ has been found to be a good ratio when low-discrepancy numbers are used throughout.

The final result, by virtue of the inequality (10), is of course biased low since the valuation based on an optimized (implicit) functional approximation can only be as good as the exercise domain boundary is represented in the approximation. It can be shown readily, though, that for a small difference ε (defined in any suitable way) between the truly optimal exercise boundary and the one used in the numerical approximation, the difference between the numerically obtained value and the truly optimal value scales like the second order in ε , that is, like $\mathcal{O}(\varepsilon^2)$ whence small differences tend to have negligible influence on the calculation. Another mitigating factor with respect to the exercise boundary representation is that any mismatches only contribute proportionally to the probability of actually reaching this part of state space, that is, the exercise boundary only need be

matched well where probability densities are high. It is these features of second-order-only exercise domain-matching error propagation and the fact that the boundary only needs to be represented accurately in the center of the state vector distribution that makes this method so effective in practice.

References

- [1] Andersen, L. (2000). A simple approach to the pricing of Bermudan swaptions in the multifactor LIBOR market model, *Journal of Computational Finance* 3(2), 5–32.
- [2] Jäckel, P. (2002). *Monte Carlo Methods in Finance*, John Wiley & Sons.

- [3] Piterbarg, V. (2003). Computing deltas of callable Libor exotics in forward Libor models, *Journal of Computational Finance* 7(3), ssrn.com/abstract=396180.
- [4] Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical Recipes in C*, Cambridge University Press, www.library.cornell.edu/nr/cbookcpdf.html

Related Articles

Bermudan Options; Bermudan Swaptions and Callable Libor Exotics; Early Exercise Options: Upper Bounds; Finite Difference Methods for Early Exercise Options; LIBOR Market Models: Simulation; Stochastic Mesh Method.

PETER JÄCKEL & LEIF B.G. ANDERSEN

Early Exercise Options: Upper Bounds

Setup and Basic Results

We work, as usual, on a filtered probability space and consider a contingent claim with *early exercise* rights, that is, the right to accelerate payment on the claim at will. Let the claim in question be characterized by an adapted, nonnegative payout process $U(t)$, payable to the option holder at a stopping time (or *exercise policy*) $\tau \leq T$, chosen by the holder. If early exercise can take place at any time in some interval, we say that the derivative security is an *American option*; if exercise can only take place on a discrete set of dates, we say that it is a *Bermudan option*.

Let the allowed set of exercise dates larger than or equal to t be denoted $\mathcal{D}(t)$, and suppose that we are given at time 0 a particular exercise policy τ taking values in $\mathcal{D}(0)$, as well as a pricing numeraire N inducing a unique martingale measure $\mathbb{Q}^N$. Let $C^\tau(0)$ be the time 0 value of a derivative security that pays $U(\tau)$. Under technical conditions on $U(t)$, we can write the value of the derivative security as

$$C^\tau(0) = \mathbb{E}^N \left(\frac{U(\tau)}{N(\tau)} \right) \quad (1)$$

where $\mathbb{E}^N(\cdot)$ denotes expectation in measure $\mathbb{Q}^N$ and where we have assumed, with no loss of generality, that $N(0) = 1$. Let $\mathcal{T}(t)$ be the time t set of (future) stopping times taking value in $\mathcal{D}(t)$. In the absence of arbitrage, the time 0 value $C(0)$ of a security with early exercise into U is then given by the *optimal stopping problem*

$$C(0) = \sup_{\tau \in \mathcal{T}(0)} C^\tau(0) = \sup_{\tau \in \mathcal{T}(0)} \mathbb{E}^N \left(\frac{U(\tau)}{N(\tau)} \right) \quad (2)$$

reflecting the fact that a rational investor would choose an exercise policy to optimize the value of his/her claim.

With $\mathbb{E}_t^N(\cdot)$ denoting expectation conditional on the information (i.e., the filtration) at time t , we can extend equation (2) to future times t

$$C(t) = N(t) \sup_{\tau \in \mathcal{T}(t)} \mathbb{E}_t^N \left(\frac{U(\tau)}{N(\tau)} \right) \quad (3)$$

where $\sup_{\tau} \mathbb{E}_t^N (U(\tau)/N(\tau))$ is known as the *Snell envelope* of U/N under $\mathbb{Q}^N$. Here $C(t)$ must be interpreted as the value of the option with early exercise, *conditional* on exercise not having taken place before time t . To make this explicit, let $\tau^* \in \mathcal{T}(0)$ be the optimal exercise policy, as seen from time 0. We can then write, for $0 < t \leq T$,

$$C(0) = \mathbb{E}^N (1_{\{\tau^* \geq t\}} C(t)/N(t)) + \mathbb{E}^N (1_{\{\tau^* < t\}} U(\tau^*)/N(\tau^*)) \quad (4)$$

where we break the time 0 value into two components: one from the time t value of the option, should it not have been exercised before time t , and other from the right to exercise on $[0, t]$. As we can always elect —possibly suboptimally— to never exercise on $[0, t]$, from equation (4) we see that

$$C(0) \geq \mathbb{E}^N (C(t)/N(t)) \quad (5)$$

which establishes that $C(t)/N(t)$ is a *supermartingale* under $\mathbb{Q}^N$. This result also follows directly from known properties of the Snell envelope; see [13].

In numerical implementations, it is most relevant to consider the discrete-time (i.e., Bermudan) case and assume that $\mathcal{D}(0) = \{T_1, T_2, \dots, T_B\}$, where $T_1 \geq 0$ and $T_B = T$. For $t \in (T_i, T_{i+1})$, define H_i as the time t value of the Bermudan option when exercise is restricted to the dates $\mathcal{D}(T_{i+1}) = \{T_{i+1}, T_{i+2}, \dots, T_B\}$; that is,

$$H_i(t) = N(t) \mathbb{E}_t^N (C(T_{i+1})/N(T_{i+1})), \quad i = 1, \dots, B-1 \quad (6)$$

At time T_i , $H_i(T_i)$ can be interpreted as the *holding value* of the Bermudan option, that is, the value of the Bermudan option if not exercised at time T_i . If an optimal exercise policy is followed, clearly we must have at time T_i

$$C(T_i) = \max (U(T_i), H_i(T_i)), \quad i = 1, \dots, B \quad (7)$$

such that

$$H_i(t) = N(t) \mathbb{E}_t^N (\max (U(T_{i+1}), H_{i+1}(T_{i+1}))), \quad i = 1, \dots, B-1 \quad (8)$$

Starting with the terminal condition $H_B(T) = 0$, equation (8) defines an backward iteration in time for the value $C(0) = H_0(0)$.

Option Pricing Bounds

In a setting where $U(t)$ is a function of a low-dimensional diffusion process, the iteration (8) can often be solved numerically by partial differential equations (PDE) or lattice methods, for example, the finite difference method (*see Finite Difference Methods for Early Exercise Options*). In many cases of practical interest, however, these methods either do not apply or are computationally infeasible. In such situations, we may be interested in at least bounding the value of an option with early exercise rights. Providing a *lower bound* is straightforward: postulate an exercise policy τ and compute the price $C^\tau(0)$ by direct methods, for example, the Monte Carlo method. From equation (2), this clearly provides a lower bound

$$C^\tau(0) \leq C(0) \quad (9)$$

The closer the postulated exercise policy τ is to the optimal exercise policy τ^* , the tighter this bound will be. Two common strategies for approximation of τ^* in a Monte Carlo setting are discussed in **Bermudan Options** and **Exercise Boundary Optimization Methods**, the first based on regression estimates of holding values H in equation (8) and the second on optimization of parametric rules for the exercise strategy.

To produce an *upper bound*, we can rely on duality results established in [9, 14]. To present these results here, let $\mathcal{K}$ denote the space of adapted martingales π for which $\sup_{\tau \in [0, T]} E^N |\pi(\tau)| < \infty$. For a martingale $\pi \in \mathcal{K}$, we then write

$$\begin{aligned} C(0) &= \sup_{\tau \in \mathcal{T}(0)} E^N \left(\frac{U(\tau)}{N(\tau)} \right) \\ &= \sup_{\tau \in \mathcal{T}(0)} E^N \left(\frac{U(\tau)}{N(\tau)} + \pi(\tau) - \pi(\tau) \right) \\ &= \pi(0) + \sup_{\tau \in \mathcal{T}(0)} E^N \left(\frac{U(\tau)}{N(\tau)} - \pi(\tau) \right) \end{aligned} \quad (10)$$

In the second equality, we have relied on the *optional sampling theorem* to tell us that the martingale property is satisfied up to a bounded random stopping time, that is, that $E^N(\pi(\tau)) = \pi(0)$. See [12] for details. We now turn the above result into an upper

bound by forming a pathwise maximum at all possible future exercise dates $\mathcal{D}(0)$:

$$\begin{aligned} C(0) &= \pi(0) + \sup_{\tau \in \mathcal{T}(0)} E^N \left(\frac{U(\tau)}{N(\tau)} - \pi(\tau) \right) \\ &\leq \pi(0) + E^N \left(\max_{t \in \mathcal{D}(0)} \left(\frac{U(t)}{N(t)} - \pi(t) \right) \right) \end{aligned} \quad (11)$$

With equations (9) and (11) we have, as desired, established upper and lower bounds for values of options with early exercise rights. Let us consider how to make these bounds tight. As mentioned earlier, to tighten the lower bound we need to pick exercise strategies close to the optimal one. Tightening the upper bound is a bit more involved and requires usage of the *Doob–Meyer decomposition* (*see Doob–Meyer Decomposition*), which can be used here to show that

$$C(t)/N(t) = M(t) - A(t) \quad (12)$$

where $M(t)$ is a martingale and $A(t)$ an increasing, predictable process with $A(0) = 0$ (such that $C(0) = M(0)$). Given equation (12), consider taking $\pi(t) = M(t)$ in equation (11), to get

$$\begin{aligned} C(0) &\leq C(0) + E^N \left(\max_{t \in \mathcal{D}(0)} \frac{U(t)}{N(t)} - M(t) \right) \\ &= C(0) + E^N \left(\max_{t \in \mathcal{D}(0)} \frac{U(t)}{N(t)} - \frac{C(t)}{N(t)} - A(t) \right) \\ &\leq C(0). \end{aligned} \quad (13)$$

The last inequality follows from the fact that $C(t) \geq U(t)$ and $A(t) \geq 0$. As $M(0) = C(0)$, it follows that the quantity

$$\begin{aligned} M(0) + E^N \left(\max_{t \in \mathcal{D}(0)} \frac{U(t)}{N(t)} - M(t) \right), \\ M(t) = \frac{C(t)}{N(t)} + A(t) \end{aligned} \quad (14)$$

is bounded by $C(0)$ from both above and below, that is, it must equal $C(0)$. We have thereby arrived at a *dual* formulation of the option price

$$C(0) = \inf_{\pi \in \mathcal{K}} \left(\pi(0) + E^N \left(\max_{t \in \mathcal{D}(0)} \frac{U(t)}{N(t)} - \pi(t) \right) \right) \quad (15)$$

and have demonstrated that the infimum is attained when the martingale π is set equal to the martingale component M of the deflated price process $C(t)/N(t)$.

Monte Carlo Upper Bound Methods

Let us consider how we can use the upper bound results (11) and (15) in an actual Monte Carlo application. According to equation (11), to generate an upper bound for the true option price, it evidently suffices to simply pick *any* martingale process adapted to the filtration we work in, and then compute the expectation (11) by Monte Carlo methods. For instance, if the filtration is generated by a vector-valued Brownian motion $W(t)$, we can always set

$$\pi(t) = \int_0^t \sigma(t)^\top dW(t) \quad (16)$$

for some adapted vector-process $\sigma(t)$ satisfying the usual conditions required for the stochastic integral to be a proper martingale. Clearly, however, if $\sigma(t)$ is chosen arbitrarily, the resulting upper bound is likely to be very loose, and probably not very useful. While equation (15) is of little immediate practical use (since we do not know the process $C(t)/N(t)$), it does suggest that for a chosen martingale $\pi(t)$ in equation (11) to produce a tight upper bound, it needs to be “close” to $M(t)$.

Several strategies have been proposed for constructing a good martingale $\pi(t)$. When working on a simple model set up on simple payouts, sometimes one can make inspired guesses for what $\pi(t)$ should be. For instance, in a one-dimensional Black–Scholes model, Rogers [14] shows that using the numeraire-deflated European put option price (which is analytically known) as a guess for $\pi(t)$ generates good bounds for a Bermudan put option price. This approach, however, does not easily generalize to settings with more complicated dynamics and/or more complicated exercise payouts.

The Andersen–Broadie Algorithm

A general strategy for generating upper bounds is proposed in [1], which can start from any approximation to the optimal exercise strategy, perhaps

generated from either of the methods in **Bermudan Options** or **Exercise Boundary Optimization Methods**. Using a straightforward “simulation-within-a-simulation” approach, the authors construct an estimate to the value process $C^\tau(t)$ and use its estimated martingale component as $\pi(t)$ in equation (11). Specifically, working on a discrete timeline, they set

$$\begin{aligned} \pi(T_{i+1}) - \pi(T_i) &= \frac{C^\tau(T_{i+1})}{N(T_{i+1})} - E_{T_i}^N \left(\frac{C^\tau(T_{i+1})}{N(T_{i+1})} \right) \\ &= E_{T_{i+1}}^N \left(\frac{U(\tau)}{N(\tau)} \right) - E_{T_i}^N \left(\frac{U(\tau)}{N(\tau)} \mid \tau \geq T_{i+1} \right) \end{aligned} \quad (17)$$

where nested simulations are used to estimate both expectations on the right-hand side of the equation.^a The resulting Monte Carlo estimate of the upper bound is shown to be biased high always, with the bias being a decreasing function in the number of inner simulation trials. As suggested by equation (15), the upper bound produced by the algorithm in [1] strongly depends on the quality of the exercise strategy: the better the strategy, the tighter the bound.

The need for nested simulations makes the algorithm in [1] expensive: if M is the number of outer simulations and K the number of inner simulations, an option with B exercise opportunities will involve a worst case workload proportional to

$$M \cdot K \cdot B^2 \quad (18)$$

For comparison, a lower bound simulation has a workload proportional to $M \cdot B$, plus whatever work is required to estimate the exercise rule in a presimulation. In many cases the inner simulations can be stopped quickly (due to early exercise); thus, in practice the dependence on B in equation (18) is often less than quadratic and sometimes close to linear. In addition, in [3] it is shown—along with other ideas for efficiency improvements—that inner simulations are not needed on dates where it is suboptimal to exercise the option, which can lead to considerable time savings, especially for out-of-the-money options. Finally, K can often be set to a number much smaller than M without significantly affecting the quality of the upper bound, and even very small values of K (e.g., 50–100 or less) may yield informative results. With the computational improvements suggested in [3], upper bound computations on a range of different option payouts take on the order of 1–10 CPU minutes compared with 0.1–2 CPU minutes for the lower

bound. Of course, the CPU times depend on the speed of the processor, but it is safe to say that relatively tight confidence intervals for most option types can be obtained in times measured in minutes, not in hours as is commonly believed.

The strategy in [1] is generic, in that it can handle virtually any type of multidimensional process dynamics and security payouts. Despite the computational drawbacks, the use of nested simulation guarantees that the choice of π induces an upper bound estimate that is biased high. Importantly, this key property is *not* shared by many alternative estimators, such as regression, of the expectations in equation (17). One exception is discussed in [8] where a special martingale-preserving regression approach is introduced. This algorithm, however, requires strong conditions on regression basis functions that may be hard to check in practice.

The Belomestny–Bender–Schoenmakers Algorithm

In the special case where dynamics are driven only by Brownian motions, the usual martingale representation theorems show that the optimal strategy $\pi^*(t)$ must be an Ito integral, that is, of the form in equation (16). Starting again from a postulated exercise strategy, Belomestny *et al.* [2] use this observation to construct a regression on a set of basis functions to uncover an estimate for the function $\sigma(t)$. By applying regression techniques this way—rather than to directly compute expectations of $U(\tau)/N(\tau)$ —the authors are able to construct a true martingale process $\pi(t)$, which can be turned into a valid upper bound through equation (11). The resulting nonnested simulation algorithm requires careful implementation to yield stable results, in part, because the optimal integrand $\sigma(t)$ can be expected to be considerably less regular than $\pi^*(t)$ itself; this, in turn, requires additional thought in the selection of appropriate basis functions for the regression. One possibility advocated in [2] is to include, whenever available, exact or approximate expressions for the diffusion term in dynamics of several still-alive European options underlying the Bermudan option. This strategy is akin to that of [14], and its feasibility depends on the pricing problem at hand. In cases where it does apply, the authors of [2] demonstrate that their method gives good results, with the upper bound often being nearly as tight as that of the nested algorithm in [1]. They

also show how to use their technique to develop a variance-reduced version of the algorithm in [1].

Confidence Intervals and Practical Usage

Assume that we have estimated an exercise strategy using either of the approaches in **Bermudan Options** or **Exercise Boundary Optimization Methods**. Suppose that the Monte Carlo estimate for the lower bound price is $\hat{C}_{lo}(0)$ with a sample standard deviation of $\hat{s}_{lo}$ based on M_{lo} Monte Carlo trials. Using, say, the algorithm in [1], we also estimate an upper bound $\hat{C}_{hi}(0)$ with a sample standard deviation $\hat{s}_{hi}$ computed from M_{hi} (outer) simulation trials. With z_x denoting the x th percentile of a standard Gaussian distribution, asymptotically a $100(1 - \alpha)\%$ confidence interval for the true price $C(0)$ must be tighter^b than

$$\left[\hat{C}_{lo}(0) - z_{1-\alpha/2} \frac{\hat{s}_{lo}}{\sqrt{M_{lo}}}; \hat{C}_{hi}(0) - z_{1-\alpha/2} \frac{\hat{s}_{hi}}{\sqrt{M_{hi}}} \right] \quad (19)$$

Most often, upper bound simulation algorithms can be expected to be both more involved and/or more expensive than lower bound simulation methods. In many cases, the role of the upper bound simulation algorithm will therefore be to test whether postulated lower bound exercise strategies are tight or not. Specifically, starting from some guess for the exercise strategy, we can produce confidence intervals using equation (19) to test whether the lower bound estimate is of good quality, in which case the confidence interval can be made tight by using large values of M_{lo} and M_{hi} (as well as the number of inner simulation trials). In case the lower bound estimator is deemed unsatisfactory, we can iteratively refine it, by altering the choice of basis functions, say, until the confidence interval is tight. Importantly, such tests can often be done at a high level, covering entire classes of payouts and/or models. Once an exercise strategy has been validated for a particular product or model, day-to-day pricing of Bermudan securities can be done by the lower bound method, with only occasional runs of the upper bound method needed (e.g., if market conditions change markedly). If upper bound methods are predominantly used in this manner, the fact that they may sometimes be computationally intensive^c becomes less punitive.

Extensions and Related Work

The results (11) and (15) are sometimes known as *additive* duality results. Jamshidian [10] has introduced alternative *multiplicative* results. A comparative study of additive and multiplicative duality was undertaken in [7], with the authors concluding that the additive duality results are preferable in applications.

Earlier methods for producing lower and upper bounds were proposed in [5, 6]. Both methods [5, 6] have the significant feature of producing automatically convergent bounds. However, [5] is only practical when the number of exercise dates is a small finite number (e.g., less than five). The method proposed in [6] does not suffer this drawback, but is more challenging to implement and is substantially slower than most lower bounds methods.

Finally, let us note that computation of upper bounds by Monte Carlo simulation theoretically extends to option sensitivities (Greeks), as demonstrated in [11] where duality is applied to the likelihood ratio method (see [4]). As the proposed algorithm is very computationally intensive, the practical value of this result remains to be seen.

End Notes

^aIn cases where $U(t)$ is not known in closed form—as may be the case for complicated callable securities (see **Bermudan Swaptions and Callable Libor Exotics**)—nested simulation can also be used to establish estimates for $U(t)$.

^bThe confidence interval is conservative because of the low bias in $\hat{C}_{lo}(0)$ (i.e., $E^N(\hat{C}_{lo}) \leq C(0)$) and the high bias in $\hat{C}_{hi}(0)$ which originates in part from the nature of the upper bound, and in part from the earlier mentioned additional high bias introduced by the inner simulations.

^cNote, though, that in testing the viability of a class of exercise rules through an upper bound simulation, it is often acceptable to work with a reduced set of exercise opportunities—for example change a quarterly exercise schedule to an annual one—in order to save computation time (see equation (18)).

References

- [1] Andersen, L. & Broadie, M. (2004). A primal-dual simulation algorithm for pricing multi-dimensional American options, *Management Science* **50**, 1222–1234.
- [2] Belomestny, D., Bender, C. & Schoenmakers, J. (2009). True upper bounds for Bermudan products via non-nested Monte Carlo, *Mathematical Finance*, **19**(1), 53–71.
- [3] Broadie, M. & Cao, M. (2008). Improved lower and upper bound algorithms for pricing American options by simulation, *Quantitative Finance* **8**, 845–861.
- [4] Broadie, M. & Glasserman, P. (1996). Estimating security price derivatives using simulation, *Management Science* **42**, 269–285.
- [5] Broadie, M. & Glasserman, P. (1997). Pricing American-style securities using simulation, *Journal of Economic Dynamics and Control* **21**, 1323–1352.
- [6] Broadie, M. & Glasserman, P. (2004). A stochastic mesh method for pricing high-dimensional American options, *Journal of Computational Finance* **7**, 35–72.
- [7] Chen, N. & Glasserman, P. (2007). Additive and multiplicative duals for American option pricing, *Finance and Stochastics* **11**, 153–179.
- [8] Glasserman, P. & Yu, B. (2005). Pricing American options by simulation: regression now or regression later? in *Monte Carlo and Quasi-Monte Carlo Methods*, H. Niederreiter, ed, Springer Verlag.
- [9] Haugh, M. & Kogan, L. (2004). Pricing American options: a duality approach, *Operations Research* **52**, 258–270.
- [10] Jamshidian, F. (2006). The duality of optimal exercise and domineering claims: a Doob-Meyer decomposition approach to the Snell envelope, *Stochastics: an International Journal of Probability and Stochastic Processes* **79**, 27–60.
- [11] Kaniel, R., Tompaidis, S. & Zemlianov, A. (2008). Efficient computation of hedging parameters for discretely exercisable options, *Operations Research* **56**, 811–826.
- [12] Karatzas, I. & Shreve, S. (1991). *Brownian Motion and Stochastic Calculus*, 2nd Edition, Springer Verlag.
- [13] Lamberton, D. & Lapeyre, B. (2007). *Introduction to Stochastic Calculus Applied to Finance*, 2nd Edition, CRC Press.
- [14] Rogers, L.C.G. (2001). Monte Carlo valuation of American options, *Mathematical Finance* **12**, 271–286.

Related Articles

American Options; Bermudan Options; Bermudan Swaptions and Callable Libor Exotics; Doob–Meyer Decomposition; Exercise Boundary Optimization Methods; Finite Difference Methods for Early Exercise Options.

LEIF B.G. ANDERSEN & MARK BROADIE

Monte Carlo Simulation

The History of the Monte Carlo Method

The history of Monte Carlo methods goes back a long time. The generic idea of random, or “stochastic”, sampling is straightforward and appealing in its elegance and has been used for centuries. Possibly, the first systematic application of statistical sampling techniques in science and engineering was by Enrico Fermi in the early 1930s to predict the results of experiments related to the properties of the neutron [11], which had recently been discovered by James Chadwick in 1932. In 1947, Stanislaw Ulam suggested to John von Neumann that the newly developed ENIAC computer would give them the means to carry out calculations based on statistical sampling with hitherto unattained efficiency and comparative ease [17]. Their coworker Nicholas Metropolis dubbed the numerical technique the *Monte Carlo method* partly inspired by Ulam’s anecdotes of his gambling uncle who “just had to go to Monte Carlo”.

Since the deployment of the ENIAC, which could do about 5000 additions or 400 multiplications per second and which occupied the size of a large room, computing power has grown dramatically. In the early 1970s, a computer design was introduced that had at its heart an electronic component first introduced in 1958, a so-called “integrated circuit”. All of a sudden, a computer’s *central processing unit* (CPU) shrank from the size of a domestic refrigerator to that of a fingernail. The number of transistors in a single integrated circuit kept growing at an almost constant exponential rate since then,^a and with it grew the computing power of the computer. In addition to that, miniaturization and the introduction of new materials allowed for equally dramatic increases in the clock speeds of computers. At the time of this writing, on a CPU that trades for £ 25 to retail customers, over 2 billion double-precision floating-point multiplications can be carried out per second,^b which means that the kind of hardware used these days as a wordprocessor can do in one second what used to take the ENIAC over two months.

It is no surprise, then, that by now the use of Monte Carlo methods has become ubiquitous in science, technology, and business. Simulation techniques are used in oil well exploration; stellar evolution; electronic chip design; reactor design; quantum

chromo dynamics; material sciences; physical chemistry; nanostructure, protein, and polymer research; operations research, for example, when designing the relationships and control mechanisms between raw materials input, manufacturing, and delivery; ground and air traffic control systems design; communication and computer system design and testing, for example, network theory; biomolecular research, for example, cancer drug design; all areas of finance and insurance; weather forecasting (where it is referred to as *ensemble forecasting*^c); and local authorities planning and commissioning.^d

Today, the comparative ease of implementation, in combination with the readily available required computer power, makes the use of Monte Carlo techniques more often than not the method of choice. This has gone to the extent where it is deployed (almost) as a black-box tool. As an example for this, we give a quote from the local authority planning and commissioning site mentioned above:

This spreadsheet based tool gives local authorities access to Monte Carlo statistical modelling techniques. The technique allows local authorities to take account of the different factors which may affect spend levels. This will give authorities the ability to more accurately estimate expenditure to take account of uncertainty.

Monte Carlo is a recognised statistical technique, which is recommended by the Treasury.

Basic Ideas

The key defining feature of a Monte Carlo simulation may be stated as follows.

Definition 1 *A Monte Carlo method is any technique whose purpose it is to approximate a specific measure defined on a given domain by the aid of sampling according to a predetermined distributional law.*

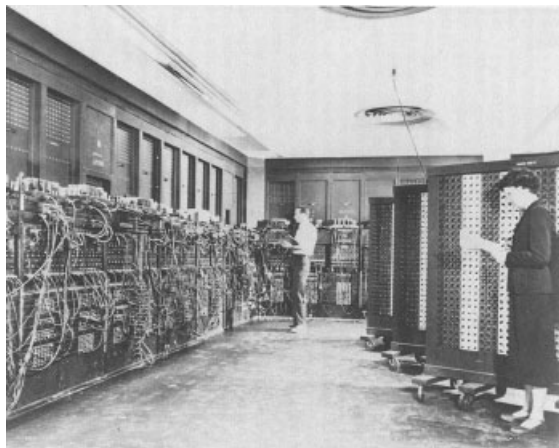
Note that this definition does *not* involve any of the following:

1. randomness
2. expectations or moments
3. stochastic processes.

It may come as a surprise that randomness is not included. It is true that randomness of some of the

input numbers is often associated with Monte Carlo methods. In truth, however, and this was already known to the pioneers of Monte Carlo methods, all that is really needed for a Monte Carlo method to succeed is *asymptotic adherence* to the desired predetermined distributional law on the given domain that is sampled upon. For the sake of explanation, but without loss of generality, we shall assume in the following that the domain is a hypercube $H_m = (0, 1)^m$ for some $m \in \mathbb{N}$ and that the desired distribution is uniform in H_m . Any “draw” in the simulation sequence is thus naturally best to be seen as an m -tuple, or a vector $\mathbf{u} \in H_m$.

Randomness is predominantly used in conceptual considerations for its convenience of guaranteeing uniform coverage as a consequence of *serial independence of draws* (Figure 1). Obviously, if one vector-valued sample drawn from H_m is, in a statistical sense, completely independent from the next draw, and so on, then, asymptotically, the set of all draws will cover H_m uniformly. However, it is both intuitive, as well as readily demonstrable, that it is fairly easy to achieve a more uniform coverage with a deliberate strategy that takes into account all the points that are already part of the simulation. It is, in fact, possible to prove that at least asymptotically (i.e., in the limit of large numbers of draws), by any self-consistent measure, any strategy (algorithm) to generate numbers that attempts to “avoid”



The ENIAC. Smithsonian Institution Photo No. 53192 reproduced from [4] with kind permission by Paul Ceruzzi

points already drawn gives more uniform coverage of the sampling domain than random numbers. In this asymptotic sense, among all number sequences one can construct, random numbers are the worst possible category of numbers one could use for Monte Carlo simulation purposes. In practical applications, though, a Monte Carlo simulation is hardly ever carried out with the intent to determine some kind of asymptotic behavior. What matters for real calculations is to have

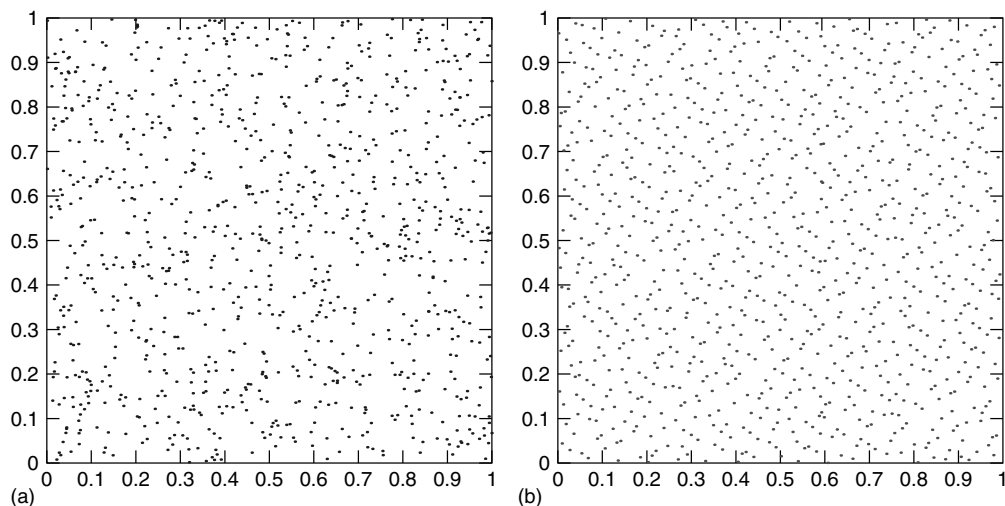


Figure 1 One thousand (a) random and (b) strategic draws from $(0, 1)^2$. Note the significantly more uniform coverage of the unit square in (b)

a result as accurate as possible with the fewest necessary iterations, and it is this latter part that makes the use of nonrandom numbers (also known as *low-discrepancy* or *quasi-random* numbers) slightly more delicate, and this is discussed in article **Quasi-Monte Carlo Methods** by one of the most accomplished researchers in this field, H. Niederreiter. Unfortunately, in the 1940s, no algorithms for the generation of usable nonrandom numbers were available when Monte Carlo simulations on the ENIAC commenced, despite the fact that the theoretical foundations for such number sequences had already been laid in 1916 by Hermann Weyl [18]. As a consequence, practically used Monte Carlo simulations relied on the generation of pseudorandom numbers for the first few decades. The term *pseudorandom* is commonly used to indicate that these are computer-generated numbers, as opposed to random numbers as we assume them to occur in natural phenomena such as radioactive decay. The distinction is somewhat philosophical, but noteworthy nevertheless, since it is, after all, not actually possible to generate randomness on a machine that is designed to give *deterministic* results. This is summarized in John von Neumann's famous words [9]:

Anyone who considers arithmetical methods of producing random numbers is, of course, in a state of sin.

As it turns out, algorithms for the generation of number sequences that can be considered *random enough* for practical use, in some sense, can be devised, though, by the aid of modern number theory, and this is the subject of article **Pseudorandom Number Generators** by P. L'Ecuyer.

At a mathematical level, Monte Carlo methods have a wide range of applicability. Of course, they can be used simply to compute an approximation for the expectation of a given function, say, $E[f(\mathbf{u})]$, and in finance, this is probably the most common application: draw N vectors $\mathbf{u}_1$ to $\mathbf{u}_N$, and evaluate the equally weighted N -iteration estimator:

$$\hat{E}_N[f(\mathbf{u})] := \frac{1}{N} \sum_{i=1}^N f(\mathbf{u}_i) \quad (1)$$

They can, however, also be used for calculations that do not fit well into the format $E[f(\mathbf{u})]$. A common use of Monte Carlo methods is to find a local or global extremum of a given function. Examples

for this application are the so-called *Metropolis* algorithm, and the related *simulated annealing* procedure, which, in finance, is sometimes used for model calibration. Another frequent application of Monte Carlo methods in finance is the calculation of a *quantile level*, that is, to find f_q such that the probability that $f(\mathbf{u}) < f_q$ is q , or, in other words, f_q is implicitly defined by:

$$E[\mathbf{1}_{\{f(\mathbf{u}) < f_q\}}] = q \quad (2)$$

The simplest algorithm to estimate f_q given q is to draw a number of, say N , iterations for $\mathbf{u}$, evaluate $f(\cdot)$ for all the drawn vectors, order the so generated f values by decreasing size, and pick the value at position $\lfloor q \cdot N \rfloor$. The quantile level is only one of various measures for *extreme events*, which are discussed in article **Rare-event Simulation**.

The Monte Carlo method is widely accepted as the designated method of choice for the following two types of calculations:

- **Mathematical problems that are not posed as a set of concise equations, but instead are only described as a procedure, or algorithm, or process.**

An example is the estimation of statistics for traffic-flow problems where the reaction of individual traffic constituents (e.g., cars) on ambient conditions is known, but, because of their nonlinearity, the net effect of many such constituents reacting to the (re-)actions of all other constituents is hard to predict. Some of the problems Fermi, von Neumann, Metropolis, and Ulam originally worked on also belong to this category: they investigated the statistics of neutrons passing through matter by simulating repeated impacts with and deflections by atoms. Some problems are given in the form of stochastic differential equations (SDEs) for which we have no analytical solutions but which can be integrated numerically. Numerical integration of an SDE is mathematically the design of a numerical algorithm that has starting values and an iterative procedure involving numbers to be sampled from a certain domain. Standard Euler integration, for instance, of a simple SDE, over k time steps, typically becomes a sampling problem on $\mathbb{R}^k$, which is usually transformed to $(0, 1)^k$. Note that the numerical integration is composed only by the step that is the conversion of the SDE into a sampling problem. Numerical integration of SDEs is technically not a Monte Carlo technique. It is, in principle, possible, though rarely done, to evaluate the

resulting sampling problem with other means such as multidimensional quadratures, and so on. In practice, however, numerical integration of SDEs is usually combined with a Monte Carlo simulation of the integration procedure. Articles **Monte Carlo Simulation for Stochastic Differential Equations** and **Stochastic Differential Equations: Scenario Simulation** by Platen are dedicated to the subject of numerical integration of SDEs, and how it interacts with the embedding Monte Carlo simulation. The mathematical background behind the design of many of these numerical integration schemes is the equivalent to Taylor's expansion for SDEs, and these *stochastic Taylor expansions* are explained in article **Stochastic Taylor Expansions**.

- **Mathematical problems that require the evaluation of an expectation on a high-dimensional domain.**

The alternative to a Monte Carlo simulation for such problems is to employ lattice methods. The commonly used error measure for lattice techniques is a term proportional to the square of the lattice discretization, say, h^2 . The number of nodes (i.e., sampling points) in a lattice scales like $N = h^{-d}$ with the dimensionality d . Putting this together shows that the error of a lattice algorithm relates to the number of evaluated points like $\sim 1/N^{(2/d)}$. As d increases, the relationship between a lattice method error and N becomes hopelessly inefficient, and this fact is sometimes referred to as the *curse of dimensionality*. When using random numbers, the commonly used error measure of a Monte Carlo estimator for N iterations scales like $1/\sqrt{N}$, irrespective of the dimensionality of the domain. A straightforward scale comparison indicates that the Monte Carlo error estimator is asymptotically superior in its behavior as a function of N as soon as $d > 4$. For low-discrepancy numbers, as explained in detail in article **Quasi-Monte Carlo Methods**, the error measure scales closer to $1/N$, which indicates that lattice methods are (asymptotically) superior to those (used within a Monte Carlo simulation) only when the dimensionality d is 1.

Beyond these two categories, there are many other applications of Monte Carlo techniques, often associated with bespoke Monte Carlo algorithms. Many of these algorithms belong to the family of *Monte Carlo Markov Chain* methods [3, 5, 10]. These methods are, to date, less commonly used in

finance. An exception is perhaps the Robbins–Monro algorithm [1] whose basic purpose is to find the root of a function that is defined as an expectation (and can practically be evaluated only by the aid of a Monte Carlo simulation itself). The Robbins–Monro algorithm is designed to avoid having to nest a Monte Carlo simulation as an objective function inside a numerical root finding procedure. It has been used in the context of variance reduction methods which are discussed in article **Variance Reduction**.

Monte Carlo Methods in Quantitative Finance

The starting point of the use of the Monte Carlo method in quantitative finance, or specifically, in financial engineering, was probably the suggestion by P. Boyle to use it for the valuation of options [2]. Despite the fact that Monte Carlo methods, initially, had the reputation of being *slow*, they started to become popular rather quickly, probably primarily owing to the enormous ease with which they could be implemented. For the valuation of exotic financial contracts, especially in the world of equity and foreign exchange derivatives, and particularly when multiple underlyings were involved, a Monte Carlo method could be implemented directly from the contract's term sheet, without the need for any further mathematics other than a model for the dynamics of the financial underlyings. This made it possible to combine multipurpose financial models, which only describe the (stochastic) dynamics of the financial underlyings, with *product description language* engines, that allowed any trader, or even structurer, to price, and risk-manage, entirely new financial products within minutes, without any new model development or any code recompilation. The use of Monte Carlo simulations is made particularly easy in this context by the fact that most financial contracts' term sheets are written in a language close to an investor's perception of the evolution of the financial contract *forward in time*, making term sheet definitions intrinsically well compatible with the Monte Carlo method's forward looking nature: *if you can describe it on a term sheet, you can probably implement it as a payoff*. Over time, computers became faster, reducing concerns some people may have had with respect to Monte Carlo methods. Eventually, multicore computers became commonplace in the financial industry,

and since Monte Carlo methods are so extremely easily amenable to multithreading, possibly more so than any other numerical technique, they could readily be adapted to take advantage of the extra computing power.

The adoption of Monte Carlo methods as a standard technique was helped not only by pure technological progress of computers, though. A whole range of mathematical developments and discoveries provided improvements both in terms of robustness and relative speed that made it possible to carry out Monte Carlo simulations much faster than before. Many of these methods are known as *variance reduction methods*, which is the subject of article **Variance Reduction**. A very important generic relative speedup technique is the use of low-discrepancy, as opposed to pseudorandom, numbers, such as the Niederreiter'88 [12], Niederreiter-Xing [13], and Sobol' [7, 14–16] numbers, and article **Quasi-Monte Carlo Methods** is dedicated to the theory behind these number sequences. We say *relative* speedup to indicate that the use of these numbers does not necessarily make the execution of N iterations actually faster (albeit that these numbers are usually faster to generate than pseudorandom numbers). They provide a relative speedup due to the fact that, for approximately the same residual numerical error, one may need only of the order $\sqrt{N}$ iterations in comparison to the use of pseudorandom numbers. Another type of speedup is associated with bespoke simulation algorithms for specific mathematical problems such as the numerical integration and simulation of square root processes described by SDEs such as

$$dX = \sigma\sqrt{X} \cdot dW \quad (3)$$

which is discussed in article **Simulation of Square-root Processes**. Simulations of numerical integrations of SDEs with jumps are described in article **Stochastic Differential Equations with Jumps: Simulation**. The simulations required for the important LIBOR market model for interest rate derivative valuation are particularly challenging because of their intrinsic high dimensionality, and are elaborated in article **LIBOR Market Models: Simulation**.

As Monte Carlo methods started being used with financial models, practitioners invented further techniques for specific purposes that are nowadays considered part of the standard toolbox. The handling

of sensitivity calculations, in the context of Monte Carlo simulations, can be tricky, owing to the inherent nature of a numerical sensitivity calculation to amplify noise. In the article **Monte Carlo Greeks**, the most common techniques to deal with this issue are covered. A mathematically elegant, though, in practice, perhaps less frequently used, framework that can help address the same issues is the method known as *integration by parts for stochastic functionals* and is presented in the article **Sensitivity Computations: Integration by Parts**. A particularly interesting development is the family of *weighted Monte Carlo* methods, which provide means for noise-reduced sensitivity calculation and noise-reduced model calibration. This framework gives a generic mathematical analysis that encompasses a proof that the popular *control variates* technique for variance reduction is, in fact, just a special case of a weighted Monte Carlo method, as discussed in article **Weighted Monte Carlo**.

Finally, we should mention that a lot has been done with respect to the handling of early exercise opportunities that may be given within financial contracts. Apart from simple one-dimensional American equity or FX options, these are very common among exotic fixed income derivatives. Since most of these exotic contracts require rather complex and high-dimensional models for realistic risk-management and hedging, it became desirable to find techniques that can be used to evaluate early exercise rights within a simulation setting. In this section of the encyclopedia, we present four articles on the most important techniques for Bermudan options: **Bermudan Options** deals with basis function regression methods for Bermudan options; in **Stochastic Mesh Method**, the Broadie–Glasserman stochastic mesh technique is discussed; article **Exercise Boundary Optimization Methods** explains Bermudan Monte Carlo methods based on exercise boundary optimization; and article **Early Exercise Options: Upper Bounds** is on upper bound methods for callable products.

There is much more one could say about Monte Carlo methods in finance, some of it highly applicable in many situations, some of it bespoke for specific applications, and some of it merely for the sake of its mathematical elegance. We believe, though, that all of the most important aspects of Monte Carlo methods in quantitative finance are covered in this section of the encyclopedia. For further details beyond this, we recommend the books [7, 8] and [6].

End Notes

^aThis phenomenon was surprisingly accurately predicted by Gordon Moore in 1965, whence it is often referred to as *Moore's law*.

^bThis figure is *per se*, of course, misleading since any realistic calculation involves more than the multiplication of the same two numbers over and over again. It does help, though, to highlight the scale of speed improvements in hardware since the days of the ENIAC.

^cAccording to the UK Meteorological Office (www.metoffice.gov.uk/science/creating/daysahead/ensembles), “an ensemble forecasting system samples the uncertainty inherent in weather prediction to provide more information about possible future weather conditions.” In other words, it is a mini Monte Carlo simulation consisting of a comparatively small number (~ 24) of individually deterministic weather forecast calculations, each primed with slightly differing input scenarios based on the currently known weather system status.

^dsee, e.g., www.everychildmatters.gov.uk/resources-and-practice/IG00215.

References

- [1] Arouna, B. (2003). Robbins-Monro algorithms and variance reduction in finance, *Journal of Computational Finance* **7**(2), 35–61.
- [2] Boyle, P. (1977). Options: a Monte Carlo approach, *Journal of Financial Economics* **4**, 323–338.
- [3] Bremaud, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*, Springer.
- [4] Ceruzzi, P.E. (1983). *Reckoners, the Prehistory of the Digital Computer, from Relays to the Stored Program Concept, 1935-1945*, Greenwood Press. ISBN 0-313-23382-9.
- [5] Gilks, W.R., Richardson, S. & Spiegelhalter, D. (eds) (1996). *Markov Chain Monte Carlo in Practice*, Chapman & Hall.
- [6] Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*, Springer.
- [7] Jäckel, P. (2002). *Monte Carlo Methods in Finance*, John Wiley & Sons.
- [8] Kloeden, P.E. & Platen, E. (1992,1995,1999). *Numerical Solution of Stochastic Differential Equations*, Springer.
- [9] Knuth, D. (1969,1981). *The Art of Computer Programming: Seminumerical Algorithms*, Addison-Wesley, Vol. 2.
- [10] Liu, J.S. (2001). *Monte Carlo Strategies in Scientific Computing*, Springer.
- [11] Metropolis, N. (1987). *The Beginning of the Monte Carlo Method*, Los Alamos Science, Special Issue 15.
- [12] Niederreiter, H. (1988). Low-discrepancy and low-dispersion sequences, *Journal of Number Theory* **30**, 51–70.
- [13] Niederreiter, H. & Xing, C.P. (1995). Low-discrepancy sequences obtained from algebraic function fields over finite fields, *Acta Arithmetica* **72**, 281–298.
- [14] Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical Recipes in C*, Cambridge University Press, www.library.cornell.edu/nr/cbookcpdf.html.
- [15] Sobol', I.M. (1967). On the distribution of points in a cube and the approximate evaluation of integrals, *USSR Computational Mathematics and Mathematical Physics* **7**, 86–112.
- [16] Sobol', I.M. (1976). Uniformly distributed sequences with an additional uniform property, *USSR Computational Mathematics and Mathematical Physics* **16**(5), 236–242.
- [17] Ulam, S.M., Richtmyer, R.D. & von Neumann, J. (1947). *Statistical Methods in Neutron Diffusion*. Technical report, Los Alamos Scientific Laboratory report LAMS-551.
- [18] Weyl, H. (1916). Über die Gleichverteilung von Zahlen mod. eins, *Mathematische Annalen* **77**, 313–352.

PETER JÄCKEL & ECKHARD PLATEN

Mutual Funds

After a major financial crisis in 1773–1774, a Dutch broker founded an investment company that would buy bonds from issuers located in different countries. (For a detailed history of mutual funds, see [1].) Shareholders received an annual report and could periodically check assets and company books. This first example of a mutual fund contains all the elements of the modern collective investment vehicle: a professional manager, a pool of assets, diversification, and governance.

While modern mutual funds can take several legal shapes, such as partnerships and limited-liability corporations, they all perform two important functions: (i) asset management carried out by an expert and (ii) diversified exposure to investment assets.

Mutual funds are beneficial for people who lack the time or expertise necessary to select individual securities. By buying shares in a mutual fund, investors can capitalize on a professional fund manager's expertise and experience. The manager is responsible for building a portfolio consistent with her mandate.

In addition to expert management, mutual funds give investors, of all skill levels, the benefit of higher diversification. For example, Person A has \$1000 to invest, and one share of Company B costs \$1000. To invest in Company B, Person A would have to invest his entire portfolio in that single stock. If Company B were to go bankrupt, the investor would lose everything. However, if Person A were to pool his money with 10 other investors with \$1000 each, they could purchase shares of several different companies. The resulting portfolio's returns would reflect the average return of several stocks rather than just one. Since stocks can move differently from one another (i.e., one stock may gain as another loses), this pooling will reduce the portfolio's risk overall. Mutual funds provide these benefits on a large scale (see **Diversification; Modern Portfolio Theory**).

Mutual funds manage tens of trillions of dollars worldwide. Morningstar, an investment analysis and rating company, had over 54 000 different mutual fund portfolios in its worldwide database in February 2008. There were 7054 open-end mutual funds in the Morningstar database for the United States, 287 of which (4%) were pure index funds. According to

Morningstar, in February 2008, index funds had about 10% of the US open-end mutual fund assets—not including “enhanced index” and similar hybrid strategies (the Morningstar data cited in this paragraph do not include money market funds).

Legal Types of Mutual Fund

Mutual funds have different legal forms. Open-end mutual funds are investment companies that sell a variable number of shares of a pool of assets. In this type of fund, when an investor wants to buy shares, she contacts the investment company, which issues new shares of the fund. If the investor needs to liquidate her position, the company liquidates some of the fund's assets and pays the investor the share's current market price, called *net asset value* (NAV).

Closed-end mutual funds are just like holding companies. The number of shares is predetermined, and shares are bought and sold on the stock market. No contact between issuer and investors occurs after the initial public offering.

Exchange-traded funds (ETFs) are a special type of closed-end funds (see **Exchange-traded Funds (ETFs)**). The peculiarity of ETFs is that their number of shares is variable, and can be increased when an institutional investor, called *authorized participant*, presents the issuer a bundle of assets replicating the ETF's portfolio. The bundle is exchanged for new shares of the ETF. Similarly, authorized participants can redeem ETF shares in kind, receiving portfolio assets from the issuer. As a result, ETF prices reflect the prices of the underlying assets exactly.

Since they are sold to the public, both open- and closed-end funds are regulated in most countries and have strict standards regarding their mandates and disclosures. In contrast, hedge funds are a kind of fund that is not available to most investors; therefore, it is generally unregulated. This allows hedge funds to have more flexible mandates. As a result, hedge funds often are both potentially lucrative and risky (see **Hedge Funds**).

Investment-based Classification

Mutual funds differ by investment mandate. The first type of mandate specifies whether the manager must strictly reproduce the returns of a given index (see **Electricity Markets**). As explained in [2], an

index is a market average. By definition, about half of the investors will have returns exceeding the market average and half will have returns that fall short. Therefore, a reasonable strategy is to pursue the market average. A fund that adopts this strategy is called an *index fund*. Index funds have several advantages. First, because the market is diversified, an index fund's portfolio will have limited idiosyncratic risk. Second, since replicating a market index is rather inexpensive, index funds' management fees are low. Third, in countries taxing realized capital gains, index funds trigger limited capital gains because indexes turn over their portfolios relatively rarely.

The opposite of an index (or passive) fund is an actively managed fund. In an actively managed fund, the index's returns act as a benchmark rather than a target. The fund manager aims to exceed the benchmark, and he/she is free to choose different assets within his/her mandate to meet this goal (*see Capital Asset Pricing Model*).

ETFs started as passive investments, that is, index-based products. More recently, ETFs have branched out into direct investment in commodities such as gold bullion as well as in indexes based on complex trading strategies. While this increases the choices for investors, it often increases fees, too.

Index funds and passive ETFs initially tracked very broad indexes to provide diversified market exposure at low cost. It is now possible to find index funds and ETFs tracking very specialized indexes such as single countries, sectors, or currencies. In this way, an investor can overweight a narrow segment of the market. With ETFs, it is possible to also short market sectors that the investor considers overvalued (*see Hedging*).

A second mandate classification categorizes funds in terms of the kinds of securities that the fund manager can purchase. Equity funds invest in stocks, bond funds invest in debt instruments, and money market funds invest in short-term bonds. In some countries, including the United States, money market funds have legal limits on the credit risk of the bonds they can purchase.

Hybrid funds, also called *balanced* or *asset allocation funds*, invest in a mix of stocks, bonds, and money market instruments. Hybrid funds are often intended to be the only fund in a portfolio, because they contain a balanced exposure to several asset classes.

Funds in each of these broad categories—equity, bond, money market, and hybrid—can be further subdivided according to their specialization. For example, a fund manager may invest exclusively in US large-cap equity, Australian Dollar money market instruments, or Euro Zone high-yield bonds; another fund manager may have a different mandate and run a conservative balanced fund. Management style (*see Style Analysis*) is another common asset-based categorization of equity portfolios.

Expense-based Classification

Open-end funds are called *no-load* if they only charge an annual management fee, which is a percentage of assets. A load is a commission charged either when the investor buys (front load) or sells (back load) the shares. Loads compensate brokers and management companies. In general, a front-load fund has lower annual management fees. Many back-load funds waive loads for investors who have kept their shares for several years.

Closed-end funds do not have loads, but they do have annual management fees. Also, they may have a bid–ask spread and brokerage commission just like a regular stock.

Hedge funds generally have a management fee as well as a performance fee. For example, a hedge fund charging “2 and 20” has an annual management fee of 2% of its assets and a performance fee of 20% of all profits exceeding a certain threshold (called the *hurdle*).

Performance Evaluation

Specialized companies, such as Morningstar and Lipper, collect information about mutual funds' returns, fees, investment styles, and costs. They also provide peer comparisons, which rely on proprietary measures (e.g., the Morningstar Rating™) or on risk-adjusted measures envisioned by academics (*see Sharpe Ratio; Performance Measures*). The purpose of these comparisons is to help investors select skilled managers and avoid those who do not add value.

Modern portfolio theory suggests that the riskier assets should, on average, have higher returns. Hence, it is appropriate to use risk-adjusted measures to avoid confusing managers who have higher long-term returns because they take more risk (*see Capital*

Asset Pricing Model) from managers who add value without adding risk.

References

- [1] Goetzmann, W.N. & Rouwenhorst, K.G., (eds) (2005). *The Origins of Value: The Financial Innovations that*

Created Modern Capital Markets, Oxford University Press, Oxford, pp. 249–270.

- [2] Malkiel, B.G. (2007). *A Random Walk Down Wall Street*, 9th Edition, Norton, New York.

MICHELE GAMBERA

Style Analysis

Asset allocation is a major determinant of the return variation in fund portfolios. Hence, an investor who delegates portfolio management to a fund manager is particularly interested in the exposure to different types of asset classes. These exposures characterize the *style* of a fund. Fund names and self-declared fund objectives stated in the prospectus are typically not very informative about the style of mutual funds. Moreover, funds often change their style over time (see, e.g., [2]). Investors therefore turn to alternative ways to infer fund style. Return-based style analysis, introduced by Nobel Prize laureate William F. Sharpe ([5, 6]), determines the style of a fund by regressing the fund returns against the returns of a set of benchmark indices. These benchmark indices typically include style indices for large and small capitalization stocks, value and growth stocks, and bond indices. The coefficients on the style benchmark indices measure the exposures to the risks of these asset classes. Holdings-based style analysis infers the style of a fund directly from the characteristics of the individual positions of the fund portfolio. Style analysis can help characterize the asset allocation, evaluate the performance, and monitor the style consistency of a fund.

Return-based Style Analysis

Methodology

Factor models are a common way in finance to decompose the forces that drive security returns into common and firm-specific influences. A factor model differs from a simple regression exercise in that the residual firm-specific influence for one asset is assumed to be uncorrelated with the residual of all other assets. Hence, the correlation of asset returns is due to their exposure to common factors. Return-based style analysis is a special case of a *linear factor model*. The factors are returns on different asset classes, such as returns on stocks of different style or bond-market indices, which are also called *style benchmark indices*. Owing to this property, return-based style analysis belongs to the category of *asset class factor models*. The return of a fund, $r_{i,t}$, is decomposed into two parts: the fraction that can be

explained by a linear combination of the factors, and a residual term that is uncorrelated with the factors. The residual of fund i , $e_{i,t}$, is uncorrelated with the residual of any other fund j , $e_{j,t}$. Sharpe's ([5, 6]) return-based style analysis differs in a few ways from a standard factor model: (i) There is no intercept; (ii) the loadings on asset classes sum up to 1 to provide an intuitive interpretation as allocations to the asset classes; (iii) in its original form, and when applied to mutual funds, the coefficients are required to be nonnegative to capture short-selling constraints. This last constraint (2) to nonnegative exposures can be relaxed for hedge funds (see, e.g., [4]). The coefficients $b_{i,n}$ of equation (1) are estimated by minimizing the variance of the error terms $e_{i,t}$ subject to these constraints.

$$r_{i,t} = b_{i,1}f_{1,t} + b_{i,2}f_{2,t} + \cdots + b_{i,N}f_{N,t} + e_{i,t} \quad (1)$$

$$\text{s.t. } b_{i,n} \geq 0 \text{ for } n = 1, 2, \dots, N \quad (2)$$

$$\sum_{n=1}^N b_{i,n} = 1 \quad (3)$$

The coefficients $b_{i,n}$ measure the exposure to the n th factor, $f_{n,t}$. These coefficients help the investor to identify the *effective style* of an active portfolio manager. Return-based style analysis extends easily to a multimanager portfolio.

The R^2 from the constrained regression (1–3) measures the fraction of the variation in fund returns that can be explained by the variation of the style benchmark indices.

$$R^2 = 1 - \frac{\text{Var}(e_i)}{\text{Var}(r_i)} \quad (4)$$

When the set of style benchmarks is accurately defined, the difference $(1 - R^2)$ serves as a measure of active management. Stock picking, sector bets, and timing the exposure to different style benchmark indices results in lower R^2 . To penalize the inclusion of additional style benchmarks and to favor a parsimonious specification, the adjusted R^2 or Akaike information criterion can be used as a maximization criterion instead of R^2 (see, e.g., [1]).

Selection of Style Benchmark Indices

The choice of asset classes determines the quality of any asset class factor model:

1. Sharpe [5] recommends using a set of asset classes that are (i) mutually exclusive, (ii) comprehensive; and (iii) where correlations between different asset classes are preferably low.
2. When return-based style analysis is used to evaluate fund performance, a few more properties are desirable: asset classes should represent strategies that can be implemented (i) at low cost and (ii) by a passive, value-weighted, and investable portfolio, that is, the strategy can be implemented *ex ante*.

As an example, average returns on peer groups violate this last property since the median fund manager is not known at the beginning of the evaluation period.

Common choices for equity mutual funds include style indices along the two dimensions size and value/growth. Value (growth) stocks are defined by high (low) book-to-market ratios. Chan *et al.* [3] show that the two dimensions size and book-to-market capture the essential style of mutual funds. The key is a parsimonious set of asset classes that explains the time-series variation of different fund returns, not necessarily their average returns. A larger number of asset classes increases coverage, but, at the same time, reduces the reliability of the estimated exposures.

Use as a Performance Measure

A style-consistent portfolio manager allocates constant fractions $b_{i,n}$ of his or her investments to a given asset class n . The portfolio manager's performance can thus be evaluated against a passive portfolio with the same fractions invested in the N benchmark indices. The difference

$$e_{i,t} = r_{i,t} - b_{i,1}f_{1,t} + b_{i,2}f_{2,t} + \cdots + b_{i,N}f_{N,t} \quad (5)$$

using the optimal weights for the factors, measures the *selection* skills of a manager. The part of the return that is explained by the manager's effective style, $b_{i,1}f_{1,t} + b_{i,2}f_{2,t} + \cdots + b_{i,N}f_{N,t}$, is called the *style benchmark return*. The standard deviation of the selection component $e_{i,t}$ measures the *style benchmark tracking error*.

In a return-based style analysis, the nonfactor residuals $e_{i,t}$ are determined in-sample. To evaluate the performance of a fund manager, we need to determine his or her style and then compare the fund's

out-of-sample return to a passive style benchmark. This requires estimating first the coefficients $\tilde{b}_{i,t-1}$ in equation (1) using the returns from $t - 1 - k$ through $t - 1$, where k is the length of the estimation window. Then, the fund's *selection return* for period t is measured by the difference

$$r_{i,t} - [\tilde{b}_{1,t-1}f_{1,t} + \tilde{b}_{2,t-1}f_{2,t} + \cdots + \tilde{b}_{N,t-1}f_{N,t}] \quad (6)$$

A common approach to monitor style changes of a fund is to apply a rolling window. Typical choices are a 36-month or 60-month window.

Holdings-based Style Analysis

A critique of return-based style analysis is that it does not necessarily uncover what the fund actually holds but rather how it behaves, as expressed by Sharpe's often quoted words "if it walks like a duck and quacks like a duck, it is a duck." Take a position in a hybrid security such as convertible bonds. It likely shows up simultaneously in the loadings of a bond and equity style benchmark index. Furthermore, return-based style analysis reflects the average style over the estimation window and for funds with style drift the coefficient estimates become noisy and less meaningful.

Alternatively, investors may scrutinize fund portfolio holdings. Prominent data providers in the mutual fund industry, such as Morningstar and Lipper, use portfolio holdings to classify funds along the two dimensions size and value/growth. A holdings-based style analysis provides a snapshot of the current asset allocation. Information on portfolio positions is more difficult to collect in a timely manner than that on returns on funds and style benchmark indices, it often lacks uniformity, and processing the information is not an easy task. Holdings may also be affected by window dressing, a behavior of fund managers to adjust their portfolio positions before disclosure at the end of a quarter in an attempt to mimic stock picking skills.

References

- [1] Ben Dor, A., Jagannathan, R. & Meier, I. (2003). Understanding mutual fund and hedge fund styles using return-based style analysis, *Journal of Investment Management* 1(1), 94–134.

-
- [2] Brown, S.J. & Goetzmann, W.N. (1997). Mutual fund styles, *Journal of Financial Economics* **43**(3), 373–399.
- [3] Chan, L.K.C., Chen, H.-L. & Lakonishok, J. (2002). On mutual fund investment styles, *Review of Financial Studies* **15**(5), 1407–1437.
- [4] Fung, W. & Hsieh, D.A. (1997). Empirical characteristics of dynamic trading strategies: the case of hedge funds, *Review of Financial Studies* **10**(2), 275–302.
- [5] Sharpe, W.F. (1988). Determining a fund’s effective asset mix, *Investment Management Review* **2**(6), 59–69.
- [6] Sharpe, W.F. (1992). Asset allocation: management style and performance measurement, *Journal of Portfolio Management* **18**(2), 7–19.

Related Articles

Capital Asset Pricing Model; Factor Models; Performance Measures.

IWAN MEIER

Hedge Funds

A *hedge fund* is a private investment fund that pools money from a limited number of qualified investors and may use leverage to take long and short positions in various financial instruments. The first hedge fund was created by Alfred Winslow Jones in 1949, which was designed to invest in equity long positions hedged through short positions.^a

Owing to flexible trading strategies, incentive-aligned structures, low correlations with traditional asset classes, and loose regulatory oversight, hedge funds have gained popularity in the past decade. In 2007, it has been estimated that there were more than 9000 hedge funds worldwide with assets of about \$2.0 trillion under management, compared with only \$39 billion in 1990.^b Initially, most investments in hedge funds were made by high-net-worth individuals. Of late, institutional investors are increasingly investing in hedge funds.

Structure

Hedge funds are normally organized as limited partnerships for onshore funds in a country like the United States. Under the limited partnership, the investment manager acts as the general partner while investors are limited partners; offshore hedge funds are organized as corporations without a limit on the number of investors.

Offshore hedge funds are often registered in offshore tax havens such as Cayman Island, the British Virgin Island, Bermuda, and Bahamas [31].

Fees

Hedge funds usually charge both asset management fees, and incentive fees. The management fee is typically 1–2% of assets under management, while the incentive fee is typically 20% of the profit. The incentive fee feature makes hedge funds different from mutual funds and offers better incentive alignment between the managers and investors (see [25]).

The minimum rate above which the incentive fee can be charged is called the *hurdle rate*. It could be zero, the T-Bill rate, London Interbank Offered Rate (LIBOR), or some fixed benchmark.

High Watermark

This is a provision under which a manager has to recoup the previous losses before the incentive fee can be charged, that is, the cumulative returns have to be above the hurdle rate in order to collect an incentive fee. Further, it is possible that the manager could “owe” the investors a rebate of fees charged in previous years.

Strategy

Hedge funds follow dynamic trading strategies that are different from buy-and-hold strategies:

- **Global macro**

Follow a “top-down” approach to analyze the world economy and financial markets, and invest in corresponding stock, bond, and currency markets. George Soros’ Quantum Fund is a typical global macro fund.

- **Convertible arbitrage**

Purchase the convertible securities and simultaneously short sell the same issuer’s common stock.

- **Merger arbitrage/risk arbitrage**

Short sell the acquirer’s stock and buy the target’s stock to make arbitrage profit.

- **Fixed income arbitrage**

Purchase undervalued fixed income securities and short sell overvalued fixed income securities.

- **Long/short equity**

Combine long and short positions in equity securities. It could be long biased, short biased, or equity market neutral.

- **Event driven**

Aim to profit from some special events like merger and acquisition, reorganizations, or distressed companies.

- **Equity market neutral**

Take long and short position in equities with the goal of having no net exposure to equity markets.

- **Multistrategy**

A single hedge fund that is managed by multiple managers each following a separate strategy.

• Fund of hedge funds

Invest in a basket of underlying hedge fund managers for the purpose of diversification and operational due diligence. Investors in a fund of hedge funds pay two layers of fees: one at the underlying hedge fund level and one at the fund of hedge funds level.

Minimum Investment

Since hedge funds are designed for accredited/qualified investors, the minimum initial investment can be very high. For example, Eton Park Capital Management LLC has an initial investment of \$5 million.^c

Lockup Period

It is a time period during which an investor's money cannot be redeemed. An advance notice period is typically accompanied by the lockup period for redemption. Hedge funds usually impose a shorter lockup period of less than two years, which is typically the minimum lockup period for private equity funds.

Leverage

In addition to investors' own money, a hedge fund can borrow additional money from prime brokers or other institutions for investment. This can increase both the return and risk. The exact amount of leverage used depends on the investment strategy. For example, at the end of 1997, Long Term Capital Management had a leverage ratio of 28 to 1 for its fixed income arbitrage fund.^d

US Regulation

Unlike heavily regulated vehicles like mutual funds, hedge funds are largely exempt from the US Securities and Exchange Commission (SEC) regulation (at the time of writing). Hedge funds issue securities in "private offerings" and are not required to be registered with the SEC under the Securities Act of 1933. Furthermore, hedge funds are not required to make periodic reports under the Securities Exchange Act of 1934.

Like mutual funds and other securities market participants, hedge funds are subject to prohibitions against fraud, and their managers have the same fiduciary duties as other investment advisers. Depending upon their activities, in addition to complying with the federal securities laws, hedge funds and their advisers may have to comply with other laws including the Commodity Exchange Act or register with the US Commodity Futures Trading Commission (CFTC).

Performance and Risk*Performance Measures*

Unlike traditional investment vehicles, hedge funds are not required to report their performance to any database or regulatory organization. This leads to the question of the accuracy of the reported figures and whether the resulting selection bias prevents investors from having a complete picture of hedge fund industry. To date, three databases have been collecting data for longer than 10 years. They are CISDM (Center for International Securities and Derivatives Markets, University of Massachusetts, Amherst), HFR (Hedge Fund Research), and Lipper-TASS.

Return Distribution

Table 1 presents certain statistical properties of the CISDM indices. First, hedge fund indices provide mean returns that are similar to those provided by traditional stock and bond indices but at lower volatility. Second, hedge fund indices tend to display large negative skewness and large positive excess kurtosis. These two indicate that hedge fund indices are likely to experience extreme negative returns. Hence, traditional measures such as standard deviation and Sharpe ratio may not be suitable for hedge funds. Alternative measures such as the downside deviation, VaR, expected short fall, and maximum drawdown have been proposed [30]. Finally, certain strategies display significant positive autocorrelation, which may indicate that hedge funds representing these strategies carry illiquid and infrequently traded securities.

Data Biases

Returns reported by the public databases are subject to the following biases [1, 16, 29]:

Table 1 Statistical properties of CISDM indices (January 1990–November 2007)

Index provider	Launch date	Base date	Index weighting	Number of funds in the database	Number of funds in the indices	Rebalancing frequency
Altvest/Morningstar	2000	1993	Equal weighted	+4000	+2000	Monthly
Barclay Group	2003	1997	Equal weighted	+4000	+2000	Monthly
CISDM	1994	1990	Median	+4000	+2000	Monthly
CS/Lipper/TASS	1999	1994	Value weighted	+4000	+500	Quarterly
Hennessee	1987	1987	Equal weighted	+4000	+700	Annual
HF Net	1998	1976–1995	Equal weighted	+4000	+3000	Continual
HFR	1994	1990	Equal weighted	+4000	+2000	Monthly
MSCI	2002	2002	Equal and value weighted	+2000	+1500	Quarterly
Van Hedge	1994	1988	Equal weighted	+5000	+1500	Monthly

• Selection bias

Bias resulting from managers reporting their returns on a voluntary basis. This bias arises if the hedge funds that report to databases are systematically different from the universe. The size of this bias is not known since the hedge fund universe is not reported. It is argued that better performing funds may be closed to new investors and therefore do not need to report to public databases. This indicates that reported returns underestimate the true returns. On the other hand, poor performing hedge funds may decide not to report to any database. In this case, the reported returns may overestimate returns to the hedge fund universe.

• Survivorship bias

Databases drop hedge funds that do not survive (i.e., stop reporting). Those that survive tend to have better performance.^c The bias can be measured as the average return of the survived funds in excess of the average return of all funds. It must be noted that not all hedge funds that stop reporting to a database are doing so because of poor performance.^f Some databases maintain records of those funds who decided to stop reporting to them. Using these databases, survivorship bias is estimated to be around 0.16–3.4% per year [1, 11, 16, 17, 29].

• Backfill bias

When a manager decides to report his or her performance to a database, typically after obtaining a reasonable track record, the entire history of the manager may be added to the database (i.e., returns are backfilled). These backfilled returns are estimated to be about 1.3% per year higher than the returns

reported to the database after the manager is added to the database [8, 14, 16, 17].

Liquidity Risk

This refers to risk originating from restrictions on money withdrawal or illiquid assets. It can be estimated by examining share restriction parameters such as lockup as well as return autocorrelations. Returns on certain hedge fund strategies tend to display significant autocorrelation. This estimated autocorrelation may be caused by the presence of illiquid assets in hedge fund portfolios or could be due to return smoothing performed by managers in order to reduce their estimated risk [7, 9, 22]. Estimates of volatility and other risk measures should be adjusted to account for the presence of autocorrelation in reported returns.^g

Operational Risk

It is estimated that one of the main reasons for hedge fund failures is operational risk. Owing to their private nature, hedge funds have low degree of transparency, which may further contribute to operational risk. A recent study by Brown *et al.* [12] proposed to use the public filing information to build a numerical measure for operational risk. A further study shows that operational due diligence is an important source of fund of hedge funds alpha [10].

Risk Exposures

Hedge funds are called *absolute return investments* because they are supposed to provide positive returns

Table 2 Uninvestible indices

January 1990–November 2007	Equal weighted hedge fund	Convertible arbitrage	Distressed securities	Emerging markets	Equity market neutral	Equity long/short	Event driven multistrategy	Fixed income arbitrage	Global macro
Annualized mean (%)	14.03	10.00	13.29	14.36	8.81	13.09	12.48	7.12	11.22
Annualized standard (%)	6.77	2.74	6.28	14.01	1.83	8.14	5.22	1.76	6.16
Sharpe ratio	1.47	2.15	1.46	0.73	2.57	1.10	1.60	1.72	1.16
Skewness	-0.33	-1.06	-0.83	-1.16	-0.01	-0.27	-1.03	-2.35	1.09
Kurtosis	3.43	2.25	4.98	11.62	2.22	2.48	3.75	10.05	3.57
Autocorrelation	0.23	0.56	0.26	0.28	0.24	0.20	0.37	0.20	0.13
Correlation against S&P 500	0.70	0.27	0.49	0.49	0.35	0.76	0.63	0.12	0.43
Russell 2000	0.84	0.36	0.62	0.54	0.48	0.88	0.77	0.21	0.49
MSCI Emerging Mkts	0.74	0.32	0.55	0.74	0.28	0.65	0.62	0.24	0.48
LB Aggregate Bond	0.07	0.13	0.07	0.03	0.19	0.10	0.03	-0.04	0.27
LB Hi-Yld TR	0.55	0.47	0.54	0.45	0.31	0.55	0.66	0.39	0.36
GS Commodity	0.12	0.07	0.05	0.03	0.13	0.04	0.05	0.10	0.05

January 1990–November 2007	Merger arbitrage	S&P 500	Russell 2000	MSCI emerging markets	LB aggregate bond	LB Hi-Yld TR	GS commodity
Annualized mean (%)	10.07	11.10	11.60	14.83	7.03	8.74	9.25
Annualized standard (%)	4.05	13.76	18.10	22.45	3.77	7.49	19.61
Sharpe ratio	1.48	0.51	0.41	0.48	0.78	0.62	0.26
Skewness	-1.09	-0.47	-0.49	-0.76	-0.41	-0.39	0.43
Kurtosis	6.24	0.93	0.93	1.79	0.65	4.74	1.09
Autocorrelation	0.37	-0.04	0.10	0.16	0.16	0.29	0.10
Correlation against S&P 500	0.48	1.00	0.74	0.62	0.13	0.50	-0.08
Russell 2000	0.59	0.74	1.00	0.65	0.01	0.57	0.02
MSCI Emerging Mkts	0.46	0.62	0.65	1.00	-0.07	0.49	0.04
LB Aggregate Bond	0.06	0.13	0.01	-0.07	1.00	0.23	0.00
LB Hi-Yld TR	0.57	0.50	0.57	0.49	0.23	1.00	-0.08
GS Commodity	0.02	-0.08	0.02	0.04	0.00	-0.08	1.00

Convertible Arbitrage: January 1992–November 2007. Fixed Income Arbitrage: January 1998–November 2007

regardless of market conditions. Several recent studies have examined factor exposures of hedge funds with the goal of determining whether hedge funds are able to earn positive risk-adjusted excess returns (or alpha). The biases discussed above generally do not have significant impact on the estimates of factor exposures of hedge funds. Rather, they are likely to affect the estimates of alphas.

Using linear multifactor models, available research shows that, depending on strategy, hedge funds have exposures to both equity and fixed income factors.^h Though earlier studies reported significant exposures to these risk factors, the resulting R-squares were relatively small. Lack of explanatory power of traditional risk factors can be due to several reasons. First, most hedge fund returns are reported on monthly basis, while hedge fund managers change their factor exposures more often. This will reduce the explanatory power of linear factor models and may create nonlinear exposures in monthly returns.ⁱ Second, the multifactor models may be missing certain factors that are traded by hedge fund managers leading to low explanatory power. Third, the option-like incentive fee structure and investing in derivative securities can cause nonlinearity in risk exposures. Studies of factor exposures of hedge funds are generally conducted in order to determine if hedge funds offer positive excess risk-adjusted return. These studies generally show that top managers are able to generate risk-adjusted abnormal returns.^j Fung *et al.* [21] find that fund-of-funds alphas are shrinking over time as the markets become more competitive.

Hedge Fund Indices

Active Indices

Similar to traditional assets, there also exist a number of manager-based indices for hedge funds. These indices are not directly investible.

• Uninvestible indices

Traditionally, hedge fund indices have been based on a particular database. For example, CISDM hedge fund indices represent equally weighted average performance of managers who report to CISDM. In general, most indices do not impose restrictions on the managers that are added to these databases.^k Also, most indices do not impose a minimum track record and new hedge funds are allowed to be included in strategy indices. Not all indices maintain a database of those managers who at one point reported to their database, but decided at some point in the past to stop reporting. CISDM, HFR, and Lipper-TASS databases have maintained such databases of “dead funds” in the past. Depending on when a database becomes public, it may contain a different degree of backfilled bias. Generally, databases created more recently tend to display a higher amount of backfilled bias.^l Table 2 provides a summary of uninvestible indices.

• Investible indices

Since 2000, a number of manager-based hedge fund indices have been created that may be regarded as investible. Unlike their uninvestible counterparts, these investible indices impose several restrictions on hedge funds that are included in these indices. Foremost among these restrictions is that the managers must be open to new investments and must have adequate capacity. These managers then go through a rigorous due diligence process and at the end only a very small number of them are included in these indices.^m Typically, there are between 6 and 12 managers in each strategy investible index. This may indicate that these indices are not representative of the hedge fund universe. Further, since performance figures of these investible indices are typically reported on a daily basis, financial institutions that offer these indices require these managers to carry more liquid positions when compared to other managers. This selection bias

Table 3 Investible indices

Index provider	Launch date	Base date	Fund weighting	Number of funds in the index	Rebalancing frequency
CS-Termont	August, 2003	January, 2000	Value weighted	+70	Semiannually
Dow Jones	November, 2003	January, 2002	Equal weighted	+70	Quarterly
HFRX	March, 2003	January, 2000	NA	NA	Quarterly
MSCI	July, 2003	January, 2000	Equal weighted	+90	Quarterly
S&P	May, 2002	January, 1998	Equal weighted	+40	Annual

may affect the risk-return profiles of these managers. Table 3 provides a summary of investible indices.

End Notes

^a. See <http://www.awjones.com/historyofthefirm.html>

^b. See <http://www.hfr.com>

^c. See the Wall Street Journal, November 3, 2004.

^d. See <http://treas.gov/press/releases/reports/hedgfund.pdf> and [26]

^e. Survivorship bias has long been recognized in the mutual fund literature (e.g., see [23] and [32]).

^f. Reference [2] discusses overlap among major hedge fund databases.

^g. References [6] and [31] report that those hedge funds that display significant autocorrelation in their reported returns tend to outperform other hedge funds that follow the same strategy.

^h. Mitchell and Pulvino [33] study risk factors of merger arbitrage strategy, Fung and Hsieh [18] and Duarte *et al.* [13] study fixed income hedge funds, Agarwal and Naik [5] and Fung and Hsieh [19] study equity long/short strategy and Agarwal *et al.* [3] examine risk factors of convertible arbitrage strategy.

ⁱ. See [24]. These nonlinearities may be captured by using returns on option positions as explanatory factors (see [17]).

^j. Among others, Fung and Hsieh [15], Schneeweis and Spurgin [34], Liang [28], Agarwal and Naik [4]; Fung and Hsieh [20]; Fung *et al.* [21] and Kosowski *et al.* [27] use multifactor models to measure the performance of hedge funds.

^k. Credit Suisse, Hennessee, and MSCI indices impose a minimum size for assets under management. Credit Suisse and Hennessee indices require a minimum track record of one year.

^l. CISDM, HFR, and Lipper-TASS are among the oldest databases.

^m. Some of the more well-known investible indices are Dow Jones Hedge Fund Strategy Benchmarks, HFR, Credit Suisse, and MSCI investible indices.

References

- [1] Ackermann, C., McEnally, R. & Ravenscraft, D. (1999). The performance of hedge funds: risk, return and incentive, *Journal of Finance* **54**, 833–874.
- [2] Agarwal, V., Daniel, N. & Naik, N. (2004). *Flows, Performance, and Managerial Incentives in Hedge Funds*. Working Paper, London Business School. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=424369.
- [3] Agarwal, A., Fung, W., Loon, Y. & Naik, N. (2006). *Liquidity Provision in the Convertible Bond Market: Analysis of Convertible Arbitrage Hedge Funds*. Working Paper, London Business School. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=885945.
- [4] Agarwal, N. & Naik, N. (2001). *Performance Evaluation of Hedge Funds with Option-Based and Buy-and-Hold Strategies*. Working Paper, London Business School.
- [5] Agarwal, V. & Naik, N. (2004). Risks and portfolio decisions involving hedge funds, *Review of Financial Studies* **17**, 63–98.
- [6] Aragon, G. (2007). Share restriction and asset pricing: evidence from the hedge fund industry, *Journal of Financial Economics* **83**, 33–58.
- [7] Asness, C., Krail, R. & Liew, J. (2001). Do hedge funds hedge? *Journal of Portfolio Management* **20**, 6–19.
- [8] Barry, R. (2003). *Hedge Funds: A Walk through the Graveyard*. Working Paper, Macquarie Applied Finance Center. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=333180
- [9] Bollen, N. & Pool, V. (2007). *Conditional Return Smoothing in the hedge fund Industry*. Working Paper, Vanderbilt University - Owen Graduate School of Management. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=937990, (2008). *Journal of Financial and Quantitative Analysis* **43**, 267–298.
- [10] Brown, S.J., Fraser, T. & Liang, B. (2008). Hedge fund due diligence: a source of alpha in a hedge fund portfolio strategy, *Journal of Investments Management* **6**, 23–33.
- [11] Brown, S.J., Goetzmann, W. & Ibbotson, R. (1999). Off-shore hedge funds: survival & performance 1989–1995, *Journal of Business* **72**, 91–118.
- [12] Brown, S.J., Goetzmann, W., Liang, B. & Schwarz, C. (2008). Mandatory disclosure and operational risk: evidence from hedge fund registration, *Journal of Finance* **63**, 2785–2815.
- [13] Duarte, J., Longstaff, F. & Yu, F. (2005). *Risk and Return in Fixed Income Arbitrage: Nickels in Front of a Steamroller?* Working paper, UCLA.
- [14] Edwards, F.R. & Caglayan, M. (2001). Hedge fund performance and manager skill, *Journal of Futures Markets* **21**, 1003–1028.
- [15] Fung, W. & Hsieh, D.A. (1997). Empirical characteristics of dynamic trading strategies: the case of hedge funds, *Review of Financial Studies* **10**, 275–302.
- [16] Fung, W. & Hsieh, D.A. (2000). Performance characteristics of hedge funds and commodity funds: Natural vs. Spurious biases, *Journal of Financial and Quantitative Analysis* **35**, 291–307.
- [17] Fung, W. & Hsieh, D.A. (2001). The risk in hedge fund strategies: theory and evidence from trend followers, *Review of Financial Studies* **14**, 313–341.
- [18] Fung, W. & Hsieh, D.A. (2002). The risk in fixed-income hedge fund styles, *Journal of Fixed Income* **12**, 16–27.
- [19] Fung, W. & Hsieh, D.A. (2004). Extracting portable alphas from equity long-short hedge funds, *Journal of Investment Management* **2**, 57–75.
- [20] Fung, W. & Hsieh, D.A. (2004a). Hedge fund benchmarks: a risk based approach, *Financial Analyst Journal* **60**, 65–80.

-
- [21] Fung, W., Hsieh, D., Naik, N. & Ramadorai, T. (2008). Hedge funds: performance, risk and capital formation, *Journal of Finance* **63**, 1777–1803.
 - [22] Getmansky, M., Lo, A. & Makarov, I. (2004). An econometric model of serial correlation and liquidity in hedge fund returns, *Journal of Financial Economics* **74**, 529–609.
 - [23] Goetzmann, W., Brown, S. & Ross, S. (1995). Survival, *Journal of Finance* **50**, 853–873.
 - [24] Goetzmann, W., Ingersoll, J. & Ivkovic, Z. (2000). Monthly measurement of daily timers, *Journal of Financial and Quantitative Analysis* **35**(3), 257–290.
 - [25] Goetzmann, W., Ingersoll, J. & Ross, S. (2003). High-water marks and hedge fund management contracts, *Journal of Finance* **58**(4), 1685–1717.
 - [26] Kazemi, H., Karavas, V., Martin, G. & Schneeweis, T. (2005). The impact of leverage on hedge fund risk and return, *Journal of Alternative Investments* **7**, 1–12.
 - [27] Kosowski, R., Naik, N. & Teo, M. (2005). *Do Hedge Funds Deliver Alpha? A Bayesian and Bootstrap Analysis*. Working Paper, London Business School. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=829025
 - [28] Liang, B. (1999). On the performance of hedge funds, *Financial Analysts Journal* **12**, 72–85.
 - [29] Liang, B. (2000). Hedge funds: the living and the dead, *The Journal of Financial and Quantitative Analysis* **35**, 309–326.
 - [30] Liang, B. & Park, H. (2007a). Risk measures for hedge funds: a cross-sectional approach, *European Financial Management* **13**, 333–370.
 - [31] Liang, B. & Park, H. (2007). *Share Restrictions, Liquidity Premium, and Offshore Hedge Funds*. Working Paper, University of Massachusetts, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=967788
 - [32] Malkiel, B. (1995). Returns from investing in equity mutual funds 1971 to 1991, *Journal of Finance* **50**, 549–572.
 - [33] Mitchell, M. & Pulvino, T. (2001). Characteristics of risk in risk arbitrage, *Journal of Finance* **56**, 2135–2175.
 - [34] Schneeweis, T. & Spurgin, R. (1999). Managed futures, hedge fund and mutual fund return estimation: a multi-factor approach, *Journal of Alternative Investments* **1**, 1–24.

HOSSEIN KAZEMI & BING LIANG

Performance Measures

The measurement of performance is the cornerstone of the evaluation of a fund. Since the advent of modern portfolio theory, it has been the focus of extensive practitioner and academic literature. We start by examining the hypothesis underlying the most commonly used measure, namely, the Sharpe ratio. We then present performance measures avoiding normality assumptions. The section on factor models analyzes the addition of multiple benchmarks for performance measurement. Finally, performance measures considering portfolio holdings are discussed.

Sharpe Ratio

The definition of the *ex post* Sharpe ratio (*see also Sharpe Ratio*) is as follows:

$$Sharpe = \frac{\bar{r}}{\sigma} \quad (1)$$

where $\bar{r}$ is the average excess return over the risk-free rate and σ is the standard deviation of the excess return.

From a theoretical point of view, the Sharpe ratio finds its roots in the Capital Market Line, which weighs expected returns against the volatility. As a consequence, measuring performance with the *ex post* Sharpe ratio implicitly assumes that the return time series are normally distributed or that investor preferences can be described by a quadratic utility function. Moreover, the return time series need to be independently and identically distributed (i.i.d.) as shown in [11].

Performance Measures in the Absence of Normality

Many asset classes, for example hedge funds, do not exhibit normally distributed returns. The use of derivatives has increased the potential of asymmetric behavior of investment strategies. The Sharpe ratio may be significantly increased in such circumstances ([18]). Moreover, this ratio does not reflect the preferences of risk-averse investors who tend to dislike negative skewness and high kurtosis (as shown in [14] and [16]).

As a remedy, new performance measures have emerged in order to cope with the absence of normality. The Sortino ratio, proposed by Sortino and Price [17], has been advocated to capture the asymmetry of the return distribution. It replaces the standard deviation in the Sharpe ratio by the downside deviation, which captures only the downside risk. The Sortino ratio is expressed as follows:

$$\text{Sortino} = \frac{\bar{R} - RMAR}{\text{Downside}(R, RMAR)} \quad (2)$$

where $\bar{R}$ is the average return of the fund and $RMAR$ is the minimum acceptable rate of return.

$$\begin{aligned} \text{Downside}(R, RMAR) \\ = \sqrt{\frac{1}{T} \sum_t 1_{R_t < RMAR} (R_t - RMAR)^2} \end{aligned} \quad (3)$$

The Stutzer index ([19]) does not rely explicitly on a utility function but on a behavioral hypothesis related to the “safety-first” principle of Roy [15]. It assumes that investors aim at minimizing the probability that the excess returns over a given threshold will be negative over a long time horizon. When the portfolio has a positive expected excess return, the aforementioned probability will exponentially decay to zero as the time horizon increases without limit irrespective of the threshold. The maximum possible decay is defined as the Stutzer index. More precisely, if the excess return process is i.i.d., the Stutzer index is expressed as follows:

$$I = - \lim_{T \rightarrow \infty} \frac{1}{T} \log (\text{Prob}[\bar{r}_T \leq 0]) > 0 \quad (4)$$

where $\bar{r}_T$ is the average excess return earned over the threshold periods.

It is interesting to note that for normal return distributions, the ranking of investments is exactly the same as with the Sharpe ratio. This is due to the fact that the Stutzer index is equal to half of the square of the Sharpe ratio in this case. In more general cases, higher moments of the distribution will have an impact on the value of the index. For example, a distribution with negative skewness and high kurtosis will result in a lower Stutzer index value than a normal distribution with the same mean and variance.

The Omega measure suggested by Keating and Shadwick [8] incorporates all the moments of the

distribution by a direct transformation of the latter. More specifically, Omega provides a risk-reward measure for which returns are weighted by their probability of occurrence. In practice, the Omega function at a given threshold can be estimated as

$$\Omega(\text{Thres}) = \frac{\frac{1}{T} \sum_t \max(0, R_t - \text{Thres})}{\frac{1}{T} \sum_t \max(0, \text{Thres} - R_t)} \quad (5)$$

where R_t is the return of the fund at time t and Thres is the threshold defined by the investor.

The Omega function encodes the cumulative density function of the returns (see [8] for a proof). As a consequence, all moments are embodied in the function. In practice, investments may be difficult to rank with this performance measure. When the Omega function of a given investment dominates the Omega function of another, the investment second order stochastically dominates the other. All other possibilities lead to unclear results without defining risk preferences of the investors or without defining a specific metric on the Omega function.

Performance Measures versus Factor Models

Factor models are widely used to decompose and analyze the performance of funds. A general model can be expressed as follows:

$$r_t = \alpha + \sum_i \beta_i F_{it} + \varepsilon_t \quad (6)$$

where r_t is the excess return of the fund over the risk-free rate, F_{it} is the excess return of factor i , and α denotes the alpha of the fund.

Guided by the capital asset pricing model, Jensen [6, 7] uses a single market index as a benchmark. Merton's Intertemporal Capital Asset Pricing Model (ICAPM) [12] provides a theoretical framework for the presence of additional factors in equation (6). Fama and French [3] discovered purely empirical support for including a size (market capitalization) factor and a value (book-to-market ratio) factor, while Carhart [1] added a momentum factor. The emergence of hedge funds as an asset class has further refined the use of factor models. Fung and Hsieh [4] document that factor models are much

poorer at decomposing hedge fund returns than at decomposing mutual fund returns. They show that hedge fund returns contain exposure to additional alternative factors.

The estimation of alpha is severely impacted by the length of the track record under consideration. Pastor and Stambaugh [13] propose a new methodology to increase the precision of the estimation and distinguish good managers from lucky ones. This methodology relies on information (returns) extracted from additional "seemingly unrelated" assets not used in the definition of the alpha. The additional non-benchmark assets help estimate alpha if they are priced by the benchmarks or if their return histories are longer than that of the fund. Kosowski *et al.* [10] advocate a robust bootstrap approach for coping with the cross-sectional nonnormal distribution of individual fund alpha. Using both approaches, Kosowski *et al.* [9] show that persistence can be found in the alpha of top hedge fund managers.

Analysis Based on Portfolio Holdings Information

When portfolio holdings are known, the analysis can be enhanced by looking at the relationship between the weights and the future returns of the portfolio components. More specifically, the performance measure is defined as

$$\text{PerfHolding} = \sum_i \text{cov}(w_{it-1}; R_{it}) \quad (7)$$

where w_{it-1} is the weight of underlying i at the end of period $t-1$ and R_{it} is the return of underlying i between $t-1$ and t .

Grinblatt and Titman [5] estimate equation (7) as

$$\text{PerfHolding} = \sum_i \frac{1}{T} \sum_t [w_{it-1} - E(w_{it-1})] R_{it} \quad (8)$$

where $E(w_{it-1})$ refers to the benchmark weight of underlying i . In practice, the benchmark weight may be chosen as a long-run average.

Daniel *et al.* [2] introduce benchmarks to proxy characteristics of stocks in this context. They show that the gross return of an equity fund can be decomposed into three components:

- Characteristic selectivity:

$$CS_t = \sum_i w_{it-1} (R_{it} - R_t^{b_{it-1}}) \quad (9)$$

- Characteristic timing:

$$CT_t = \sum_i (w_{it-1} R_{it}^{b_{it-1}} - w_{it-k-1} R_{it}^{b_{it-k-1}}) \quad (10)$$

- Average style measure:

$$AS_t = \sum_i w_{it-k-1} R_{it}^{b_{it-k-1}} \quad (11)$$

where $R_t^{b_{it-1}}$ is the return of a value weighted portfolio that matches the characteristics of stock i at the end of period $t - 1$. Wermers [20] documents the superiority of this approach *versus* the factor decomposition.

References

- [1] Carhart, M. (1997). On persistence in mutual fund performance, *Journal of Finance* **52**, 57–82.
- [2] Daniel, K., Grinblatt, M., Titman, S. & Wermers, R. (1997). Measuring mutual fund performance with characteristic-based benchmarks, *Journal of Finance* **52**, 1035–1058.
- [3] Fama, E. & French, K. (1992). The cross-section of expected stock returns, *Journal of Finance* **47**(2), 427–465.
- [4] Fung, W. & Hsieh, D. (1997). Empirical characteristics of dynamic trading strategies: the case of hedge funds, *Review of Financial Studies* **10**, 275–302.
- [5] Grinblatt, M. & Titman, S. (1993). Performance measurement without benchmarks: an examination of mutual fund returns, *Journal of Business* **66**, 47–68.
- [6] Jensen, M. (1968). The performance of mutual funds in the period 1945–1964, *Journal of Finance* **23**, 389–416.
- [7] Jensen, M. (1969). Risk, the pricing of capital assets, and the evaluation of investment portfolios, *Journal of Business* **42**, 167–247.
- [8] Keating, C. & Shadwick, F. (2002). A universal performance measure, *The Journal of Performance Measurement* **6**(3), 59–84.
- [9] Kosowski, R., Naik, N. & Teo, M. (2007). Do hedge funds deliver alpha? A Bayesian and bootstrap analysis, *Journal of Financial Economics* **84**(1), 229–264.
- [10] Kosowski, R., Timmerman, A., Wermers, R. & White, H. (2006). Can mutual fund “stars” really pick stocks? New evidence from a bootstrap analysis, *Journal of Finance* **61**(6), 2551–2595.
- [11] Lo, A. (2002). The statistics of Sharpe ratios, *Financial Analysts Journal* **58**, 36–52.
- [12] Merton, R. (1973). An intertemporal asset pricing model, *Econometrica* **41**, 867–887.
- [13] Pastor, L. & Stambaugh, R. (2002). Mutual fund performance and seemingly unrelated assets, *Journal of Financial Economics* **63**(3), 315–349.
- [14] Pratt, J. & Zeckhauser, R. (1987). Proper risk aversion, *Econometrica* **55**(1), 143–154.
- [15] Roy, A. (1952). Safety-first and the holding of assets, *Econometrica* **20**, 431–449.
- [16] Scott, R. & Horvath, P. (1980). On the direction of preference for moments of higher order than the variance, *Journal of Finance* **35**(4), 915–919.
- [17] Sortino, F. & Price, L. (1994). Performance measurement in a downside risk framework, *Journal of Investing* **3**(3), 59–65.
- [18] Spurgin, R. (2001). How to game your sharpe ratio, *Journal of Alternative Investments* **4**(3), 38–46.
- [19] Stutzer, M. (2000). A portfolio performance index, *Financial Analysts Journal* **56**(3), 52–61.
- [20] Wermers, R. (2006). Performance evaluation with portfolio holdings information, *North American Journal of Economics and Finance* **17**, 207–230.

Further Reading

- Ferson, W. & Schadt, R. (1996). Measuring fund strategy and performance in changing economic conditions, *Journal of Finance* **51**, 425–461.
- Fung, W. & Hsieh, D. (2004). Hedge fund benchmarks: a risk based approach, *Financial Analyst Journal* **60**, 65–80.
- Henriksson, R. & Merton, R. (1981). On market timing and investment performance II. Statistical procedures for evaluating forecasting skills, *Journal of Business* **54**, 513–533.
- Lehmann, B. & Modest, D. (1987). Mutual fund performance evaluation: a comparison of benchmarks and benchmark comparisons, *Journal of Finance* **42**, 233–265.
- Mamaysky, H., Spiegel, M. & Zhang, H. (2008). Estimating the dynamics of mutual fund alphas and betas, *Review of Financial Studies* **21**, 233–264.
- Treynor, J. & Mazuy, K. (1966). Can mutual funds outguess the market? *Harvard Business Review* **44**, 131–136.

Related Articles

Sharpe Ratio; Style Analysis.

JEAN-FRANCOIS BACMANN & STEFAN SCHOLZ

Sharpe Ratio

The Sharpe ratio (SR) is a popular measure of portfolio performance that was introduced as the reward-to-variability ratio in [6]. For a given set of returns during an evaluation period, portfolio p 's *ex post* SR is

$$SR_p = \frac{\bar{R}_p - \bar{R}_f}{s_p} \quad (1)$$

where $\bar{R}_p$ and s_p denote, respectively, the historical average and standard deviation of p 's returns; $\bar{R}_f$ denotes the average risk-free return.^a A portfolio's SR is typically compared to a market index m 's SR:

$$SR_m = \frac{\bar{R}_m - \bar{R}_f}{s_m} \quad (2)$$

Figure 1 shows that SR_p represents the slope of a line originating at the average risk-free rate and extending to point p where the portfolio is located. Similarly, SR_m represents the slope of a line extending to point m where the index is located. This line is a proxy for the *ex post* capital market line (CML) in [5]. Portfolio p has had superior performance relative to m since $SR_p > SR_m$, resulting in it plotting above the CML; if it had plotted below, then $SR_p < SR_m$ and it would have had inferior performance.

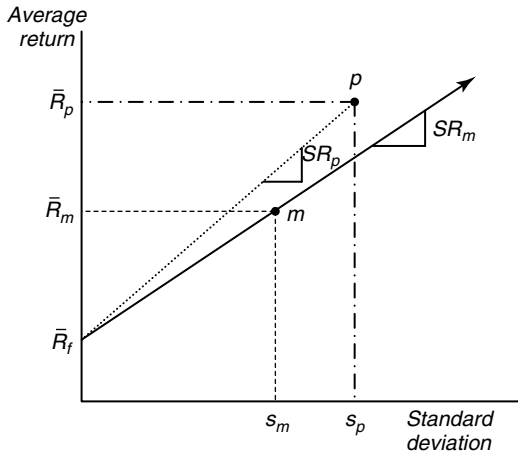


Figure 1 Evaluating the performance of portfolio p with the Sharpe ratio

The SR can also be used to make portfolio construction decisions as in [7]. For a given holding period, portfolio p 's *ex ante* SR is

$$\frac{E[R_p] - R_f}{\sigma[R_p]} \quad (3)$$

where $E[R_p]$ and $\sigma[R_p]$ denote, respectively, the expected value and standard deviation of p 's future return, and R_f denotes the risk-free return. In frictionless markets with a risk-free security, the tangency portfolio in [10] has the highest SR (in absolute value). Optimal portfolios of investors using mean-variance analysis consist of combinations of the risk-free security and the tangency portfolio, which is the market portfolio in [5].

The SR has limitations. First, it assumes that either (i) investors' objective functions are defined solely over the first two moments of portfolio return distributions or (ii) such distributions are characterized by only these moments. Second, proper use of the SR requires knowledge of investors' investment horizons. Levy [4] shows that a bias emerges when using it with a horizon that is different from that of investors. Cvitanic *et al.* [2] find that maximizing the short-term *ex ante* SR is suboptimal for long-term investors. Third, the SR can be manipulated by using derivatives. Ingersoll *et al.* [3] define and characterize a manipulation-proof performance measure. Their measure resembles the average value of a power utility function defined over portfolio returns.

End Notes

^aFor other reward-to-risk ratios of portfolio performance that use VaR and beta instead of standard deviation as risk measures in the denominator of equation (1), see references [1] and [11], respectively. Also see references [8] and [9].

References

- [1] Alexander, G.J. & Baptista, A.M. (2003). Portfolio performance evaluation using Value-at-Risk, *Journal of Portfolio Management* **29**, 93–102.
- [2] Cvitanic, J., Lazrak, A. & Wang, T. (2008). Implications of the Sharpe ratio as a performance measure in multi-period settings, *Journal of Economic Dynamics and Control* **32**, 1622–1649.
- [3] Ingersoll, J., Spiegel, M., Goetzmann, W. & Welch, I. (2007). Portfolio performance manipulation and manipulation-proof performance measures, *Review of Financial Studies* **20**, 1503–1546.

- [4] Levy, H. (1972). Portfolio performance and the investment horizon, *Management Science* **18**, B645–B653.
- [5] Sharpe, W.F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk, *Journal of Finance* **19**, 425–442.
- [6] Sharpe, W.F. (1966). Mutual fund performance, *Journal of Business* **39**, 119–138.
- [7] Sharpe, W.F. (1975). The Sharpe ratio, *Journal of Portfolio Management* **21**, 49–58.
- [8] Sortino, F., van der Meer, R. & Plantinga, A. (1999). The Dutch triangle, *Journal of Portfolio Management* **26**, 50–58.
- [9] Stutzer, M. (2000). A portfolio performance index, *Financial Analysts Journal* **56**, 52–61.
- [10] Tobin, J. (1958). Liquidity preference as behavior towards risk, *Review of Economic Studies* **25**, 65–86.
- [11] Treynor, J. (1965). How to rate management of investment funds, *Harvard Business Review* **43**, 63–75.

Related Articles

Capital Asset Pricing Model; Performance Measures; Sharpe, William F.

GORDON J. ALEXANDER & ALEXANDRE M.
BAPTISTA

Risk–Return Analysis

Mean–Variance Analysis

While the idea of trade-off curves goes back to, Pareto, the notion of a trade-off curve (later dubbed *efficient frontier*) in finance was introduced by Markowitz [15] in 1952. Markowitz proposed expected return and variance as both a hypothesis about how investors act and as a rule for guiding action in fact. By 1959 [18], he had given up the notion of mean and variance as a hypothesis but continued to propose them as criteria for action.

Tobin [32] said that the use of mean and variance as criteria assumed either a quadratic utility function or a Gaussian probability distribution. This view is sometimes ascribed to Markowitz, but he never justified the use of mean and variance in this way. His views evolved considerably from 1952 to 1959 [15] and [18]. Consequently, his early views [15] should be ignored. In his book published in 1959, Markowitz [18] accepts the views of Von Neumann and Morgenstern [33] when probability distributions are known, and Leonard J. Savage [29] when probabilities are not known. The former asserts that one should maximize expected utility, whereas the latter asserts that when probabilities are not known one should maximize expected utility using probability beliefs when objective probabilities are not known.

A utility function $U(R)$ is any function of return R . If the function is concave, that is,

$$U(aX + (1 - a)Y) \geq aU(X) + (1 - a)U(Y) \quad (1)$$

for $a \in [0, 1]$ then an investor prefers certainty to random returns. If

$$U(aX + (1 - a)Y) \leq aU(X) + (1 - a)U(Y) \quad (2)$$

then the investor prefers gambling to certainty. An intermediate case is when the utility function is linear. In this case, the gambler is neither risk averse as in equation (1) nor risk seeking as in equation (2). Rather, the investor acts to maximize expected return.

Markowitz [18] seeks to serve the cautious investor whose utility function satisfies equation (1). In this respect, Markowitz, assumes a different shape of utility function in [15] from the one in [16]. In the latter, he assumes a utility function that is convex

in some places and concave in others. This was proposed as a hypothesis about human action consistent with gambling and insurance. It builds upon and, the author believes, improves upon [7].

Critical Line Algorithm

The critical line algorithm (CLA) is presented in [17] and again in Appendix A of [18]. The latter shows that the algorithm works even if the matrix of covariances, required by the problem, is singular. This is important since the problem may include short as well as long positions (treated as separate “investments” to represent real-world constraints on the selection of the portfolio when short-sales are permitted), or covariances based on historical returns when there are more securities than observations. In either case, a singular covariance matrix will result.

The CLA makes use of the fact that, in portfolio space, the set of efficient portfolios is piecewise linear. In mean–variance space it is piecewise parabolic; in mean–standard deviation space it is piecewise linear or hyperbolic (see [19] or [22], Chapter 7). The CLA generates, one after the other, all the linear pieces of the portfolio-space set without groping or iterating for the right answer. In this manner, CLA generates the entire mean–variance efficient set almost as quickly as the best methods for finding a single point in this set.

The CLA uses the George Dantzig simplex algorithm [6] of linear programming to determine the first critical line. The portfolio selection constraint may be bounded or unbounded. In the latter case, the first critical line is a ray that proceeds without bounds in the direction of increasing expected return. Again, see [22, Chapter 8], Chapter 9 of [22] discusses the degenerate case in which variables go to zero simultaneously.

Mean–Variance Approximations to Expected Utility

Markowitz [18] accepts the justifications of Von Neumann and Morgenstern [33], and Leonard J. Savage [29] for expected utility using personal probabilities when objective probabilities are not known. He conjectures that a suitably chosen point from the efficient frontier will approximately maximize expected utility for the kinds of utility functions that

are commonly proposed for cautious investors, and for the kinds of probability distributions that are found in practice. Levy and Markowitz [14] expand on this notion considerably. Specifically, Levy and Markowitz show that for such probability distributions and utility functions there is typically a correlation between the actual expected utility and the mean–variance approximation in excess of 0.99. Levy and Markowitz also show that the Pratt [27] and Arrow [1] objections to quadratic utility do not apply to the kind of approximations used by them, or to those in [18]. Specifically, Pratt and Arrow assume that the investor has one fixed-forever utility function and, as her or his wealth changes, the investor moves up and down this utility function. Under this assumption, if the investor becomes wealthy enough, quadratic utility no longer increases with wealth. This is clearly absurd. The Levy and Markowitz [14] and the Markowitz [18] approximations are based on Taylor expansions about current wealth, or expected end-of-period wealth.

Models of Covariance

If covariances are computed from historical returns with more securities than there are observations, for example, 5000 securities and 60 months of observations, then the covariance matrix will be singular. A preferable alternative is to use a linear model of covariance where the return on the i th security is assumed to obey the following relationship:

$$r_i = \alpha_i + \Sigma \beta_{ik} f_k + u_i \quad (3)$$

where u_i are independent of each other and the f_k . The f_k may be either factors or scenarios or some of each. These ideas are carried out in, for example [28, 30], and [20, 21].

Semivariance

In [18, Chapter 9], Markowitz defines two forms of semivariance, namely, about expected value E or about some fixed number a :

$$S_E = E(\min(r, E)^2) \quad (4)$$

or

$$S_a = E(\min(r, a)^2) \quad (5)$$

In [18], Markowitz says that some form of semi-variance seems preferable to mean–variance analysis, but computation is a problem. At present, this computational problem has disappeared. In [18, Chapter 9], he presents a variant of CLA, which traces out the mean–semivariance frontier. Sortino [31] champions the use of semivariance. Mean–variance analysis has been dubbed *modern portfolio theory (MPT)*. Sortino refers to risk–return analysis with semivariance as *postmodern portfolio theory*.

The chief argument in favor of semivariance, as opposed to variance, is that the investor is not concerned with upside deviations; she or he is concerned only with downside deviations. Arguments in favor of mean–variance rather than semivariance are as follows: variance requires only the covariance matrix as input rather than historical returns, or synthetic history generated randomly; and, the mean–variance approximations of expected utility do so well when probability distributions are not spread out “too much”.

Other Measures of Risk

Konno [11] recommends absolute deviation as a criterion for risk in a risk–return trade-off analysis. An advantage of these criteria is that the frontier may be traced out using linear programming.

In [18, Chapter 13], Markowitz objects to these criteria because the function that they imply as the approximation to the utility function does not seem plausible. A similar, but even stronger, objection is raised there to the use of probability of loss as the measure of risk.

Time

In [18, Chapters 11 and 13], Markowitz notes that mean–variance analysis is a single-period analysis, but that does not mean that it is useless in a many-period world. Bellman [2] shows that the optimum strategy for a many-period or infinite-period game consists of maximizing a sequence of single-period utility functions where the utility function is the “derived” utility for the game. If assets are perfectly liquid, the end-of-period derived utility function depends only on end-of-period wealth, and if the Levy–Markowitz approximations to expected utility are good enough, then one may use mean–variance

for a many-period game. If the end-of-period utility function depends on other state variables, and the utility function may be adequately approximated by a quadratic, then the action should depend on mean and variance, and covariance with the other state variables.

If assets are not perfectly liquid, then state variables include the holding of each asset. This results in the problem referred to as the *curse of dimensionality*. Markowitz and van Dijk [19] propose a heuristic for solving this problem. This heuristic approximates the unknown derived utility function by a quadratic in the various state variables. Kritzman [12] report as follows:

Our tests reveal that the quadratic heuristic provides solutions that are remarkably close to the dynamic programming solutions for those cases in which dynamic programming is feasible and far superior to solutions based on standard industry heuristics. In the case of five assets, in fact, it performs better than dynamic programming due to approximations required to implement the dynamic programming algorithm. Moreover, unlike the dynamic programming solution, the quadratic heuristic is scalable to as many as several hundred assets.

Estimation of Parameters

Covariance matrices are sometimes estimated from historical returns and sometimes from factor or scenario models such as the one-factor model of Sharpe [30], the many-factor model of Rosenberg [28], or the scenario models of Markowitz and Perold [20, 21].

Expected returns are estimated in a great variety of ways. It is unlikely that anyone would suggest that the expected returns of individual stocks be estimated from the historical average returns. The Ibbotson [9] series are frequently used to estimate expected returns for asset classes. Black and Litterman [3, 4] propose a very interesting Bayesian approach to the estimation of expected returns. Richard Michaud [25] proposes to use estimates for asset classes based on what he refers to as a *resampled frontier*. Markowitz and Usmen [23] test the resampled frontier idea against a diffuse Bayesian approach. By and large, they find that the Michaud approach outperformed the diffuse Bayesian approach. However, Markowitz and Usmen noted that had they increased the variance estimated by the Bayesian they would have done approximately as well as the Michaud approach. Somehow,

Michaud's patented process, which averages repeated drawings of frontiers generated from a Gaussian distribution with historical covariances, seems to essentially replicate a supercautious Bayesian.

Additional methods for estimating expected return are based on statistical methods for "disentangling" various anomalies [10], or estimates based on factors that [8] might use. See [13, 26], and [5]. The latter paper is based on results obtained by back-testing many alternate hypotheses concerning how to achieve excess returns. When many estimation methods are tested, the expected future return for the best of the lot (assuming that nature will sample from the same population as before) should not be estimated as if this were the only procedure tested (see [24]).

References

- [1] Arrow, K. (1971). *Aspects of the Theory of Risk Bearing*, Markham Publishing Company, Chicago, IL.
- [2] Bellman, R.E. (1957). *Dynamic Programming*, Princeton University Press, Princeton, NJ.
- [3] Black, F. & Litterman, R. (1991). Asset allocation: combining investor views with market equilibrium, *Journal of Fixed Income* 1(2), 7–18.
- [4] Black, F. & Litterman, R. (1992). Global portfolio optimization, *Financial Analysts Journal* 48(5), 28–43.
- [5] Bloch, M., Guerard, J., Markowitz, H., Todd, P. & Xu, G. (1993). A comparison of some aspects of the U.S. and Japanese equity markets, *Japan and the World Economy* 5, 3–26.
- [6] Dantzig, G.B. (1954). *Notes on Linear Programming: Parts VIII, IX, X—Upper Bounds, Secondary Constraints, and Block Triangularity in Linear Programming*, The RAND Corporation, Research Memorandum RM-1367, October 4, 1954. Published in *Econometrica*, Vol. 23, No. 2, April 1955, pp. 174–183.
- [7] Friedman, M. & Savage, L.P. (1948). The utility analysis of choices involving risk, *Journal of Political Economy*, 56, 279–304.
- [8] Graham, B. & Dodd, D.L. (1940). *Security Analysis*, 2nd Edition, McGraw-Hill, New York.
- [9] Ibbotson R.G. (2009). *Market Results for Stock, Bonds, Bills, and Inflation 1926–2008*. Classic Yearbook, Morningstar, Inc, Chicago, IL.
- [10] Jacobs, B.I. & Levy, K.N. (1988). Disentangling equity return regularities: new insights and investment opportunities, *Financial Analysts Journal* 44(3), 18–44.
- [11] Konno, H. & Yamazaki, H. (1991). Mean-absolute deviation portfolio optimization model and its applications to Tokyo stock market, *Management Science* 37(5), 519–531.
- [12] Kritzman, M., Myrgren, S. & Page, S. (2007). *Portfolio Rebalancing: A Test of the Markowitz-van Dijk Heuristic*, Pending Publication.

- [13] Lakonishok, J., Shleifer, A. & Vishny, R.W. (1994). Contrarian investment, extrapolation and risk, *Journal of Finance* **49**(5), 1541–1578.
- [14] Levy, H. & Markowitz, H.M. (1979). Approximating expected utility by a function of mean and variance, *American Economic Review* **69**(3), 308–317.
- [15] Markowitz, H.M. (1952). Portfolio selection, *The Journal of Finance* **7**(1), 77–91.
- [16] Markowitz, H.M. (1952). The utility of wealth, *The Journal of Political Economy* **2**, 152–158.
- [17] Markowitz, H.M. (1956). The optimization of a quadratic function subject to linear constraints, *Naval Research Logistics Quarterly* **3**, 111–133.
- [18] Markowitz, H.M. (1959, 1991). *Portfolio Selection: Efficient Diversification of Investments*, 2nd Edition, Wiley, Yale University Press, Basil Blackwell.
- [19] Markowitz, H.M. & van Dijk, E. (2003). Single-period mean–variance analysis in a changing world, *Financial Analysts Journal* **59**(2), 30–44.
- [20] Markowitz, H.M. & Perold, A.F. (1981). Portfolio analysis with factors and scenarios, *The Journal of Finance* **36**(4), 871–877.
- [21] Markowitz, H.M. & Perold, A.F. (1981). Sparsity and piecewise linearity in large portfolio optimization problems, in *Sparse Matrices and Their Uses*, I.S. Duff ed, Academic Press, pp. 89–108.
- [22] Markowitz, H.M. & Todd, P. (2000). *Mean-Variance Analysis in Portfolio Choice and Capital Markets*, Frank J. Fabozzi Associates, New Hope, PA. (revised reissue of Markowitz (1987) with chapter by Peter Todd).
- [23] Markowitz, H.M. & Usmen, N. (2003). Resampled frontiers versus diffuse Bayes: an experiment, *Journal of Investment Management* **1**(4), 9–25.
- [24] Markowitz, H.M. & Xu, G.L. (1994). Date mining corrections, *The Journal of Portfolio Management* **21**, 60–69.
- [25] Michaud, R.O. (1989). The Markowitz optimization enigma: is optimized optimal? *Financial Analysts Journal* **45**(1), 31–42.
- [26] Ohlson, J.A. (1979). Risk, return, security-valuation and the stochastic behavior of accounting numbers, *Journal of Financial and Quantitative Analysis* **14**(2), 317–336.
- [27] Pratt, J.W. (1964). Risk aversion in the small and in the large, *Econometrica* **32**, 122–136.
- [28] Rosenberg, B. (1974). Extra-market components of covariance in security returns, *Journal of Financial and Quantitative Analysis* **9**(2), 263–273.
- [29] Savage, L.J. (1954). *The Foundations of Statistics*, 2nd Revised Edition, John Wiley & Sons, Dover, New York.
- [30] Sharpe, W.F. (1963). A simplified model for portfolio analysis, *Management Science* **9**(2), 277–293.
- [31] Sortino, F. & Satchell, S. (2001). *Managing Downside Risk in Financial Markets: Theory, Practice and Implementation*, Butterworth-Heinemann, Burlington, MA.
- [32] Tobin, J. (1958). Liquidity preference as behavior towards risk, *Review of Economic Studies* **25**(1), 65–86.
- [33] Von Neumann, J. & Morgenstern, O. (1944, 1953). *Theory of Games and Economic Behavior*, 3rd Edition, Princeton University Press.

Further Reading

Ibbotson, R.G. & Sinquefeld, R.A. (2007). *Stocks, Bonds, Bills and Inflation Yearbook*, Morningstar, New York.

Related Articles

Behavioral Portfolio Selection; Black–Litterman Approach; Diversification; Expected Utility Maximization; Markowitz, Harry; Mean–Variance Hedging.

HARRY M. MARKOWITZ

Expected Utility Maximization

The continuous-time portfolio problem consists of maximizing total expected utility of consumption over the trading interval $[0, T]$ and/or of terminal wealth $X(T)$. We look at this problem in the standard continuous-time diffusion market setting. In other words, we consider a security market with $n + 1$ assets, one of which is a money market account with rate of return $r(t)$, and the other n of which are stocks, whose prices are driven by an m -dimensional Brownian motion. More precisely, the prices are given as the unique solutions of the equations

$$dP_0(t) = P_0(t)r(t) dt, \quad P_0(0) = 1 \quad (1)$$

$$dP_i(t) = P_i(t) \left(b_i(t) dt + \sum_{j=1}^m \sigma_{ij}(t) dW_j(t) \right), \quad (2)$$

$$P_i(0) = p_i$$

Here, we assume that the market coefficients r , b , and σ are progressively measurable with respect to the Brownian filtration, component-wise bounded, and that $\sigma\sigma'$ is uniformly positive definite, that is, we have

$$x'\sigma(t)\sigma(t)'x \geq c x'x \quad \text{a.s.} \quad (3)$$

for a positive constant c , and all $0 \neq x \in IR^n$, $t \in [0, T]$. This, in particular, implies

$$m \geq n \quad (4)$$

The investor specifies his/her investment and consumption strategy at time t by choosing the rate $c(t)$ at which he/she consumes and the different fractions $\pi_i(t)$ of his/her wealth that he/she invests in the risky asset i . The remaining fraction of his/her wealth $1 - \sum_{i=1}^n \pi_i(t)$ has to be invested in the money market account. We assume that the investor bases his investment decisions on the observation of past and present security prices. We therefore require both the consumption and the portfolio process to be progressively measurable with respect to the filtration $(f_t)_{t \in [0, T]}$ generated by the security prices. Of course, the consumption should always be nonnegative. We define the investor's wealth process $X^{\pi, c}(t)$ using the strategy (π, c) as the sum of the value of his/her total

holdings at time t . By requiring the investor to act in a self-financing way (i.e., the investor's wealth only changes due to gains/losses from trading and due to consumption), the return of the wealth is given as

$$\frac{dX^{\pi, c}(t)}{X^{\pi, c}(t)} = \left(1 - \sum_{i=1}^n \pi_i(t) \right) \frac{dP_0(t)}{P_0(t)} + \sum_{i=1}^n \pi_i(t) \frac{dP_i(t)}{P_i(t)} - \frac{c(t) dt}{X^{\pi, c}(t)} \quad (5)$$

leading to the stochastic differential equation (the "wealth equation")

$$dX^{\pi, c}(t) = X^{\pi, c}(t) \left(\left(r(t) + \sum_{i=1}^n \pi_i(t)(b_i - r) \right) dt + \sum_{i=1}^n \pi_i(t) \sum_{j=1}^m \sigma_{ij} dW_j(t) \right) - c(t) dt \quad (6)$$

with initial condition

$$X^{\pi, c}(0) = x \quad (7)$$

We call a self-financing pair (π, c) admissible for an initial wealth of x (and write $(\pi, c) \in A(x)$) if the wealth equation (6) has a unique positive solution with $c \geq 0$. By this, we implicitly assume that (π, c) satisfies integrability conditions ensuring that the stochastic integrals in equation (6) are well defined.

To judge the performance of such a pair, we introduce the notion of a utility function:

Definition 1 A strictly concave function $U : (0, \infty) \rightarrow IR$ with $U \in C^2$ satisfying

$$U'(0) := \lim_{x \downarrow 0} U'(x) = +\infty, \quad (8)$$

$$U'(+\infty) := \lim_{x \rightarrow +\infty} U'(x) = 0$$

is called a utility function.

Remarks and Examples

1. Note that the definition implies that a utility function has to be increasing ("more is better than less"), but the advantage of one additional unit decreases with increasing amount x ("decreasing marginal utility"). Popular examples of utility

functions are $U(x) = \frac{1}{\gamma}x^\gamma$ for $\gamma \in (0, 1)$ or $U(x) = \ln(x)$.

2. By slight misuse of the above definition, we will also call a family of functions $U(t, \cdot)$, $t \in [0, T]$ a utility function if for fixed $t \in [0, T]$ $U(t, \cdot)$ is a utility function as a function of the second variable. An obvious example would be the product of a utility function $F(\cdot)$ as in Definition 1 and a discount factor, for example, $U(t, x) = \exp(-\rho t)F(x)$.
3. A utility function is introduced to model the investors attitude toward risk. This can be seen by considering two investment alternatives. One results in the constant payment of A , the other one in a random payment B with $E(B) = A$, both at the same time T . Then by Jensen's inequality, we obtain $U(A) = U(E(B)) > E(U(B))$. Thus, an investor with such a utility function U would automatically go for the less risky alternative, the reason why his behavior (again characterized by the utility function) is called **risk-averse** (read more on this in Chapter 3 of [3]).

Definition 2 *The continuous-time portfolio problem of an investor with an initial wealth of x consists of maximizing his expected utility from final wealth and consumption on $[0, T]$ by choosing the best possible portfolio and consumption process, that is, by solving*

$$\max_{(\pi, c) \in A'(x)} E \left(\int_0^T U_1(t, c(t)) dt + U_2(X^{\pi, c}(T)) \right) \quad (9)$$

Here, U_1 and U_2 are utility functions. The restricted set $A'(x)$ consists of all those elements of $A(x)$ where the expectation over the negative part of the utility functions in equation (9) is finite (ensuring that the expected value in expression (9) is defined).

Remarks

1. Although $U \equiv 0$ is no utility function, we introduce the *pure consumption problem* by setting $U_2 \equiv 0$ and the *pure terminal wealth problem* by setting $U_1 \equiv 0$.
2. In the literature, there are mainly two different classes of methods to solve the continuous-time portfolio problem. The *stochastic control method* pioneered by Merton (see, e.g., [7, 8]) is based on the identification of the expected utility

maximization problem as a stochastic control problem. The optimal solution is then computed by solving the so-called Hamilton–Jacobi–Bellman equation of dynamic programming (see **Stochastic Control** for details on the stochastic control method). The second method, the *martingale method*, is tailored around the specific properties of the market model. We describe it in more details below.

The Martingale Method in the Complete Market Case

The martingale method goes back to [1, 2] and [9]. It relies on decomposing the dynamic portfolio problem (9) into two simpler subproblems, a static optimization problem and a representation problem. We demonstrate the martingale method in a complete market setting (i.e., we assume $m = n$). Here, every contingent claim (i.e., the sum of a nonnegative terminal payment B and a payout process $g(t)$) can be replicated by following a suitable portfolio strategy. Let us introduce the state-dependent discount process

$$H(t) := \exp \left(- \int_0^t (r(s) - 1/2 \|\theta(s)\|^2) ds - \int_0^t \theta(s)' dW(s) \right) \quad (10)$$

with $\theta(t) := \sigma(t)^{-1}(b(t) - r(t)\mathbf{1})$.

The completeness of the market allows to treat the problems of finding the optimal final wealth and finding the corresponding optimal portfolio process separately. More precisely, if a final payment of B and a consumption process $c(t)$ have an initial price satisfying

$$x = E \left(H(T)B + \int_0^T H(t)c(t) dt \right) \quad (11)$$

then a portfolio process $(\pi, c) \in A(x)$ resulting in a final wealth of B always exists.

We demonstrate this in the case of a pure terminal wealth problem (i.e., $U_1 = 0$): we decompose the portfolio problem into a *static optimization problem*

$$\max_{B \in B(x)} E(U(B)) \quad (12)$$

with $B(x) := \{B | B \geq 0, f_T\text{-measurable}, E(H(T)B) = x, E(U(B^-)) < \infty\}$ being the set of all contingent claims B maturing at T and having an initial price of x , and the *representation problem* to find a portfolio process $\pi^* \in A'(x)$ with

$$X^{\pi^*}(T) = B^* \quad a.s. \quad (13)$$

where B^* solves the static optimization problem (12).

Note that the optimization problem (12) is simpler than the original portfolio problem (9) as we only have to optimize over a set of distributions and not over a class of stochastic processes. It can be solved directly with the help of convex optimization and Lagrangian multiplier considerations, even in the general case including consumption. To state the explicit form of the optimal final wealth and consumption we need the following notation:

$$\begin{aligned} I_2(\lambda) &:= (U_2')^{-1}(\lambda), \\ I_1(t, \lambda) &:= \left(\frac{\partial}{\partial x} U_1 \right)^{-1}(t, \lambda) \end{aligned} \quad (14)$$

$$\begin{aligned} G(\lambda) &:= E \left(H(T) I_2(\lambda H(T)) \right. \\ &\quad \left. + \int_0^T H(t) I_1(t, \lambda H(t)) dt \right) \end{aligned} \quad (15)$$

Theorem 1 *Let $x > 0$. Under the assumption of*

$$G(\lambda) < \infty \quad \forall \lambda > 0 \quad (16)$$

the inverse function $\Lambda(x)$ to $G(\lambda)$ exists and the optimal terminal wealth B^ and the optimal consumption process $c^*(t)$, $t \in [0, T]$, are given by*

$$B^* = I_2(\Lambda(x)H(T)), \quad c^*(t) = I_1(t, \Lambda(x)H(t)) \quad (17)$$

Further, there exists a portfolio process $\pi^(t)$, $t \in [0, T]$ with $(\pi^*, c^*) \in A'(x)$,*

$$X^{\pi^*, c^*}(T) = B^*(T) \quad a.s. \quad (18)$$

and (π^, c^*) is the optimal strategy for the portfolio/consumption problem. Moreover, the optimal wealth process is given by*

$$X^{\pi^*, c^*}(t) = E \left(\frac{H(T)}{H(t)} B^* + \int_t^T \frac{H(s)}{H(t)} c^*(s) ds \middle| \mathcal{F}_t \right) \quad (19)$$

Remarks

1. While the optimization problem (12) is now explicitly solved up to the (possibly numerical) determination of the number $\Lambda(x)$, we have only stated existence of an optimal portfolio process π^* . Its explicit determination can be quite complicated and has to rely on very sophisticated methods such as Malliavin derivatives if the market coefficients are not constant and if the utility function is different from the logarithmic one of the example below.
2. We obtain the solution to the pure terminal wealth problem from Theorem 1 by setting U_1 and I_1 equal to zero. Setting U_2 and I_2 equal to zero, we get the optimal solution of the consumption problem.

Example 1 Log-Utility In the case of $U_1(t, x) = U_2(x) = \ln(x)$, we can perform all the calculations explicitly to obtain the optimal final wealth B^* and the optimal consumption c^* as

$$B^* = \frac{x}{T+1} \frac{1}{H(T)}, \quad c^*(t) = \frac{x}{T+1} \frac{1}{H(t)} \quad (20)$$

From this, the optimal wealth process as given in equation (19) can be determined as

$$\begin{aligned} X^{\pi^*, c^*}(t) &= E \left(\frac{H(T)}{H(t)} B^* + \int_t^T \frac{H(s)}{H(t)} c^*(s) ds \middle| \mathcal{F}_t \right) \\ &= \frac{x(T-t+1)}{(T+t)} \frac{1}{H(t)} \end{aligned} \quad (21)$$

By deriving the differential representation for it with the help of Itô's formula and comparing this representation with the general wealth equation (6), we can identify

$$\begin{aligned} c^*(t) &= \frac{1}{T+1-t} X^{\pi^*, c^*}(t), \\ \pi^*(t) &= (\sigma(t)\sigma(t)')^{-1}(b(t) - r(t)\mathbf{1}) \end{aligned} \quad (22)$$

Note that here the optimal consumption is proportional to the current wealth with an increasing proportionality factor ("Increase relative consumption with growing time.") The optimal fractions of wealth invested in the different securities only depend on the market coefficients and not on the wealth process itself. In particular, the fractions remain constant if the market coefficients are also constant.

The Martingale Method and Mean–Variance Optimization

In this section, we mention some possible generalizations of the martingale method. They will, in particular, allow us to treat a continuous-time version of the classical mean–variance problem (see [6] for the one-period model). For simplicity, we concentrate on the pure terminal wealth problem. Let us also relax the requirements on a utility function:

Definition 3 A strictly concave function $U : (0, \infty) \rightarrow \mathbb{R}$ with $U \in C^2$ satisfying

$$\begin{aligned} U'(0) &:= \lim_{x \rightarrow 0} U'(x) > 0, \\ U'(z) &= 0 \text{ for } a \text{ } z \in (0, +\infty] \end{aligned} \quad (23)$$

is called a (weak) utility function.

Remarks

1. The most popular examples of utility functions in the sense of Definition 3, which do not satisfy the conditions of Definition 1, are $U(x) = \alpha - \exp(-\lambda x)$ for $\alpha, \lambda > 0$, the exponential utility function, and $U(x) = -1/2(x - K)^2$ for $K > 0$, the quadratic utility function.
2. One can show (see [3], Chapter 3) that with the generalized inverse function

$$I(\lambda) := \begin{cases} (U')^{-1}(\lambda), & \lambda \in [0, U'(0)] \\ 0, & \lambda \geq U'(0) \end{cases} \quad (24)$$

the results of Theorem 1 remain true in the case that the investor cannot afford to reach the point of maximum utility for sure. This is always the case if we have

$$x < z \cdot E(H(T)) \quad (25)$$

Example 2 Quadratic Utility In the case of $U(x) = -1/2(x - K)^2$ and $x < z \cdot E(H(T))$ using Theorem 1 in the sense of the preceding remark leads to the optimal terminal wealth of

$$B^* = I(\Lambda(x)H(T)) = (K - \Lambda(x)H(T))^+ \quad (26)$$

where the number $\Lambda(x)$ is the unique solution of the equation

$$G(\lambda) = E(H(T)(K - \lambda H(T))^+) = x \quad (27)$$

In the particular case of constant market coefficients and only one stock, that is, $n = m = 1$ we can

calculate $G(\lambda)$ explicitly as

$$\begin{aligned} G(\lambda) &= K e^{-rT} \Phi\left(\frac{\ln(K/\lambda) + (r - 1/2\theta^2)T}{\theta\sqrt{T}}\right) \\ &\quad - \lambda e^{(-2r+\theta^2)T} \Phi\left(\frac{\ln(K/\lambda) + (r - 3/2\theta^2)T}{\theta\sqrt{T}}\right) \end{aligned} \quad (28)$$

with $\theta := (b_1 - r)/\sigma_{11}$. We can then easily solve equation (27) by numerical methods. One can compute the corresponding portfolio process π^* as indicated in [4]. This, however, is quite technical and requires the solution of a partial differential equation.

For dealing with constrained terminal wealth problems of the form

$$\max_{\pi \in A_T^*(x)} E(U(X^\pi(T))) \quad (29)$$

subject to $E(G_i(X^\pi(T))) \leq 0, i = 1, \dots, k$

with the functions G_i being real valued and convex, we introduce the Lagrangian function

$$\begin{aligned} L(\pi, d) &:= E((U - d'G)(X^\pi(T))), \\ \pi &\in A'(x), \quad d \in [0, \infty)^k \end{aligned} \quad (30)$$

Here, the subscript T of the set of admissible portfolio processes indicates that they also have to be admissible for the constrained problem, that is, the expectations in the constraints have to be defined and have to be satisfied (in addition to the usual requirement of a nonnegative wealth process). Then, in [5] under assumption (25) it is proved that—given the existence of a solution to the constrained problem (29)—we obtain a solution to this problem by using the following algorithm:

Step 1. Solve the unconstrained portfolio problem

$$\max_{\pi \in A'(x)} L(\pi, d) \quad (31)$$

by the martingale method for fixed, but arbitrary $d \in [0, \infty)^k$

Step 2. Minimize the function $L(\pi^*(d), d)$ with respect to $d \in [0, \infty)^k$ where $\pi^*(d)$ is an optimal portfolio process for the maximization problem of Step 1.

Example 3 Continuous-Time Mean–Variance Optimization As an application, we solve a continuous-time version of the famous Markowitz mean–

variance problem. The special feature of our solution is that we can guarantee nonnegativity of the final wealth process. This problem is intensively dealt with in the literature starting with [5] and generalized in a series of papers by Zhou *et al.* (see, e.g., [10]). Again, we assume that we are in the setting of constant market coefficients and $n = m = 1$. Then, one can show that under the assumption of

$$x < K e^{-rT} \quad (32)$$

the continuous-time mean–variance problem

$$\begin{aligned} \min_{\pi \in A_T'(y), y \leq x} \quad & \text{Var}(X^\pi(T)) \\ \text{subject to} \quad & E(X^\pi(T)) \geq K \end{aligned} \quad (33)$$

(with $K > 0$ a given constant) is equivalent to solving

$$\begin{aligned} \max_{\pi \in A_T'(x)} \quad & E(-1/2(X^\pi(T) - K)^2) \\ \text{subject to} \quad & K - E(X^\pi(T)) \leq 0 \end{aligned} \quad (34)$$

This, however, is a problem that can be solved by the above two-step algorithm. If in addition one realizes that we have

$$\begin{aligned} L(\pi, d) &= 1/2d^2 \\ &+ \max_{\pi \in A_T'(x)} E(-1/2(X^\pi(T) - (K + d))^2) \end{aligned} \quad (35)$$

we can use the calculations of the quadratic utility case to solve the unconstrained portfolio problems. A deterministic minimization in $d \geq 0$ then yields the solution to the continuous-time mean–variance problem under nonnegativity constraints on the terminal wealth.

The martingale method can also be applied to portfolio problems in incomplete markets. Then, however, convex duality methods have to be used to deal with constraints on the portfolio process or with the fact that the underlying market model is not

complete (see Chapter 4 of [3] for a survey on such constrained problems)

References

- [1] Cox, J. & Huang, C.F. (1989). Optimal consumption and portfolio policies when asset prices follow a diffusion process, *Journal of Economic Theory* **49**, 33–83.
- [2] Karatzas, I., Lehoczky, J.P. & Shreve, S.E. (1987). Optimal portfolio and consumption decisions for a “small investor” on a finite horizon, *SIAM Journal on Control and Optimization* **27**, 1157–1186.
- [3] Korn, R. (1997). *Optimal Portfolios*, World Scientific, Singapore.
- [4] Korn, R. (1997). Some applications of L^2 -hedging with a non-negative wealth process, *Applied Mathematical Finance* **4**, 65–79.
- [5] Korn, R. & Trautmann, S. (1995). Continuous-time portfolio optimization under terminal wealth constraints, *Zeitschrift für Operations Research* **42**, 69–92.
- [6] Markowitz, H. (1952). Portfolio selection, *Journal of Finance* **7**, 77–91.
- [7] Merton, R.C. (1969). Lifetime portfolio selection under uncertainty: the continuous case, *Reviews of Economical Statistics* **51**, 247–257.
- [8] Merton, R.C. (1971). Optimum consumption and portfolio rules in a continuous time model, *Journal of Economic Theory* **3**, 373–413.
- [9] Pliska, S. (1986). A stochastic calculus model of continuous trading: optimal portfolios, *Mathematics of Operations Research* **11**, 371–382.
- [10] Zhou, X. & Li, D. (2000). Continuous-time mean–variance portfolio selection: a stochastic LQ framework, *Applied Mathematics and Optimization* **42**, 19–33.

Related Articles

Expected Utility Maximization: Duality Methods; Mean–Variance Hedging; Merton Problem; Risk–Return Analysis; Stochastic Control; Utility Function.

RALF KORN

Black–Litterman Approach

The approach by Black and Litterman (BL) [2] blends a reference market distribution with subjective views on the market and yields allocations that smoothly reflect those views.

We present the original BL model and we review the related literature. A longer version of this article with all the proofs and more details is available at www.symmys.com > Research > Working Papers.

The Original Model

Here we follow [4], see also [5, 12] and [13].

The Market Model

We consider a market of N securities or asset classes, whose returns are normally distributed:

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (1)$$

The covariance $\boldsymbol{\Sigma}$ is estimated by exponential smoothing of the past return realizations. Since $\boldsymbol{\mu}$ cannot be known with certainty, BL model it as a normal random variable

$$\boldsymbol{\mu} \sim N(\boldsymbol{\pi}, \tau \boldsymbol{\Sigma}) \quad (2)$$

where $\boldsymbol{\pi}$ represents the best guess for $\boldsymbol{\mu}$ and $\tau \boldsymbol{\Sigma}$ the uncertainty on this guess.

To set $\boldsymbol{\pi}$, BL invoke an equilibrium argument. Assuming there is no estimation error, that is, $\tau \equiv 0$ in equation (2), the reference model (1) becomes

$$\mathbf{X} \sim N(\boldsymbol{\pi}, \boldsymbol{\Sigma}) \quad (3)$$

Assume that, consistent with this normal market, all investors maximize a mean–variance trade-off and that the optimization is unconstrained:

$$\mathbf{w}_\lambda \equiv \underset{\mathbf{w}}{\operatorname{argmax}} \{ \mathbf{w}' \boldsymbol{\pi} - \lambda \mathbf{w}' \boldsymbol{\Sigma} \mathbf{w} \} \quad (4)$$

The solution of this optimization problem yields the relationship between the equilibrium portfolio $\tilde{\mathbf{w}}$, which stems from an average risk-aversion level $\bar{\lambda}$ and the reference expected returns:

$$\boldsymbol{\pi} \equiv 2\bar{\lambda} \boldsymbol{\Sigma} \tilde{\mathbf{w}} \quad (5)$$

Therefore, $\boldsymbol{\pi}$ can be set in terms of $\tilde{\mathbf{w}}$, where BL set exogenously $\bar{\lambda} \approx 1.2$. Giacometti *et al.* [3] generalize this argument to stable-distributed markets. Notice that historical information does not play a direct role in the determination of $\boldsymbol{\pi}$: this is an instance of the shrinkage approach to estimation risk, see more details in the extended online version of this article.

To calibrate the overall uncertainty level τ in equation (2), we can compare this specification with the dispersion of the sample mean in a market where returns are distributed as in equation (3) independently across time: this implies $\tau \approx 1/T$. Satchell and Scowcroft [12] propose an ingenious model where τ is stochastic, but extra parameters need to be calibrated. In practice, a tailor-made calibration that spans the interval $(0, 1)$ is called for in most applications, see also the discussion in [13].

To illustrate, we consider the oversimplified case of an international stock fund that invests in the following six stock market national indexes: Italy, Spain, Switzerland, Canada, United States, and Germany. The covariance matrix of daily returns on the above classes $\boldsymbol{\Sigma}$ is estimated as follows in terms of the (annualized) volatilities $\boldsymbol{\sigma} \approx (21\%, 24\%, 24\%, 25\%, 29\%, 31\%)$ and the correlation matrix

$$\mathbf{C} \approx \begin{pmatrix} 1 & 54\% & 62\% & 25\% & 41\% & 59\% \\ \cdot & 1 & 69\% & 29\% & 36\% & 83\% \\ \cdot & \cdot & 1 & 15\% & 46\% & 65\% \\ \cdot & \cdot & \cdot & 1 & 47\% & 39\% \\ \cdot & \cdot & \cdot & \cdot & 1 & 38\% \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix} \quad (6)$$

To determine the prior expectation $\boldsymbol{\pi}$, we start from the market-weighted portfolio $\tilde{\mathbf{w}} \approx (4\%, 4\%, 5\%, 8\%, 71\%, 8\%)'$ and obtain from equation (5) the annualized expected returns $\boldsymbol{\pi} \approx (6\%, 7\%, 9\%, 8\%, 17\%, 10\%)'$. Finally, we set $\tau \approx 0.4$ in equation (2).

The Views

BL consider views on expectations. In the normal market (1), this corresponds to statements on the parameter $\boldsymbol{\mu}$. Furthermore, BL focus on linear views: K views are represented by a $K \times N$ “pick” matrix $\mathbf{P}$, whose generic k th row determines the relative weight of each expected return in the respective view. To associate uncertainty with the views, BL use a normal model:

$$\mathbf{P}\boldsymbol{\mu} \sim N(\mathbf{v}, \boldsymbol{\Omega}) \quad (7)$$

where the metaparameters $\mathbf{v}$ and $\mathbf{\Omega}$ quantify views and uncertainty thereof, respectively.

If the user has only qualitative views, it is convenient to set the entries of $\mathbf{v}$ in terms of the volatility induced by the market:

$$v_k \equiv (\mathbf{P}\boldsymbol{\pi})_k + \eta_k \sqrt{(\mathbf{P}\boldsymbol{\Sigma}\mathbf{P}')_{k,k}}, \quad k = 1, \dots, K \quad (8)$$

where $\eta_k \in \{-\beta, -\alpha, +\alpha, +\beta\}$ defines “very bearish”, “bearish”, “bullish”, and “very bullish” views, respectively. Typical choices for these parameters are $\alpha \equiv 1$ and $\beta \equiv 2$. Also, it is convenient to set as in [6]

$$\mathbf{\Omega} \equiv \frac{1}{c} \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}' \quad (9)$$

where the scatter structure of uncertainty is inherited from the market volatilities and correlations and c represents an overall level of confidence in the views.

To continue with our example, the manager might assess two views: the Spanish index will rise by 12% on an annualized basis, and the spread United States–Germany will experience a negative annualized change of 10%. Therefore, the pick matrix reads

$$\mathbf{P} \equiv \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \quad (10)$$

and the annualized views vector becomes $\mathbf{v} \equiv (12\%, -10\%)'$. We set the uncertainty in the views to be of the same order of magnitude as that of the market, that is, $c \equiv 1$ in equation (9).

The Posterior

With the above inputs, we can apply Bayes’ rule to compute the posterior market model:

$$\mathbf{X}|\mathbf{v}; \mathbf{\Omega} \sim \mathcal{N}(\boldsymbol{\mu}_{BL}, \boldsymbol{\Sigma}_{BL}) \quad (11)$$

where

$$\boldsymbol{\mu}_{BL} = \boldsymbol{\pi} + \tau \mathbf{\Sigma} \mathbf{P}' (\tau \mathbf{P} \boldsymbol{\Sigma} \mathbf{P}' + \mathbf{\Omega})^{-1} (\mathbf{v} - \mathbf{P} \boldsymbol{\pi}) \quad (12)$$

$$\boldsymbol{\Sigma}_{BL} = (1 + \tau) \mathbf{\Sigma} - \tau^2 \mathbf{\Sigma} \mathbf{P}' (\tau \mathbf{P} \boldsymbol{\Sigma} \mathbf{P}' + \mathbf{\Omega})^{-1} \mathbf{P} \mathbf{\Sigma} \quad (13)$$

See the proof in the extended online version of this article.

The normal posterior distribution (11) represents the modification of the reference model (3) that incorporates the views (7).

The Allocation

With the posterior distribution, it is now possible to set and solve a mean–variance optimization, possibly under a set of linear constraints, such as boundaries on securities/asset classes, or a budget constraint. This quadratic programming problem can be easily solved numerically. The ensuing efficient frontier represents a gentle twist to equilibrium that reflects the views without extreme corner solutions.

In our example, we assume the standard long-only and budget constraints, that is, $\mathbf{w} \geq \mathbf{0}$ and $\mathbf{w}' \mathbf{1} \equiv 1$. In Figure 1, we plot the efficient frontier from the reference model (3) and from the posterior model (11). Consistently with the views, the exposure to the Spanish market increases for lower risk values; the exposure to Germany increases across all levels of risk aversion; and the exposure to the US market decreases.

Related Literature

The BL posterior distribution (11) presents two puzzles. On one extreme, when the views are uninformative, that is, $\mathbf{\Omega} \rightarrow \infty$ in equation (7), one would expect the posterior to equal the reference model (3). On the other extreme, when the confidence in the views $\mathbf{v}$ is full, that is, $\mathbf{\Omega} \rightarrow \mathbf{0}$, one would expect the posterior to yield scenario analysis: the user inputs deterministic scenarios $\mathbf{v} \equiv (v_1, \dots, v_K)'$ for the factor combinations, resulting in a conditional distribution

$$\mathbf{X}|\mathbf{v} \sim \mathcal{N}(\boldsymbol{\mu}|\mathbf{v}, \boldsymbol{\Sigma}|\mathbf{v}) \quad (14)$$

where

$$\boldsymbol{\mu}|\mathbf{v} \equiv \boldsymbol{\pi} + \mathbf{\Sigma} \mathbf{P}' (\mathbf{P} \boldsymbol{\Sigma} \mathbf{P}')^{-1} (\mathbf{v} - \mathbf{P} \boldsymbol{\pi}) \quad (15)$$

$$\boldsymbol{\Sigma}|\mathbf{v} \equiv \mathbf{\Sigma} - \mathbf{\Sigma} \mathbf{P}' (\mathbf{P} \boldsymbol{\Sigma} \mathbf{P}')^{-1} \mathbf{P} \mathbf{\Sigma} \quad (16)$$

see, for example, [9].

The BL model does not satisfy the above two limit conditions. However, this issue can be fixed as in [6], by rephrasing the model in terms of views on the market $\mathbf{X}$, instead of the parameter $\boldsymbol{\mu}$. As we show in the extended online version of this article, the posterior then reads

$$\mathbf{X}|\mathbf{v}; \mathbf{\Omega} \sim \mathcal{N}(\boldsymbol{\mu}_{BL}, \boldsymbol{\Sigma}_{BL}) \quad (17)$$

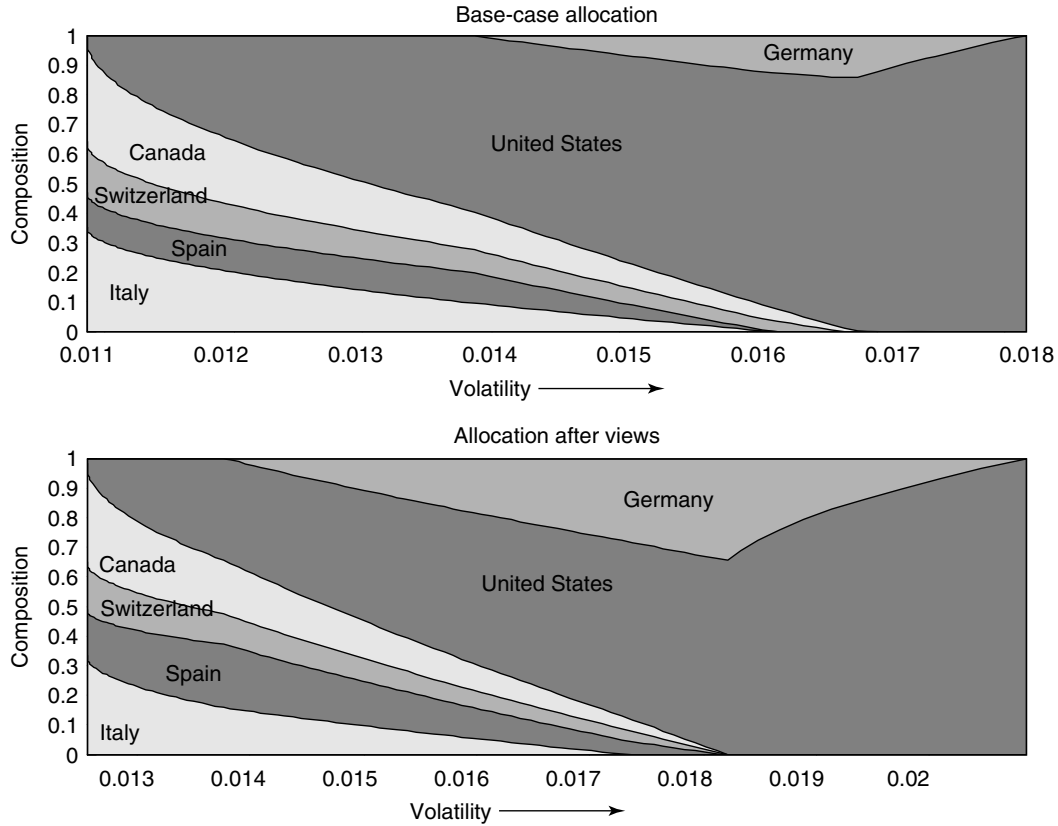


Figure 1 BL: efficient frontier twisted according to the views

where

$$\mu_{BL} \equiv \pi + \Sigma P'(\mathbf{P}\Sigma P' + \Omega)^{-1}(\mathbf{v} - \mathbf{P}\pi) \quad (18)$$

$$\Sigma_{BL} \equiv \Sigma - \Sigma P'(\mathbf{P}\Sigma P' + \Omega)^{-1}\mathbf{P}\Sigma \quad (19)$$

These formulas are very similar to their counterparts (12)–(13) in the original BL model. However, the parameter τ in equation (2) is no longer present and the two limit conditions are now satisfied.

Therefore, scenario analysis processes views on market expectations with infinite confidence and the BL model overlays uncertainty to the views by means of Bayesian formulas. Qian and Gorman [11] use a conditional/marginal factorization to input views on volatilities and correlations in addition to expectations. Pezier [10] processes full and partial views on expectations and covariances by least discrimination. Almgren and Chriss [1] provide a framework to express ranking, “lax” views on expectations.

The posterior formulas (12)–(13), their modified versions (18)–(19), as well as the formulas in the above literature can be applied to any *normal* distribution, not necessarily the equilibrium (5). Accordingly, Meucci [7] applies the above approaches to fully generic risk factors that map nonlinearly into the final P&L, instead of securities returns.

To further extend all the above approaches to nonnormal markets, as well as fully general nonlinear views from possibly multiple users, Meucci [8] uses entropy minimization and opinion pooling: since no costly repricing is ever necessary, this technique covers even the most complex derivatives.

Acknowledgments

The author gratefully acknowledges the very helpful feedback from Bob Litterman, Ninghui Liu, and Jay Walters.

References

- [1] Almgren, R. & Chriss, N. (2006). Optimal Portfolios from Ordering Information, *Journal of Risk* **9**, 1–47.
- [2] Black, F. & Litterman, R. (1990). Asset allocation: combining investor views with market equilibrium, *Goldman Sachs Fixed Income Research* September.
- [3] Giacometti, M., Bertocchi, I., Rachev, T.S. & Fabozzi, F. (2007). Stable distributions in the Black-Litterman approach to asset allocation, *Quantitative Finance* **7**, 423–433.
- [4] He, G. & Litterman, R. (2002). *The Intuition Behind Black-Litterman Model Portfolios*. ssm.com.
- [5] Idzorek, T.M. (2004). *A Step-by-Step Guide to the Black-Litterman Model*, Zephyr Associates Publications.
- [6] Meucci, A. (2005). *Risk and Asset Allocation*, Springer.
- [7] Meucci, A. (2009). Enhancing the Black-Litterman and Related Approaches: Views and Stress-Test on Risk Factors, *Journal of Asset Management* **10**(2), 89–96.
- [8] Meucci, A. (2008). Fully flexible views: theory and practice, *Risk* **21**, 97–102. Available at symmys.com > Research > Working Papers.
- [9] Mina, J. & Xiao, J.Y. (2001). *Return to RiskMetrics: The Evolution of a Standard*, Risk Metrics Publications.
- [10] Pezier, J. (2007). *Global Portfolio Optimization Revisited: A Least Discrimination Alternative to Black-Litterman*. ICMA Centre Discussion Papers in Finance.
- [11] Qian, E. & Gorman, S. (2001). Conditional distribution in portfolio theory, *Financial Analyst Journal* **57**, 44–51.
- [12] Satchell, S. & Scowcroft, A. (2000). A demystification of the Black-Litterman model: managing quantitative and traditional construction, *Journal of Asset Management* **1**, 138–150.
- [13] Walters, J. (2008). *The Black-Litterman Model: A Detailed Exploration*. blacklitterman.org.

Related Articles

Capital Asset Pricing Model; Risk–Return Analysis.

ATTILIO MEUCCI

Fixed Mix Strategy

There are a number of advantages of adopting multi-period models over the traditional single-period, static models in portfolio managements [12]. One of the more important benefits, among others, is the improved performance on portfolio investments *via* the fixed mix rule [3–5]. The buy-and-hold rule, which represents single-period models, does not rebalance the portfolio at any intermediate juncture; hence, the weight on each component might change as asset prices fluctuate in different proportions. In contrast, when a portfolio is constructed based on the fixed mix rule, it is rebalanced at every time point so that component weights remain the same as the initial state. To keep the weights unchanged, investors should sell assets whose prices have gone up and buy ones whose prices have dropped. Therefore, in some sense, the fixed mix rule is analogous to the “buy low/sell high” strategy. Possibly because of such an analogy, there seems to be a widely spread misconception regarding the fixed mix strategy and its benefits—it requires mean-reverting processes for assets. Of course, because of its nature, it is not difficult to see that it would be helpful to have such processes to achieve better performance. However, the truth is that mean reversion is *not* necessary for the fixed mix to accomplish superior performance.

Theoretical Background

We first recall performance of the buy-and-hold strategy. Suppose that there are n stocks whose mean return is $r \in \mathbf{R}^n$ and covariance matrix $\Sigma \in \mathbf{R}^{n \times n}$. Assuming that normality, r^{BH} the average buy-and-hold portfolio return with weight $w \in \mathbf{R}^n$, is normally distributed with mean $w^T r$ and variance $\sigma_p^2 = w^T \Sigma w$. That is,

$$r^{\text{BH}} \sim \mathbf{N}(w^T r, \sigma_p^2) \equiv \mathbf{N}(w^T r, w^T \Sigma w) \quad (1)$$

Next, let us consider a fixed mix portfolio constructed from the same stocks with the same weight (w) as the previous buy-and-hold portfolio. Since it is rebalanced at every intermediate juncture, it is required to model stock prices as processes. Thus, we model them as an n -dimensional geometric Brownian motion whose return distribution for a unit time

length would be the same as the previous case. Then, the price process of stock i can be written as the following SDE:

$$\frac{dS_t^i}{S_t^i} = \left(r_i + \frac{\sigma_i^2}{2} \right) dt + dB_t^i \quad (2)$$

where σ_i^2 is the i th diagonal term of Σ (hence, variance of stock i) and for the Cholesky factorization of Σ , L and the standard n -dimensional Wiener process $(W_t^1 - W_t^n)^T$,

$$d(B_t^1 - B_t^n)^T = L d(W_t^1 - W_t^n)^T \quad (3)$$

Since the fixed mix portfolio is rebalanced at each time point to the initial weight (w), its instantaneous growth rate is the same as the weighted sum of instantaneous growth rates of the stocks at any given juncture. Therefore, the SDE for the portfolio wealth can be written as

$$\frac{dP_t^{\text{FM}}}{P_t^{\text{FM}}} = \sum_{i=1}^n w_i \frac{dS_t^i}{S_t^i} = \sum_{i=1}^n w_i \left\{ \left(r_i + \frac{\sigma_i^2}{2} \right) dt + dB_t^i \right\} \quad (4)$$

With simple algebra, one can show that, for the standard one-dimensional Wiener process W_t ,

$$\frac{dP_t^{\text{FM}}}{P_t^{\text{FM}}} = \left(w^T r + \frac{1}{2} \sum_{i=1}^n w_i \sigma_i^2 \right) dt + \sigma_p^2 dW_t \quad (5)$$

Hence, the return of the fixed mix portfolio for a unit time length can be given as

$$\begin{aligned} r^{\text{FM}} &\sim N \left(w^T r + \frac{1}{2} \sum_{i=1}^n w_i \sigma_i^2 - \frac{1}{2} \sigma_p^2, \quad \sigma_p^2 \right) \\ &\equiv N \left(w^T r + \frac{1}{2} \sum_{i=1}^n w_i \sigma_i^2 - \frac{1}{2} w^T \Sigma w, w^T \Sigma w \right) \end{aligned} \quad (6)$$

Therefore, returns of both buy-and-hold (r^{BH}) and fixed mix (r^{FM}) are normally distributed with the same variance (σ_p^2), whereas the mean of the latter contains extra terms ($\sum_{i=1}^n w_i \sigma_i^2 - \sigma_p^2/2$). These extra terms, which are often referred to as *rebalancing gains* or *volatility pumping*, represent the value of having an option to constantly rebalance the portfolio to initial weights.

To observe its effects more closely, let us consider the following simple example: suppose that we have n stocks where the expected return and the volatility of each are r and σ , and the correlation is given as ρ . Assuming the portfolio is equally weighted, the amount of the rebalancing gain

$$RG = \frac{1}{2} \left\{ \sum_{i=1}^n \frac{1}{n} \sigma^2 - \left(\frac{1}{n} - \frac{1}{n} \right) \Sigma \left(\frac{1}{n} - \frac{1}{n} \right)^T \right\} \\ = \frac{(n-1)\sigma^2(1-\rho)}{2n} \quad (7)$$

Now it is evident that the fixed mix strategy has benefit over the static buy-and-hold rule, *even without mean-reversion*; the rebalancing gain is always positive, except the case that all stock returns are perfectly correlated, in which it becomes 0. Note that the rebalancing gain is an increasing function of the number of stocks (n) and the volatility (σ) and is a decreasing function of the correlation (ρ). See Figure 1 for the illustrations of simulation results for the effects of σ and ρ to rebalancing gains. Therefore, with the wisdom from the portfolio theory, one can see that volatile stocks should *not* be penalized when a portfolio is constructed with the fixed mix rule, as long as their correlations to other stocks are low and they possess reasonable expected returns; they can serve as good sources of rebalancing gains. The portfolio risks can be effectively reduced *via* dynamic diversification. For more complete discussion, see [3–5, 10, 15].

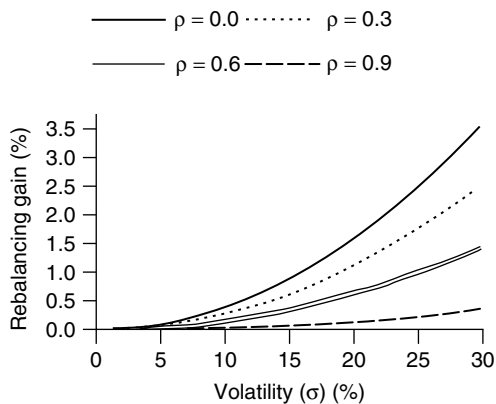


Figure 1 Effects of volatility (σ) and correlation (ρ) to rebalancing gains ($n = 5$)

Practical Examples

Under certain conditions, the fixed mix rule has been proved to be optimal in multiperiod settings. Early on, Mossin [7] showed that it is the optimal strategy when an investor maximizes the expected power utility of her terminal wealth, assuming IID asset returns and no intermediate consumption. Samuleson [17] analyzed the problem in more generalized settings: using an additive intertemporal power utility function of consumption over time, he proved that it is still optimal to adopt the fixed mix rule when the investor is allowed to consume at intermediate junctures. Merton [6] also concluded the same in the continuous time setting model.

Indeed, there are many practical applications that are successfully taking advantage of rebalancing gains by employing fixed mix rules. Among others [9, 14, 16], one of good examples is the S&P 500 equal-weighted index (S&P EWI) by Rydex Investments (Figure 2) [8]. Unlike traditional cap-weighted S&P 500 index, it applies the fixed mix rule to the same stocks as S&P 500, rebalancing them every six month to maintain the equally weighted portfolio. During 1994–2005, S&P Equal Weighted Index earned 2% excess return with mere 0.6% extra volatility over S&P 500. This added profit is partially due to superior performance of small/mid-sized stocks and also can be accounted for by rebalancing gains.

Implementations of the fixed mix rule could also lead to successful leverage. Figure 3 illustrates levered portfolios of buy-and-hold and fixed

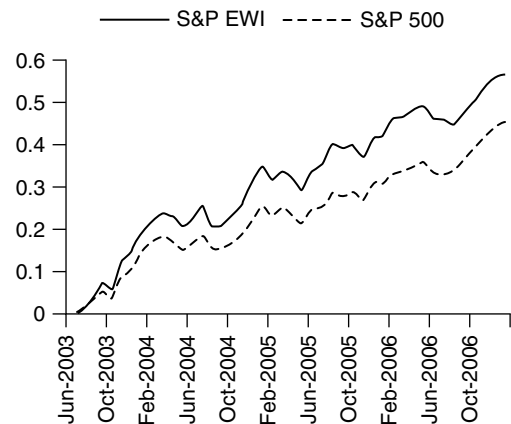


Figure 2 Log prices of S&P 500 and S&P EWI during July 2003 to December 2006

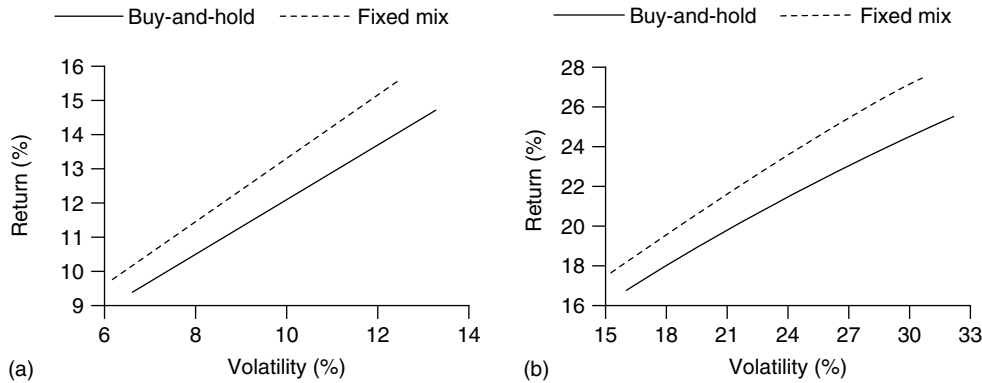


Figure 3 Efficient frontiers of levered buy-and-hold and fixed mix portfolios: (a) mix of traditional and alternative assets (1994–2005) and (b) mix of momentum strategies of five regions (1980–2006)

mix portfolios constructed in two different domains. Figure 3(a) compares efficient frontiers of buy-and-hold and fixed mix portfolios, which are constructed with six traditional assets (S&P 500, EAFE, Lehman long-term bond index, Strips, NAREIT, and Goldman Sachs commodity index) and four alternative assets (hedge fund index, managed futures index, Tremont long-short equity index, and currency index) [11]. Both are equally weighted and levered up to 100% *via* t-bill rate. Although the buy-and-hold portfolio is not rebalanced, monthly rebalancing rule is adopted for the fixed mix for the entire sample period (1994–2005). In addition, Figure 3(b) depicts results from portfolios of industry-level momentum strategies across international stock markets for 27-year sample period (1980–2006) [10]. Momentum strategies are constructed in five nonoverlapping regions (US, EU, Europe except EU, Japan, and Asia except Japan) and aggregated into equally weighted portfolios with leverage up to 100%. Similar to the previous case, the fixed mix portfolio is rebalanced monthly. In both the cases, the efficient frontiers from the fixed mix dominate ones from the buy-and-hold.

Implementation Issues

The fixed mix rule is now becoming a norm in various financial domains. For instance, it is now commonplace for large pension plans, such as TIAA-CREF, to automatically rebalance client-selected portfolios back to client-selected weights, at the client's requests. Given the circumstances, it is imperative to address issues regarding practical implementations.

First, since the best sources of rebalancing gains are volatile financial instruments with low intracorrelations, it is crucial to find a set of relatively independent assets. However, the task is very unlikely to be perfectly achieved in the real world. Second, even such a set exists at certain time point, correlations could change over time. For instance, it is well known that stock indices across international markets become highly correlated upon serious market distress. In addition, one should consider transaction costs such as capital gain taxes upon deciding the rebalancing intervals. Although frequent rebalancing could lead to investment performance close to the theoretical values, it may deteriorate performance due to transaction costs. Careful analysis on this trade-off is required. Good references regarding practical implementations of the fixed mix rules include [1, 2, 13, 18].

References

- [1] Davis, M.H.A. & Norman, A.R. (1990). Portfolio selection with transaction costs, *Mathematics of Operations Research* **15**, 676–713.
- [2] Dumas, B. & Luciano, E. (1991). An exact solution to a dynamic portfolio choice problem under transaction costs, *Journal of Finance* **46**, 577–595.
- [3] Fernholz, R. (2002). *Stochastic Portfolio Theory*, Springer-Verlag, New York.
- [4] Fernholz, R. & Shay, B. (1982). Stochastic portfolio theory and stock market equilibrium, *Journal of Finance* **37**, 615–624.
- [5] Luenberger, D. (1997). *Investment Science*; Oxford University Press, New York.

- [6] Merton, R.C. (1969). Lifetime portfolio selection under uncertainty: the continuous-time case, *Review of Economics Statistics* **51**, 247–257.
- [7] Mossin, J. (1968). Optimal multi-period portfolio policies, *Journal of Business* **41**, 215–229.
- [8] Mulvey, J.M. (2005). *Essential Portfolio Theory*, A Rydex Investment White Paper (also Princeton University Report), 14–17.
- [9] Mulvey, J.M., Gould, G. & Morgan, C. (2000). An asset and liability management system for Towers Perrin-Tillinghast, *Interfaces* **30**, 96–114.
- [10] Mulvey, J.M. & Kim, W.C. (2007). *Constructing a Portfolio of Industry-level Momentum Strategies Across Global Equity Markets*, Princeton University Report.
- [11] Mulvey, J.M. & Kim, W.C. (2007). The role of alternative assets in portfolio construction, *Encyclopedia of Quantitative Risk Assessment*, John Wiley & Sons, to be published.
- [12] Mulvey, J.M., Pauling, B. & Madey, R.E. (2003). Advantages of multi-period portfolio models, *Journal of Portfolio Management* **29**, 35–45.
- [13] Mulvey, J.M. & Simsek, K.D. (2002). Rebalancing strategies for long-term investors. *Computational Methods in Decision-Making, Economics and Finance: Optimization Models*, E.J. Kontoghiorghes, B. Rustem & S. Siokos, eds, Kluwer, pp. 15–33.
- [14] Mulvey, J.M. & Thorlacius, A.E. (1998). The Towers Perrin global capital market scenario generation system: CAP Link, in *World Wide Asset and Liability Modeling*, W. Ziemba & J. Mulvey, eds, Cambridge University Press, Cambridge, pp. 286–312.
- [15] Mulvey, J.M., Ural, C. & Zhang, Z. (2007). hbox-Improving performance for long-term investors: wide diversification, leverage, and overlay strategies, *Quantitative Finance* **7**, 175–187.
- [16] Perold, A.F. & Sharpe, W.F. (1998). Dynamic strategies for asset allocation, *Financial Analysts Journal* **44**, 16–27.
- [17] Samuelson, P.A. (1969). Lifetime portfolio selection by dynamic stochastic programming, *Review of Economics Statistics* **51**, 239–246.
- [18] Shreve, S.E. & Soner, H.M. (1991). Optimal investment and consumption with two bonds and transaction costs, *Mathematical Finance* **1**, 53–84.

Further Reading

Mulvey, J.M., Kaul, S.S.N. & Simsek, K.D. (2004). Evaluating a trend-following commodity index for multi-period asset allocation, *Journal of Alternative Investments* **7**, 54–69.

Related Articles

Diversification; Expected Utility Maximization; Mutual Funds; Transaction Costs.

JOHN M. MULVEY & WOO CHANG KIM

Stochastic Control

The theory of stochastic optimal control concerns the control of a dynamical system in the presence of random noises so as to optimize a certain performance criterion. The early development of stochastic optimal control theory began in the period between the late 1950s and early 1960s. The earlier stage of the development focused on the quadratic performance criterion. Around the same period, the classic work of Bellman [1] first introduced one of the main approaches in stochastic optimal control theory, namely, the dynamic programming principle. The dynamic programming approach plays a significant role in modern finance, in particular, continuous-time finance. The key idea of the dynamic programming principle is to consider a family of stochastic optimal control problems with different starting times and states and to relate the problems through the Hamilton–Jacobi–Bellman (HJB) equation. An HJB equation is a nonlinear second-order partial differential equation (*see Partial Differential Equations*) that describes the local behavior of the performance criterion evaluated at the optimal control. For detailed discussion of the dynamic programming approach, we refer to [11, 12, 21].

Together with the HJB approach, the stochastic maximum principle provides the second main approach to stochastic control. The key idea of the stochastic maximum principle is to derive a set of necessary conditions satisfied by any optimal control. The stochastic maximum principle basically states that any optimal control must satisfy forward–backward stochastic differential equations (SDEs), called the *optimality system*, and a maximum condition of a functional, called the *Hamiltonian*. The novelty of the stochastic maximum principle is to make the stochastic optimal control problem, which is infinite dimensional, more tractable. It leads to explicit solutions for the optimal controls in some cases. References [2] and [21] provided excellent discussions on the stochastic maximum principle.

Merton [16, 17] pioneered the study of an optimal consumption–investment problem in a continuous-time economy. He first explored the state of the art of the stochastic optimal control theory to develop an elegant (closed-form) solution to the problem (*see Merton Problem*). The stochastic control approach adopted by Merton uses the HJB equation. Another

approach is the martingale approach, which uses the martingale method for risk-neutral valuation of options to provide an elegant solution to the optimal consumption–investment problem. The martingale approach was pioneered by the important contributions of Cox and Huang [4] and Karatzas *et al.* [13]. It was then extended by a number of authors, (*see, e.g., [14]*).

Each of the three main approaches in stochastic optimal control has its own merits. For example, dynamic programming works well in the case when (i) the state processes and optimal controls are Markov (*see Markov Processes*), (ii) the state processes have deterministic coefficients, and (iii) state constraints are absent. The stochastic maximum principle can deal with the situations when the state processes have random coefficients and state constraints are present. The martingale approach is applicable when one considers a general state process, for example, when the state process is not Markov. It works well when the market is complete though there are some works that consider the case when the market is incomplete (*see [14]*). The martingale approach is suitable for the situations when there are nonnegative constraints on consumption and wealth. It is difficult to say which one is uniformly better or more general than the other two. However, the three approaches are related to each other in some way. For example, the relationship between the dynamic programming approach and the stochastic maximum principle can be established by relating the solutions of the forward–backward SDEs associated with the stochastic maximum principle to those of the HJB equation from the dynamic programming ([21], Chapter 5). The relationship between the martingale approach and the stochastic maximum principle stems from two facts. Firstly, the solutions of the adjoint equations can be related to the density process for the change of measures in the martingale approach. Secondly, the first-order condition of the constrained maximization problem in the martingale approach is related to that of the first-order condition of the Hamiltonian in the stochastic maximum principle [3]. It is interesting to note that the three approaches end up with the same result of the optimal consumption–investment problem in some cases.

We discuss three methods, namely, the martingale method, the HJB Method, and the stochastic maximum principle to solve the optimal consumption–investment problem. Here, we focus on the case

of a power utility function. For general cases, refer to [10, 14].

The Martingale Approach

The development here is based on the contributions of Karatzas, Lehoczky, Sethi, Shreve, and Xu [13]. Here, we just present some main results and highlight some key steps. For a more comprehensive discussion, we refer to [10; Chapter 10].

We consider a popular model for a financial market consisting of one risk-free asset and n risky assets. These assets are tradable over a finite time horizon $[0, T^*]$, where $T^* < \infty$. Fix a complete probability space $(\Omega, \mathcal{F}, \mathcal{P})$, where $\mathcal{P}$ is a real-world probability measure.

The dynamics of the risk-free asset or bond, B , and the risky assets $S_1, S_2, \dots, S_n$, under $\mathcal{P}$, are governed by

$$dB(t) = r(t)B(t) dt, \quad B(0) = 1 \quad (1)$$

$$dS_i(t) = S_i(t) \left(\mu_i(t) dt + \sum_{j=1}^n \sigma_{ij}(t) dW_j(t) \right) \quad (2)$$

$$S_i(0) = s_i, \quad i = 1, 2, \dots, n \quad (3)$$

Here $W(t) := (W_1(t), W_2(t), \dots, W_n(t))^T$ is an n -dimensional Brownian motion defined on $(\Omega, \mathcal{F}, \mathcal{P})$, where y^T is the transpose of a vector y . Write $\{\mathcal{F}(t)\}$ for the right-continuous and complete filtration generated by $\{W(t)\}$. For a treatment of SDEs, see [8].

The market interest rate $r(t)$, the vector of appreciation rates $\mu(t) := (\mu_1(t), \mu_2(t), \dots, \mu_n(t))^T$, and the volatility matrix $\sigma(t) := [\sigma_{ij}(t)]_{i,j=1,2,\dots,n}$ of the risky assets are supposed to be measurable, $\{\mathcal{F}(t)\}$ -adapted, and bounded processes. The market is complete.

Let $a(t) := \sigma(t)\sigma^T(t)$. Suppose there is an $\epsilon > 0$ such that

$$\xi^T a(t) \xi \geq \epsilon |\xi|^2, \quad \forall \xi \in \mathbb{R}^n, \quad (t, \omega) \in [0, T^*] \times \Omega,$$

where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^n$.

Then, the inverses of σ and σ^T exist and are bounded, and the market is complete. The filtration $\{\mathcal{F}(t)\}$ is equivalent to the $\mathcal{P}$ -completion of the filtration generated by the price process $\{S(t)\}$.

We define the market price of risk by

$$\theta(t) := \sigma^{-1}(t)(\mu(t) - r(t)\mathbf{1}) \quad (4)$$

where $\mathbf{1} := (1, 1, \dots, 1)^T \in \mathbb{R}^n$; θ is bounded and $\{\mathcal{F}(t)\}$ -progressively measurable.

Now, we introduce an exponential process:

$$\Lambda(t) := \exp \left(- \int_0^t \theta^T(s) dW(s) - \frac{1}{2} \int_0^t |\theta^T(s)|^2 ds \right) \quad (5)$$

Define a new probability measure $\mathcal{P}^\theta \sim \mathcal{P}$ on $\mathcal{F}(T^*)$ by setting

$$\frac{d\mathcal{P}^\theta}{d\mathcal{P}} := \Lambda(T^*) \quad (6)$$

By Girsanov's theorem,

$$W^\theta(t) := W(t) + \int_0^t \theta(s) ds \quad (7)$$

is an n -dimensional standard Brownian motion under $\mathcal{P}^\theta$.

Also, under $\mathcal{P}^\theta$,

$$dS_i(t) = S_i(t) \left(r(t) dt + \sum_{j=1}^n \sigma_{ij}(t) dW_j^\theta(t) \right) \quad (8)$$

Here $\mathcal{P}^\theta$ is called the *risk-neutral* or *equivalent martingale measure*.

Consider a power utility as below:

$$U(c) = \frac{c^\gamma}{\gamma}, \quad 0 < \gamma < 1 \quad (9)$$

where γ represents the risk-aversion parameter. The relative degree of risk aversion is $1 - \gamma$, which indicates how risk averse the investor is. The higher $1 - \gamma$ is, the more risk averse the investor is.

Let U' denote the first derivative of $U(c)$ with respect to c . Write $I(\cdot)$ for the inverse of $U'(\cdot)$. Then, for any $y \in (0, \infty)$,

$$I(y) = y^{1/(\gamma-1)} \quad (10)$$

Consider a measurable $\mathbb{R}^n$ -valued, $\{\mathcal{F}(t)\}$ -adapted process $\pi := (\pi_1, \pi_2, \dots, \pi_n)^T$ and a consumption

process $\{c(t)\}$ as a nonnegative, measurable, $\{\mathcal{F}(t)\}$ -adapted process such that

$$\int_0^{T^*} (c(t) + |\pi(t)|^2) dt < \infty, \quad \mathcal{P} \text{ a.s.} \quad (11)$$

Here $\pi_i(t)S_i(t)$ represents the amount invested in the i th risky asset, for $i = 1, 2, \dots, n$. So, π is called a *portfolio process* or a *trading strategy*. Note that the adapted condition of (π, c) implies that the investor cannot anticipate the future. One financial implication is that “insider trading” is not allowed.

We assume that π is self-financing. A trading strategy is said to be self-financing if the changes in the value of the wealth result entirely from net gains or losses from the investments in the risk-free asset and the risky assets. In other words, there is no net inflow or outflow of funds. Let $\{V(t)\}$ denote the wealth process of the investor, where $V(t) := \sum_{i=1}^n \pi_i(t)S_i(t) + (1 - \sum_{i=1}^n \pi_i(t))B(t) - \int_0^t c(s) ds$. Then, under $\mathcal{P}$, the evolution of $\{V(t)\}$ is governed by

$$dV(t) = \sum_{i=1}^n \pi_i(t) dS_i(t) + \left(1 - \sum_{i=1}^n \pi_i(t)\right) dB(t) - c(t) dt \quad (12)$$

Let $\beta(t) := B^{-1}(t) = \exp\left(-\int_0^t r(u) du\right)$. Then, under $\mathcal{P}^\theta$, the evolution of the discounted wealth process $\{\beta(t)V(t)\}$ is governed by

$$\begin{aligned} \beta(t)V(t) &= v - \int_0^t \beta(s)c(s) ds \\ &\quad + \int_0^t \beta(s)\pi^T(s)\sigma(s) dW^\theta(t) \end{aligned} \quad (13)$$

where $v = V(0)$ represents the initial wealth of the investor.

Let $\mathcal{A}(v)$ denote the class of control processes (π, c) , for the initial wealth v , with wealth process satisfying equation (13) and that the wealth process $\{V(t)\}$ is nonnegative at all times in $[0, T^*]$.

It can be shown that for any $(\pi, c) \in \mathcal{A}(v)$,

$$E^\theta \left[\int_0^{T^*} \beta(t)c(t) dt \right] \leq v \quad (14)$$

The utility maximization problem is to select $(\pi, c) \in \mathcal{A}(v)$ so as to maximize the expected discounted utility *see Expected Utility Maximization, Expected Utility Maximization: Duality Methods* from consumption over $[0, T^*]$:

$$J(v, \pi, c) := E \left[\int_0^{T^*} \frac{(c(t))^\gamma}{\gamma} dt \right] \quad (15)$$

Note that the utility only depends on the consumption level. So, in order to maximize the utility from consumption, we should increase the consumption level up to a certain bound. In other words, we only consider the consumption processes c such that $E^\theta \left[\int_0^{T^*} \beta(t)c(t) dt \right] = v$. The utility maximization problem is to solve the following maximization problem:

$$J(v, \hat{\pi}, \hat{c}) = \sup_{(\pi, c) \in \mathcal{A}(v)} J(v, \pi, c) \quad (16)$$

subject to the budget constraint

$$E \left[\int_0^{T^*} \Lambda(t)\beta(t)c(t) dt \right] = v \quad (17)$$

To solve the problem, we need to find the optimal portfolio process $\hat{\pi}$, the optimal consumption process $\hat{c}$, and the value function defined by $\Phi(v) := J(v, \hat{\pi}, \hat{c})$.

Let λ denote the Lagrange multiplier of the constrained maximization problem (3). Then, the first-order conditions of the maximization problem imply that the optimal consumption rate $c^*(t)$ satisfy

$$c^*(t) = (\lambda \Lambda(t)\beta(t))^{1/(\gamma-1)} \quad (18)$$

$$E \left[\int_0^{T^*} \Lambda(t)\beta(t)c^*(t) dt \right] = v \quad (19)$$

So, the optimal consumption process is

$$c^*(t) = (\lambda \Lambda(t)\beta(t))^{1/(\gamma-1)}, \quad \forall t \in [0, T^*] \quad (20)$$

with the Lagrange multiplier λ determined by

$$\lambda = \left(\frac{v}{E \left[\int_0^{T^*} (\Lambda(t)\beta(t))^{\gamma/(\gamma-1)} dt \right]} \right)^{\gamma-1} \quad (21)$$

Since the market is complete, the optimal wealth process $\{V^*(t)\}$ is given by

$$\begin{aligned} V^*(t) &= E^\theta \left[\int_0^{T^*} c^*(s) \beta(s) ds | \mathcal{F}(t) \right] \\ &= \frac{1}{\Lambda(t)} E \left[\int_0^{T^*} \Lambda(s) c^*(s) \beta(s) ds | \mathcal{F}(t) \right] \end{aligned} \quad (22)$$

The HJB Method

In this section, we illustrate the use of the dynamic programming approach, also called the *HJB method*, to solve the optimal consumption–investment problem described in the section The Martingale Approach. We consider the asset price dynamics in the section The Martingale Approach with time-dependent coefficients replaced by constant coefficients. We impose the same assumptions and notation for control policies, utility function, and probability measures as those in the section The Martingale Approach, unless otherwise stated.

We consider the same problem as that in the last section on the interval $[t, T^*]$ instead of $[0, T^*]$. For any $t \in [0, T^*]$, we consider admissible policies $(\pi, c) \in \mathcal{A}(t, v)$ for which the wealth process $\{V(t)\}$ satisfies

$$\begin{aligned} &e^{-ru} V(u) \\ &= ve^{-rt} - \int_t^u e^{-rs} c(s) ds \\ &\quad + \int_t^u e^{-rs} \pi^T(s) \sigma dW^\theta(s) \quad \forall u \in [t, T^*] \end{aligned} \quad (23)$$

Here we require that the wealth process $\{V(t)\}$ is nonnegative at all times in $[0, T^*]$.

The value function, which is an indirect utility *see Expected Utility Maximization, Expected Utility Maximization: Duality Methods* function, is defined by

$$\Phi(t, v) := \sup_{(\pi, c) \in \mathcal{A}(t, v)} E \left[\int_t^{T^*} \frac{(c(s))^\gamma}{\gamma} ds | \mathcal{F}(t) \right] \quad (24)$$

Let Φ_t , Φ_v , and Φ_{vv} denote the derivative of Φ with respect to t , the first and second derivatives of Φ with

respect to v respectively. Then, the value function satisfies the following HJB equation:

$$\begin{aligned} 0 &= \Phi_t + \sup_{\pi \in \mathfrak{M}^u, c \in [0, \infty)} \left[\frac{1}{2} |\pi^T \sigma|^2 \Phi_{vv} \right. \\ &\quad \left. + [(rv - c) + \pi^T (\mu - r\mathbf{1})] \Phi_v + \frac{c^\gamma}{\gamma} \right] \end{aligned} \quad (25)$$

with the boundary and terminal conditions

$$\Phi(t, 0+) = 0, \quad \forall t \in [0, T^*] \quad (26)$$

$$\Phi(T^*, v) = 0, \quad \forall v \in (0, \infty) \quad (27)$$

With the power utility, the solution of the HJB equation is

$$\Phi(t, v) = \frac{(g(t))^{1-\gamma}}{\gamma} v^\gamma \quad (28)$$

where

$$g(t) = \frac{1}{K} (1 - e^{-K(T-t)}) \quad (29)$$

with

$$K = -\frac{\gamma}{1-\gamma} \left(r + \frac{|\theta|^2}{2(1-\gamma)} \right) \quad (30)$$

$$\theta = \sigma^{-1}(\mu - r\mathbf{1}) \quad (31)$$

In this case, the optimal consumption and portfolio processes are, respectively, given by

$$c^*(t, v) = \frac{v}{g(t)} \quad (32)$$

and

$$\pi^*(t, v) = (\sigma^T)^{-1} \left(\frac{v}{1-\gamma} \right) \theta \quad (33)$$

The HJB method for the optimal consumption–investment problem is discussed in different monographs, such as [6, 15, 18, 19], and others. These monographs focus on discussing the verification theorem for the HJB solution to the problem. Loosely speaking, if a bounded, continuous, and smooth enough function satisfies the HJB equation with associated boundary conditions, the function

is identical to the value function. The verification theorem also provides a sufficient condition for an optimal control. For detailed discussion on the verification theorem, we refer to [18] for the diffusion case and [19] for the jump-diffusion case.

One fundamental result behind the HJB method is called the *principle of optimality*. Informally speaking, the principle of optimality states that if you do not know the optimal expected reward at the current time t , but have knowledge of how well you can achieve at some later time, say $t + h$, you can evaluate the expected reward associated with the policy of adopting the control u during $(t, t + h)$, acting optimally from $t + h$ onward and minimizing over the set of controls. Indeed, the HJB equation for a stochastic control problem follows “in principle” from the principle of optimality for dynamic programming. By the principle of optimality and Itô’s differentiation rule, it can be shown that the value function of a stochastic optimal control problem satisfies the HJB equation if the value function satisfies certain differentiability or smoothness conditions *see Monotone Schemes*. However, the principle of optimality and its derivation are often overlooked in some recent literature on the stochastic optimal control theory and its financial applications, but they are certainly important. Some fundamental contributions to these aspects were due to Davis and Varaiya [5] and Elliott [7], (see also [8]). In these works, the martingale method was used to deduce the principle of optimality.

The Stochastic Maximum Principle

Firstly, we suppose that the state $X(t) := X^{(u)}(t)$ of a controlled diffusion in $\mathbb{R}^n$ is

$$\begin{aligned} dX(t) = & b(t, X(t), u(t)) dt \\ & + \sigma(t, X(t), u(t)) dW(t) \end{aligned} \quad (34)$$

where the coefficients b and σ satisfy some regularity conditions.

Here, the control u enters both the drift coefficient b and the diffusion coefficient σ . We assume that (i) the control $u(t) := u(t, \omega)$ takes value in $U \subset \mathbb{R}^k$, for some positive integer k ; (ii) u is $\{\mathcal{F}(t)\}$ -progressively measurable and right continuous with left-hand limit (RCLL); and (iii) the controlled diffusion has a unique solution $\{X^{(u)}(t)\}$. These controls are called

admissible and we write $\mathcal{U}$ for the set of admissible controls.

We consider the same performance criterion as the one in the principle of optimality in the section The HJB Method and impose the same set of assumptions for the performance criterion as those in that section.

Let $H : [0, T^*] \times \mathbb{R}^n \times U \times \mathbb{R}^n \times \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \rightarrow \mathbb{R}$ denote the Hamiltonian given by

$$\begin{aligned} H(t, X, u, p, q) \\ := g(t, X, u) + b^T(t, X, u)p + \text{tr}(\sigma^T(t, X, u)q) \end{aligned} \quad (35)$$

where $\text{tr}(M)$ represents the trace of a square matrix M ; we suppose that H is differentiable in X .

The adjoint equation corresponding to u and $X^{(u)}$ for the unknown processes $\{p(t)\}$ and $\{q(t)\}$ is given by the following backward stochastic differential equation:

$$\begin{aligned} dp(t) = & -\nabla H(t, X(t), u(t), p(t), q^T(t)) dt \\ & + q^T(t) dW(t) \end{aligned} \quad (36)$$

$$p(T^*) = \nabla h(X(T^*)) \quad (37)$$

where ∇G is the gradient of a function G with respect to X . h is a concave function of X .

Then, we present a sufficient maximum principle in the following proposition.

Proposition 1 *Let $u^* \in \mathcal{U}$ and the corresponding controlled state process be $X^* := X^{(u^*)}$. Suppose there exists a solution $(p^*(t), q^*(t))$ of the corresponding adjoint equation, (36) and (37) satisfying*

$$E \left[\int_0^{T^*} q^*(t)(q^*(t))^T dt \right] < \infty \quad (38)$$

Suppose, further, that

1. *for each $t \in [0, T^*]$,*

$$\begin{aligned} & H(t, X^*(t), u^*(t), p^*(t), q^*(t)) \\ & = \sup_{u \in U} H(t, X^*(t), u, p^*(t), q^*(t)) \end{aligned} \quad (39)$$

2. *$h(X)$ is a concave function of X ,*

3. for each $t \in [0, T^*]$,

$$H^*(X) := \max_{u \in U} H(t, X, u, p^*(t), q^*(t)) \quad (40)$$

exists and is a concave function of X .

Then, u^* is an optimal control.

For the necessary condition of the stochastic maximum principle, we refer to [2, 3, 7–9, 20].

Now, we illustrate the application of the stochastic maximum principle to the optimal consumption–investment problem presented. Here, we just present some heuristic arguments. For detailed discussions and proofs, we refer to The Martingale Approach [3]. We assume the same asset price dynamics and adopt the same notation as those in the section. As in that section, the market is complete here, so the market price of risk is uniquely determined. We consider the following utility maximization problem for both consumption and terminal wealth:

$$J_1(v) := \sup_{(\pi, c) \in \mathcal{A}(v)} E \left[\int_0^{T^*} \frac{(c(t))^{\gamma_1}}{\gamma_1} dt + \frac{(V(T^*))^{\gamma_2}}{\gamma_2} \right] \quad (41)$$

In this case, the Hamiltonian is

$$\begin{aligned} H(t, V, (\pi, c), p, q) \\ = \frac{c^{\gamma_1}}{\gamma_1} + p[r(t)V - c + \pi^T(\mu(t) - r(t)\mathbf{1})] \\ + q^T \sigma^T(t) \pi \end{aligned} \quad (42)$$

and the adjoint equation has the following form:

$$dp(t) = -r(t)p(t)dt + q^T(t)dW(t) \quad (43)$$

$$p(T) = (V^*(T))^{\gamma_2-1} \quad (44)$$

Define an exponential process $\{Z_{\bar{\theta}}(t)\}$ by

$$Z_{\bar{\theta}}(t) := \exp \left(- \int_0^t \bar{\theta}^T(s) dW(s) - \frac{1}{2} \int_0^t |\bar{\theta}(s)|^2 ds \right) \quad (45)$$

where $\{\bar{\theta}(t)\}$ is a measurable, $\{\mathcal{F}(t)\}$ -adapted process, which is uniformly bounded in $(t, \omega) \in [0, T^*] \times \Omega$. Write $\xi_{\bar{\theta}}(t) := \beta(t)Z_{\bar{\theta}}(t)$, for each $t \in [0, T^*]$.

It can be shown that an adapted solution to the adjoint equation (42) and (43) is given by the processes

$$p(t) = p(0)\xi_{\bar{\theta}}(t), \quad q(t) = -p(0)\xi_{\bar{\theta}}(t)\theta(t) \quad (46)$$

From the first-order conditions of maximizing the Hamiltonian (5), one obtains

$$c^*(t) = (p(0)\xi_{\bar{\theta}}(t))^{1/(\gamma_1-1)} \quad (47)$$

$$\bar{\theta}(t) = \sigma^{-1}(\mu(t) - r(t)\mathbf{1}) = \theta(t) \quad (48)$$

To simplify the notation, write $\xi(t) := \xi_{\bar{\theta}}(t)$, for each $t \in [0, T^*]$. Define a function $\mathcal{X}: (0, \infty) \rightarrow (0, \infty)$ by

$$\begin{aligned} \mathcal{X}(y) \\ := E \left[\int_0^{T^*} \xi(s)(y\xi(s))^{\frac{1}{\gamma_1-1}} ds + \xi(T^*)(y\xi(T^*))^{\frac{1}{\gamma_2-1}} \right] \end{aligned} \quad (49)$$

Since $\mathcal{X}$ is strictly decreasing and surjective, its inverse $\mathcal{Y} := \mathcal{X}^{-1}: (0, \infty) \rightarrow (0, \infty)$ exists and is strictly decreasing. We then conjecture that the optimal controls (π^*, c^*) satisfy

$$E \left[\int_0^{T^*} \xi(s)c^*(s) ds + \xi(T^*)V^*(T^*) \right] = v \quad (50)$$

Under this conjecture, $p(0) = \mathcal{Y}(v)$.

By the martingale representation theorem, there exists a progressively measurable process $\eta: [0, T^*] \times \Omega \rightarrow \mathbb{R}^n$ with $\int_0^{T^*} |\eta(s)|^2 ds < \infty$, $\mathcal{P}$ a.s. such that

$$\begin{aligned} E \left[\int_0^{T^*} \xi(s)(\mathcal{Y}(v)\xi(s))^{\frac{1}{\gamma_1-1}} ds \right. \\ \left. + \xi(T^*)(\mathcal{Y}(v)\xi(T^*))^{\frac{1}{\gamma_2-1}} | \mathcal{F}(t) \right] \\ = v + \int_0^t \eta^T(s) dW(s) \end{aligned} \quad (51)$$

Then, it can be shown that

$$\pi^*(t) = \sigma^{-1}(t) \left[\frac{\eta(t)}{\xi(t)} + V^*(t)\theta(t) \right] \quad (52)$$

$$c^*(t) = (\mathcal{V}(v)\xi(t))^{\frac{1}{\gamma_1-1}} \quad (53)$$

where the optimal wealth process $\{V^*(t)\}$ is given by

$$V^*(t) = \frac{1}{\xi(t)} E \left[\int_t^{T^*} \xi(s) (\mathcal{V}(v)\xi(s))^{\frac{1}{\gamma_1-1}} ds + \xi(T^*) (\mathcal{V}(v)\xi(T^*))^{\frac{1}{\gamma_2-1}} | \mathcal{F}(t) \right] \quad (54)$$

References

- [1] Bellman, R.S. (1957). *Dynamic Programming*, Princeton University Press, Princeton.
- [2] Bensoussan, A. (1981). *Lectures on Stochastic Control*, Lecture Notes in Mathematics, 972, Springer-Verlag, Berlin, pp. 1–62.
- [3] Cadenillas, A. & Karatzas, I. (1995). The stochastic maximum principle for linear, convex optimal control with random coefficients, *SIAM Journal of Control and Optimization* **33**(2), 590–624.
- [4] Cox, J.C. & Huang, C.-F. (1989). Optimal consumption and portfolio policies when asset prices follow a diffusion process, *Journal of Economic Theory* **49**, 33–83.
- [5] Davis, M.H.A. & Varaiya, P.P. (1973). Dynamic programming conditions for partially observable stochastic systems, *SIAM Journal on Control and Optimization* **11**, 226–261.
- [6] Duffie, D. (1996). *Dynamic Asset Pricing Theory*, 2nd Edition, Princeton University Press, Princeton.
- [7] Elliott, R.J. (1977). The optimal control of a stochastic system, *SIAM Journal on Control and Optimization* **15**(5), 756–778.
- [8] Elliott, R.J. (1982). *Stochastic Calculus and Applications*, Springer, Berlin, Heidelberg, New York.
- [9] Elliott, R.J. & Kohlmann, M. (1994). The second order minimum principle and adjoint process, *Stochastics and Stochastics Reports* **46**, 25–39.
- [10] Elliott, R.J. & Kopp, P.E. (2005). *Mathematics of Financial Markets*, Springer, Berlin, Heidelberg, New York.
- [11] Fleming, W.H. & Rishel, R.W. (1975). *Deterministic and Stochastic Optimal Control*, Springer, Berlin, Heidelberg, New York.
- [12] Fleming, W.H. & Soner, H.M. (1993). *Controlled Markov Processes and Viscosity Solutions*, Springer, Berlin, Heidelberg, New York.
- [13] Karatzas, I., Lehoczky, J.P. & Shreve, S.E. (1987). Optimal portfolio and consumption decisions for a “small investor” on a finite horizon, *SIAM Journal of Control and Optimization* **25**, 1557–1586.
- [14] Karatzas, I. & Shreve, S.E. (1998). *Methods of Mathematical Finance*, Springer, Berlin, Heidelberg, New York.
- [15] Korn, R. (1997). *Optimal Portfolios: Stochastic Models for Optimal Investment and Risk Management in Continuous Time*, World Scientific, Singapore.
- [16] Merton, R.C. (1969). Lifetime portfolio selection under uncertainty: the continuous-time model, *Review of Economics and Statistics* **51**, 247–257.
- [17] Merton, R.C. (1971). Optimal consumption and portfolio rules in a continuous time model, *Journal of Economic Theory* **3**, 373–413.
- [18] Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*, 6th Edition, Springer, Berlin, Heidelberg, New York.
- [19] Øksendal, B. & Sulem, A. (2004). *Applied Stochastic Control of Jump Diffusions*, Springer, Berlin, Heidelberg, New York.
- [20] Peng, S. (1990). A general stochastic maximum principle for optimal control problems, *SIAM Journal of Control and Optimization* **28**, 966–979.
- [21] Yong, J. & Zhou, X.Y. (1999). *Stochastic Control*, Springer, Berlin, Heidelberg, New York.

Related Articles

Expected Utility Maximization; Expected Utility Maximization: Duality Methods; Markov Processes; Merton Problem; Monotone Schemes; Partial Differential Equations.

ROBERT J. ELLIOTT & TAK KUEN SIU

Transaction Costs

Standard models for financial markets are based on the simplifying assumption that trading orders can be given and executed in continuous time with no friction. This assumption is clearly a strong idealization of the reality. In particular, securities should not be described by a single price but by a bid and ask curve. As a first approximation, one may assume that the bid and ask prices do not depend on the traded quantities, which leads to models with proportional transaction costs. These models have attracted a lot of attentions these last years, mostly because their linear structure allows to develop a nice duality theory as in frictionless models.

No Arbitrage with Proportional Costs

The Fictitious Markets Approach in the One-dimensional Case

The study of models with proportional transaction costs starts with the paper of Jouini and Kallal [22] who considered a financial market with one nonrisky asset S^1 , taken as a numéraire and normalized to 1, and one risky asset called S^2 .

To be consistent with the developments below, we use a different (but equivalent) presentation than the one used in [22]. In particular, we denote by π^{ij} the number of physical units of asset i for which an agent can buy 1 unit of asset j . With these notations, the bid and ask prices of S^2 in terms of S^1 are given by $1/\pi^{21}$ and π^{12} . They are assumed to be right-continuous and adapted to the underlying (right-continuous) filtration $(\mathcal{F}_t)_{t \leq T}$.

In this model, simple self-financing trading strategies are defined as finite sequences of trading times $(t_n)_{n \leq N}$, for some $N \geq 1$, and random vectors of traded quantities $(\xi_{t_n})_{n \leq N}$ such that ξ_{t_n} is $\mathcal{F}_{t_n}$ -measurable. The i -th component $\xi_{t_n}^i$ of ξ_{t_n} stands for the number of physical units of S^i bought at the times t_n . In this framework, the usual self-financing condition reads $\xi_{t_n}^1 - (\xi_{t_n}^2)^+ \pi_{t_n}^{12} + (\xi_{t_n}^2)^- / \pi_{t_n}^{21} \leq 0$ for each $n \leq N$. The associated portfolio starting with a zero initial holding is described as a two-dimensional process $V_t^\xi = \sum_{t_n \leq t} \xi_{t_n}$ whose i -th component is the numbers of units of S^i held.

The key observation of Jouini and Kallal [22] is the following: if $\tilde{Z}^2$ is a process such that $\tilde{Z}_t^2 \in [1/\pi_t^{21}, \pi_t^{12}]$ a.s. (almost surely for all $t \leq T$), then the liquidation value at time T of the portfolio, $\ell(V_T^\xi) := V_T^{\xi,1} - (V_T^{\xi,2})^- \pi_T^{12} + (V_T^{\xi,2})^+ / \pi_T^{21}$, is a.s. lower than the terminal value $\tilde{V}_T^\xi := \sum_{t_n \leq T} \xi_{t_n}^2 (\tilde{Z}_{t_{n+1} \wedge T}^2 - \tilde{Z}_{t_n}^2)$ associated to the same strategy in a *fictitious market* in which the risky asset has the dynamics $\tilde{Z}^2$ and where there is no transaction costs. In particular, if there is an equivalent measure $\mathbb{Q}$ such that $\tilde{Z}^2$ is a martingale, then no arbitrage (NA) is possible: $\ell(V_T^\xi) \in L^0(\mathbb{R}_+) \Rightarrow \tilde{V}_T^\xi \in L^0(\mathbb{R}_+) \Rightarrow \tilde{V}_T^\xi = 0 \Rightarrow \ell(V_T^\xi) = 0$. Thus, the existence of such a process $\tilde{Z}^2$, called *fictitious price process*, admitting an equivalent martingale measure is a sufficient condition for the absence of arbitrage in this model.

The fundamental result in [22] is that this condition is actually also necessary, whenever we replace the notion of NA by that of *no free lunch*; see the above paper for a precise definition.

As an example, let us consider the case where $1/\pi^{21} = (1 - \lambda)S^2$ and $\pi^{12} = (1 + \lambda)S^2$ where S^2 is now viewed as a right-continuous adapted process and $\lambda \in (0, 1)$. Then, a necessary and sufficient condition for the absence of *free lunch* for simple strategies is that there exists a process $\tilde{Z}^2$ such that $\tilde{Z}_t^2 \in [(1 - \lambda)S_t^2, (1 + \lambda)S_t^2]$ a.s. for all $t \leq T$ and which is a martingale under some equivalent measure $\mathbb{Q}$. An important consequence of this result is that S^2 itself need not admit a martingale measure nor be a semimartingale under the original probability measure. One could, for instance, allow S^2 to be a fractional Brownian motion as in [16].

The Multivariate Case: The Solvency Region and Its Polar

In the multivariate setting where direct exchanges between many assets is possible, which is typically the case on currency markets, a similar reasoning can be used, and the geometric structure of the problem is more apparent. In particular, the notion of *solvency region*, introduced by Kabanov [29], and its positive polar play an important role.

The *solvency region* at time t is the set $K_t(\omega)$ formed by all vectors x such that we can find nonnegative numbers a^{ij} satisfying $x^i + \sum_j (a^{ji} - a^{ij} \pi_t^{ij}(\omega)) \geq 0$. It corresponds to positions, which can be transformed into nonnegative holdings after

suitable exchanges. A portfolio process, in units of physical quantities, is defined in [29] as a cadlag bounded variation process V satisfying $dV_t \in -K_t$. This means that there is a matrix-valued cadlag adapted process L , with nondecreasing components, such that $dV_t = \sum_j (dL_t^{ji} - dL_t^{ij} \pi_t^{ij})$, that is, dL_t^{ij} is the number of units of asset j obtained by selling $dL_t^{ij} \pi_t^{ij}$ units of asset i .

In this model, a strategy is said to be admissible if the following *no-bankruptcy* condition holds: $V_t - a\mathbf{1} \in K_t$ a.s. for some real number a with $\mathbf{1} := (1, \dots, 1)$. This means that the *liquidation value* of the portfolio is bounded from below by a .

The counterpart of the key observation of Jouini and Kallal [22] is the following. Let Z be a continuous martingale with positive components such that Z_t takes values in the positive polar K_t^* of K_t : $K_t^*(\omega) := \{z \in \mathbb{R}^d : 0 \leq z^i \leq z^j \pi_t^{ji}(\omega)\}$. Then, by the integration by parts formula, $Z_t \cdot V_t = \sum_i Z_t^i dV_t^i + V_t^i dZ_t^i$. The first part is nonpositive because $Z_t \in K_t^*$ and $dV_t \in -K_t$, the second part is a local martingale, and $Z \cdot V$ is bounded from below by the martingale aZ . This implies that $Z \cdot V$ is a supermartingale, which rules out any arbitrage opportunities: $V_T \in L^0(K_T) \Rightarrow Z_T \cdot V_T \in L^0(\mathbb{R}_+) \Rightarrow Z_T \cdot V_T = 0 \Rightarrow V_T \in L^0(\partial K_T)$. Otherwise stated, if the liquidation value of the portfolio is nonnegative a.s., then it must be equal to 0.

As in [22], the process Z has a nice interpretation in terms of *fictitious price process*. Indeed, if we set $\tilde{Z}^i = Z^i / (Z^1 / \mathbb{E}[Z_T^1])$, we see that the above conditions imply that the *fictitious market*, without transaction costs and where the dynamics of the i -th asset is given $\tilde{Z}^i$, is *cheaper* than the original one ($\tilde{Z}^i / \tilde{Z}^j \in [1/\pi^{ij}, \pi^{ji}]$) and admits no arbitrage (there is at least one martingale measure $\mathbb{Q} := Z_T^1 / \mathbb{E}[Z_T^1] \cdot \mathbb{P}$ for $\tilde{Z}$). See [7] for a precise statement.

Note that in this model no asset has been taken explicitly as a numéraire, except in the last interpretation in terms of *fictitious markets*. It turns out that, when working with quantities instead of amounts, as in usual frictionless models, the only important quantity is the bid–ask spread process $\pi = (\pi^{ij})$, which directly expresses the exchange rates between two assets.

No-arbitrage Conditions

Different notions of NA have been proposed for discrete time multivariate models. In the following,

we denote by A_T the set of terminal values of portfolios starting from 0.

1. The usual NA condition can be written as $A_T \cap L^0(K_T) = L^0(\partial K_T)$ (see above). It was studied in [32] and called *weak no arbitrage condition* therein. When the probability space is finite, it is equivalent to the existence of a martingale Z such that Z_t takes values in $K_t^* - \{0\}$ for all $t \leq T$ a.s. We therefore retrieve the process Z required for the NA condition in [29]. Moreover, this condition implies that A_T is closed in probability, a desirable feature to build on a nice duality for the set of superhedgeable claims (see below). Such a process Z was called a *consistent price system* in [36]. The notion of consistency reflects the fact that the exchange rates corresponding to the induced frictionless market (see above) lie within the original bid–ask spreads: $\tilde{Z}^i / \tilde{Z}^j \in [1/\pi^{ij}, \pi^{ji}]$.

However, this condition is not strong enough when Ω is not finite (see [36]). This leads to the introduction of a second notion of NA.

2. The *strict NA condition* (NA^s) introduced in [31] reads as follows: $A_t \cap L^0(K_t) = L^0(K_t^0)$ for all $t \leq T$. Here, $K_t^0 := K_t \cap (-K_t) (= \partial K_t \cap (-\partial K_t))$. The economic interpretation is that, if a wealth process V starting from a zero initial endowment has a nonnegative liquidation value at any time, then it is *equivalent* to 0, that is, its liquidation value is 0 ($V_t \in \partial K_t$) and it can be constructed from a 0 endowment at time t by a suitable immediate exchange on the market ($V_t \in -\partial K_t$). Under the *efficient friction* condition, $1/\pi^{ij} < \pi^{ji}$ for all $i \neq j$, which means that no couple of assets can be exchanged freely and can be written as $K_t^0 = \{0\}$, this condition is equivalent to the existence of a martingale Z which lies in the relative interior $\text{ri} K_t^*$ of K_t^* , that is, $Z^i / Z^j \in \text{ri}[1/\pi^{ij}, \pi^{ji}]$. Moreover, it implies that A_T is closed in probability. Such a process Z is called a *strictly consistent price system*.

This last notion of NA is sufficient to cover the cases where the transaction costs are strictly positive. However, up to the slight (also interesting) extension proposed [35], it does not allow to show that A_T is closed or to construct a *consistent price system* in general without the extra *efficient friction* condition, see the counterexample in [36].

3. The last notion was proposed in [36]. It is based on the idea that if a martingale Z satisfies the condition $Z^i/Z^j \in \text{ri}[1/\pi^{ij}, \pi^{ji}]$, then one can construct a bid–ask spread matrix $\bar{\pi}$ defined by $\bar{\pi}^{ji} := Z^i/Z^j$, which leads to a market without arbitrage, because Z is a martingale, and satisfies $[1/\bar{\pi}^{ij}, \bar{\pi}^{ji}] \subset \text{ri}[1/\pi^{ij}, \pi^{ji}]$ by construction. Thus, the existence of a *strictly consistent price system* implies a strong notion of NA: there exists a market associated to a bid–ask spread matrix $\bar{\pi}$ satisfying $[1/\bar{\pi}^{ij}, \bar{\pi}^{ji}] \subset \text{ri}[1/\pi^{ij}, \pi^{ji}]$ in which there is no arbitrage. Otherwise stated, one can slightly reduce the transaction costs, when they are not already equal to 0, and still preserve the NA condition. This condition was called the *robust NA* condition (NA^r) in [36]. The main result of this article is that this condition is sufficient to ensure that A_T is closed in probability and is actually equivalent to the existence of a *strictly consistent price system*. This is the good condition to impose on a model:

- (a) It is equivalent to the existence of a *strictly consistent price system* without any extra assumption.
- (b) When there is no friction, that is, $1/\pi^{ij} = \pi^{ji}$ for all i, j , it is equivalent to the usual no-arbitrage condition in frictionless markets.
- (c) Similar notions can be used to study models with nonlinear frictions (see [5]).
- (d) It can be extended to continuous-time models in the following form: the absence of arbitrage opportunities for arbitrary small transaction costs is equivalent to the existence of a *strictly consistent price system* for arbitrary small transaction costs. This result was proved in a model with only one risky asset with continuous paths in [17]. The existence of a *strictly consistent price system* for arbitrarily small transaction costs in a multivariate market with continuous price processes is a result of [18].

Superhedging and No-arbitrage Price Intervals

When there are transaction costs, we generally have more than one fictitious price process and more than

one martingale measure $\mathbb{Q}$ satisfying the conditions above. Furthermore, as underlined in [1], even if a contingent claim G can be duplicated by dynamic trading, the duplication strategy does not necessarily correspond to the cheapest way to hedge this claim. They thus introduced the concept of *super-replication price* $\pi(G)$ that corresponds to the minimum amount it costs to hedge the claim G (in terms of the first asset taken as a numéraire).

As first shown in [11] and [22], it can be obtained by taking the (normalized) expected value of $\tilde{Z}_T \cdot G$ with respect to all the *fictitious markets* $\tilde{Z}$ and all measures $\mathbb{Q}$ that characterize the absence of arbitrage opportunities in the corresponding *fictitious market*. Here, G is viewed as the vector of units of the different assets to be delivered. This result can easily be understood in the light of the above discussion. If $V_T - G \in K_T$, then $\tilde{Z}_T \cdot V_T \geq \tilde{Z}_T \cdot G$. Since $\tilde{Z} \cdot V$ is a $\mathbb{Q}$ -supermartingale (see above) it follows that $\tilde{Z}_0 \cdot V_0 \geq \mathbb{E}^{\mathbb{Q}}[\tilde{Z}_T \cdot G]$. The converse implication is obtained by using a standard separation argument, once A_T is known to be closed in a suitable sense.

The no-arbitrage prices interval is then equal to $[-\pi(-G), \pi(G)]$. Using the viability concept for price systems introduced in [20], the paper [23] also proves that these bounds are the tightest bounds that can be inferred at the equilibrium on the price of a contingent claim without knowing the agent's preferences (see also [21]). This is still the case even if we assume that agents have von Neumann–Morgenstern (VNM) preferences [24]. In particular, this means that even if the super-replication price seems too high (see below) it is always possible to construct VNM agents that are willing to pay amounts arbitrarily close to this super-replication price in order to hedge the considered asset (see also [4]).

Since endowments in different assets are not equivalent in the presence of transaction costs, it is also of interest to extend the notion of *superhedging price* to that of *initial endowments* $x \in \mathbb{R}^d$ that *allow to superhedge*. In this case, the above dual formulation reads $\tilde{Z}_0 \cdot x \geq \mathbb{E}^{\mathbb{Q}}[\tilde{Z}_T \cdot G]$ for all the *fictitious markets* $\tilde{Z}$ and all associated martingale measures $\mathbb{Q}$ (see [32] and [8]). It can be restated in terms of consistent price systems: $Z_0 \cdot x \geq \mathbb{E}[Z_T \cdot G]$ for all *consistence price systems* Z .

The case of American options can be treated similarly. However, it is not sufficient to impose the above condition at any stopping time lower than T as in frictionless markets. This is due to the absence of

total order on $\mathbb{R}^d$. To overcome this problem, one has to relax the notion of stopping times and consider the more general notion of *randomized* stopping times. See [6] and [10] for discrete time models, and, [3] and [15] for a continuous-time extension.

The notion that *superhedging* typically leads to much too high prices to be useful in the market (it usually corresponds to a buy-and-hold strategy) was first conjectured in [13] for call options and then proved by different authors at different level of generality; see [7] and the reference therein.

Utility Maximization

Thanks to the above-mentioned duality between superhedgeable claims and consistence price systems, existence of optimal strategies can be obtained for general models with transaction costs. The first general result was derived by Cvitanic and Karatzas [11] in a Brownian diffusion model and then extended by Cvitanic and Wang [12] to the semimartingale case. The general multivariate case was studied in [2] and [14] under *Asymptotic Elasticity* conditions similar to the one introduced in [33]; see [19] for the necessity of this condition in models with proportional transaction costs. All these papers show that the usual duality holds once we replace the notion of *equivalent local martingale measures* by that of *consistent price systems*.

In Markovian diffusion models, the partial differential equation (PDE) approach has also attracted a lot of attention. It leads to the Hamilton–Jacobi–Bellman (HJB) equations involving constraints on the gradients of the value function. It allows to show that the optimal strategy typically consists in maintaining the dotations in a given *no-trading region*. See, for example, [30, 37] and the references therein.

Extensions

In order to take a large set of possible frictions including multivariate transaction costs into account, one can follow the approach of [9] (in discrete time) and [26] and [28] (in continuous time), where it is proposed to deal directly with the space of possible cash flows, instead of the space of terminal payoffs, and provide a characterization of the no free-lunch assumption in terms of the existence of

a separating functional. In particular, Napp [34] develops an arbitrage pricing theory and a super-replication concept in this cash-flow space.

The case of fixed transaction costs is analyzed in [25] and [27]. The conclusion drawn is that the absence of free lunch is characterized by the existence of a (family of) martingale measure(s) for the frictionless price processes. The unique difference with the frictionless case consists in the fact that these martingale measures are not necessarily equivalent to the initial probability but only absolutely continuous with respect to it.

References

- [1] Bensaid, B., Lesne, J.-P., Pagès, H. & Scheinkman, J. (1992). Derivative asset pricing with transaction costs, *Mathematical Finance* **2**, 63–86.
- [2] Bouchard, B. (2002). Utility maximization on the real line under proportional transaction costs, *Finance and Stochastics* **6**(4), 495–516.
- [3] Bouchard, B. & Chassagneux, J.-F. (2009). Representation of continuous linear forms on the set of lag processes and the pricing of American claims under proportional costs, *Electronic Journal of Probability* **14**, 612–632.
- [4] Bouchard, B., Kabanov, Y. & Touzi, N. (2001). Option pricing by large risk aversion utility under transaction costs, *Decisions in Economics and Finance* **24**, 127–136.
- [5] Bouchard, B. & Pham, H. (2005). Optimal consumption in discrete time financial models with industrial investment opportunities and non-linear returns, *Annals of Applied Probability* **15**(4), 2393–2421.
- [6] Bouchard, B. & Temam, E. (2005). On the hedging of american options in discrete time markets with proportional transaction costs, *Electronic Journal of Probability* **10**, 746–760.
- [7] Bouchard, B. & Touzi, N. (2000). Explicit solution of the multivariate super-replication problem under transaction costs, *Annals of Applied Probability* **10**, 685–708.
- [8] Campi, L. & Schachermayer, W. (2006). A super-replication theorem in Kabanov’s model of transaction costs, *Finance and Stochastics* **10**(4), 579–596.
- [9] Carassus, L. & Jouini, E. (2000). A discrete stochastic model for investment with an application to the transaction costs case, *Journal of Mathematical Economics* **33**, 57–80.
- [10] Chalasani, P. & Jha, S. (2001). Randomized stopping times and american option pricing with transaction costs, *Mathematical Finance* **11**(1), 33–77.
- [11] Cvitanic, J. & Karatzas, I. (1996). Hedging and portfolio optimization under transaction costs: a martingale approach, *Mathematical Finance* **6**(2), 133–165.

- [12] Cvitanič, J. & Wang, H. (2001). On optimal terminal wealth under transaction costs, *Journal of Mathematical Economics* **35**(2), 223–231.
- [13] Davis, M. & Clark, J.M.C. (1994). A note on super-replicating strategies, *Philosophical Transactions of the Royal Society of London A* **347**, 485–494.
- [14] Deelstra, G., Pham, H. & Touzi, N. (2002). Dual formulation of the utility maximization problem under transaction costs, *Annals of Applied Probability* **11**(4), 1353–1383.
- [15] Denis, E., De Vallière, D. & Kabanov, Y. (2009). Hedging of American options under transaction costs, *Finance and Stochastics* **13**(1), 105–119.
- [16] Guasoni, P. (2006). No arbitrage with transaction costs, fractional Brownian motion and Markov processes, *Mathematical Finance* **16**(3), 569–582.
- [17] Guasoni, P., Rásonyi, M. & Schachermayer, W. (2007). The fundamental theorem of asset pricing for continuous processes under small transaction costs. *Annals of Finance* Forthcoming.
- [18] Guasoni, P., Rásonyi, M. & Schachermayer, W. (2008). Consistent price systems and face-lifting pricing under transaction costs, *Annals of Applied Probability* **18**(2), 491–520.
- [19] Guasoni, P. & Schachermayer, W. (2004). Necessary conditions for the existence of utility maximizing strategies under transaction costs, *Statistics and Decisions* **22**(2), 153–170.
- [20] Harrison, J. & Kreps, D. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- [21] Jouini, E. (2000). Price functionals with bid-ask spreads: an axiomatic approach, *Journal of Mathematical Economics* **34**, 547–558.
- [22] Jouini, E. & Kallal, H. (1995). Martingales and arbitrage in securities markets with transaction costs, *Journal of Economic Theory* **66**, 178–197.
- [23] Jouini, E. & Kallal, H. (1999). Viability and equilibrium in securities markets with frictions, *Mathematical Finance* **9**(3), 275–292.
- [24] Jouini, E. & Kallal, H. (2001). Efficient trading strategies in the presence of market frictions, *Review of Financial Studies* **14**, 343–369.
- [25] Jouini, E., Kallal, H. & Napp, C. (2001). Arbitrage and viability in securities markets with fixed trading costs, *Journal of Mathematical Economics* **35**, 197–221.
- [26] Jouini, E. & Napp, C. (2001). Arbitrage and investment opportunities, *Finance and Stochastics* **5**, 305–325.
- [27] Jouini, E. & Napp, C. (2007). Arbitrage with fixed costs and interest rate models, *Journal of Financial and Quantitative Analysis* Forthcoming.
- [28] Jouini, E., Napp, C. & Schachermayer, W. (2005). Arbitrage and state price deflators in a general intertemporal framework, *Journal of Mathematical Economics* **41**, 722–734.
- [29] Kabanov, Y. (1999). Hedging and liquidation under transaction costs in currency market, *Finance and Stochastics* **3**(2), 237–248.
- [30] Kabanov, Y. & Kluppelberg, C. (2004). A geometric approach to portfolio optimization in models with transaction costs, *Finance and Stochastics* **8**, 207–227.
- [31] Kabanov, Y., Rásonyi, M. & Stricker, C. (2002). No arbitrage criteria for financial markets with efficient friction, *Finance and Stochastics* **6**(3), 371–382.
- [32] Kabanov, Y. & Stricker, C. (2001). The Harrison-Pliska arbitrage pricing theorem under transaction costs, *Journal of Mathematical Economics* **35**(2), 185–196.
- [33] Kramkov, D. & Schachermayer, W. (1999). The asymptotic elasticity of utility functions and optimal investment in incomplete markets, *Annals of Applied Probability* **9**, 904–950.
- [34] Napp, C. (2001). Pricing issues with investment flows Applications to market models with frictions, *Journal of Mathematical Economics* **35**, 383–408.
- [35] Penner, I. (2001). *Arbitragefreiheit in Finanzmärkten mit Transaktionskosten*. Diplomarbeit, Humboldt-Universität zu Berlin, Berlin.
- [36] Schachermayer, W. (2004). The fundamental theorem of asset pricing under proportional transaction costs in finite discrete time, *Mathematical Finance*, **14**(1), 19–48.
- [37] Zariphopoulou, T. (1999). Transaction cost in portfolio management and derivative pricing, in *Introduction to Mathematical Finance. Proceedings of Symposia in Applied Mathematics* 57, D. Heath & R. Swindle, eds, AMS, Providence, RI.

Further Reading

Delbaen, F. & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Mathematische Annalen*, **300**, 463–520.

Related Articles

Arbitrage Strategy; Bid–Ask Spreads; Execution Costs; Fundamental Theorem of Asset Pricing; Price Impact; Superhedging.

BRUNO BOUCHARD & ELYES JOUINI

Behavioral Portfolio Selection

Behavioral portfolio selection is the study of how psychology impacts investors' portfolio choices. These choices pertain to the manner in which investors structure the risk-return composition of their portfolios and modify their portfolios over time through the buying and selling of securities.

The key behavioral features about portfolios include a dual emphasis on very safe and very risky securities, a lack of full diversification, excessive trading, the salience of securities that are purchased for the portfolio, and the disposition to sell winners too early but ride losers too long.

The theoretical framework underlying behavioral portfolio selection draws on the following four elements in the literature in behavioral decision making: (i) *SP/A* theory; (ii) prospect theory; (iii) regret and self-control; and (iv) heuristics and biases. Notably, the first three elements pertain to risk preferences, while the fourth deals with erroneous judgments about risk.

All four elements involve departures from the neo-classical mean-variance approach to portfolio selection. In the neoclassical approach, investors are rational in two senses. First, their preferences conform to the axioms of expected utility, where utility is defined over total return (or wealth) and risk is measured in terms of return standard deviation. Second, their judgments about risk and return are free of error. By way of contrast, in the behavioral approach, preferences typically violate the axioms of expected utility, investors attach importance to other variables besides total return (or wealth), investors do not measure risk in terms of return standard deviation, and investors make erroneous judgments about the risks they face.

In the mean-variance approach, investors hold well-diversified portfolios that feature risk and return being rationally balanced against each other. In the behavioral approach, investors are imperfectly rational, and in the course of attempting to make the best portfolio decisions they can, they adopt piecemeal approaches to portfolio selection that leave them holding undiversified portfolios.

Theoretical Foundations

SP/A Theory

Lopes [9] developed a psychologically based approach known as *SP/A theory* to explain choice among risky alternatives. She titled her 1987 article "The Psychology of Risk: Between Hope and Fear" to capture the idea that the emotions of hope and fear play key roles in choice among risky alternatives. In this respect, the letters that make up *SP/A* refer to specific concepts that measure or impact the degree of fear or hope experienced by a decision maker. Notably, *S* stands for security, *P* stands for potential, and *A* stands for aspiration.

In order to describe Lopes' formal framework, consider some notation. Denote the set of possible outcomes by a finite set of real numbers $X = \{x_1, \dots, x_n\}$, ordered from the lowest outcome x_1 to the highest x_n . Formally, a risky alternative or prospect is a random variable that takes on values in X . One way to describe the probabilities attached to a random variable is to use a decumulative distribution function D , where $D(x)$ is the probability that the outcome payoff is at least x . The same probabilistic information is also conveyed by the cumulative distribution function, which measures the probability that the outcome payoff is no greater than x , or the probability density function, which measures the probability that the outcome payoff is exactly x .

In a typical decision task, a decision maker chooses a best alternative from a menu $\{D\}$ of risky alternatives. The role of the theory is to describe the criteria leading to the choice. In *SP/A* theory, risky alternatives are evaluated using an objective function whose arguments reflect security *S*, potential *P*, and aspiration *A*.

Roughly speaking, increased fear stems from reduced security and reduced security corresponds to an increase in the probability attached to the occurrence of some unfavorable event. Suppose we consider two decision makers who face an identical risk, or prospect D , but experience different degrees of fear. Intuitively, the decision maker who experiences more fear attaches greater importance to the probability associated with unfavorable events than the decision maker who experiences less fear.

Formally, Lopes uses rank-dependent utility to capture the effects of fear. In rank-dependent utility, a decision maker who faces a risky prospect D

acts as if the decumulative density function is the transform $h(D)$ rather than D itself. In *SP/A* theory, fear operates on the probabilities associated with unfavorable events. Because D is a decumulative distribution function, the probability attached to the least favorable event x_1 is given by $\text{Prob}\{x_1\} = D(x_1) - D(x_2)$.

Under the transform $h(D)$, the probability attached to the least favorable event x_1 is given by $\text{Prob}\{x_1\} = h(D(x_1)) - h(D(x_2))$. In Lopes' framework, a decision maker who is fearful about exposure to event x_1 overweights the probability attached to x_1 . That is, the fearful decision maker employs a transform h that satisfies

$$h(D(x_1)) - h(D(x_2)) > D(x_1) - D(x_2) \quad (1)$$

Because x_1 is the least favorable outcome, the probability that the actual outcome turns out to be x_1 or higher is 1. In other words, $D(x_1) = 1$. What equation (3) effectively states is that increased fear leads to an increase in the slope of the h -function in the neighborhood of 1.

Rank-dependent utility also captures the manner in which the decision maker experiences hope. Hope is associated with upside potential, or potential for short. Hope induces a decision maker to attach greater weight to the most favorable events. Because D is a decumulative distribution function, the probability attached to the most favorable event x_n is given by $\text{Prob}\{x_n\} = D(x_n)$. Therefore, the emotion of hope leads a decision maker to employ a transform h that satisfies $h(D(x_n)) > D(x_n)$. For the second most favorable outcome, the corresponding inequality would read as

$$h(D(x_n)) - h(D(x_{n-1})) > D(x_n) - D(x_{n-1}) \quad (2)$$

This inequality indicates that increased hope leads to an increase in the slope of the h -function in the neighborhood of 0.

In Lopes' framework, a person who neither experiences fear nor hope is associated with an h -function that is the identity function: $h(D) = D$. A decision maker who experiences only fear, but not hope, is associated with an h -function that is strictly convex in D : it is steep in the neighborhood of 1 and flat in the neighborhood of 0. Formally, Lopes uses a power function $h_S(D) = D^q$, $q > 1$ for this case. A decision maker who experiences only hope is associated with an h -function that is strictly concave in D . Formally,

Lopes uses a power function $h_P(D) = 1 - (1 - D)^p$, $p > 1$ for this case. A person who experiences both fear and hope is associated with an h -function that has an inverse-S shape. It is concave in the neighborhood of the origin and convex in the neighborhood of 1. Formally, Lopes uses a convex combination of the power functions h_S and h_P to capture this case.

In *SP/A* theory, the degree to which fear and hope are experienced depends on the degree to which risky prospects offer security S and potential P . To capture the impact of both security and potential, Lopes uses an expected utility function with probabilities derived from the h -transform. She calls the function SP for security–potential, and it has the form

$$SP = \sum_{i=1}^n (h(D_i) - h(D_{i+1}))u(X_i) \quad (3)$$

In equation (1), u is a utility function whose argument is outcome x . Although Lopes uses the assumption $u(x) = x$ in most of her analysis, Lopes and Oden [10] comment that, in practice, u might display a bit of concavity.

The A in *SP/A* denotes aspiration. Aspiration pertains to a target value α (or range) to which the decision maker aspires. Aspiration points reflect different types of goals. For example, a decision maker might wish to generate an outcome that would allow the purchase of a particular good or service. Alternatively, the aspiration point might reflect a status quo position that corresponds to the notion of no gain or loss. In Lopes' framework, aspiration risk is measured as the probability 1- A where $A = \text{Prob}\{x \geq \alpha\}$ that the random outcome x meets or exceeds the aspiration level α .

In *SP/A* theory, the decision maker maximizes an objective function $V(SP, A)$ in deciding which alternative D to choose from the menu of available prospects. V is strictly monotone increasing in both of its arguments. Therefore, there are situations in which a decision maker is willing to trade off some SP in exchange for a higher value of A .

Prospect Theory

Prospect theory is a theory of choice developed by psychologists Kahneman and Tversky [5]. Prospect theory has four distinctive features.

First, the carriers of utility are changes, meaning gains and losses relative to a reference point, not the final position.

Second, the utility function (known as a *value function* in prospect theory) is concave in gains and convex in losses, with a point of nondifferentiability at the origin so that the function is more steeply sloped for losses (to the left of the origin) than for gains (to the right of the origin). Hence, the utility function is S-shaped with a kink at the origin. Tversky and Kahneman [16] suggest using a utility function $u(x)$ with the form x^α in the domain of gains ($x \geq 0$) and $-\lambda(-x)^\beta$ in the domain of losses ($x \leq 0$).

Third, probabilities are weighted (or distorted) when prospects are evaluated. In the original 1979 version of prospect theory (original prospect theory OPT), the weighting function π has probability density p as its argument. The π -function in OPT is convex, with $\pi(p) > p$ for small positive values of p and $\pi(p) < p$ for high values of p less than 1. In 1992, Tversky and Kahneman [16] proposed a cumulative version of prospect theory (cumulative prospect theory CPT) that uses rank-dependent utility. Unlike OPT, where probability weights depend only on probability density, the weights in CPT depend on outcome rank.

Tversky and Kahneman [16] use two weighting functions, one function for gains and one function for losses. Both functions take decumulative probabilities as their arguments, where the decumulative distribution pertains to the absolute value of the gain or loss, respectively. The weighting function is similar to the h -transform used by Lopes. It features an inverse S-shape, which Tversky–Kahneman generate using the ratio of a power function to a Hölder average; that is, $p^\gamma / (p^\gamma + (1-p)^\gamma)^{1/\gamma}$. As a result, in CPT, it is the probabilities of extreme outcomes that are overweighted (very large losses and very large gains).

Both the S-shape of the utility function and the inverse S-shape of the weighting functions reflect psychophysics, meaning the diminished sensitivity to successive changes. For the utility function, the changes pertain to differences relative to the reference point. For the weighting function, the changes pertain to differences relative to the endpoints 0 and 1.

Fourth, decision makers engage in editing or framing before formally evaluating risky prospects. There are several types of editing issues. Perhaps the simplest editing issue is the choice of reference point. Kahneman and Tversky illustrate this issue by describing a medical task in which the data can be presented, or framed, in one of two ways. The first

way is in terms of lives saved, while the second way is in terms of lives lost. The “lives saved” frame implicitly sets the reference point at 100% fatalities. The “lives lost” frame implicitly sets the reference point at 0% fatalities. Although the data underlying the two frames is identical, physicians tend to act as if they are more risk averse when presented with the data framed in terms of “lives saved” than when the data is framed in terms of “lives lost”. This choice pattern is consistent with the S-shaped utility function.

A more complex framing issue is the segmentation of a complicated decision task into a series of subtasks. The structure of each subtask is called a *mental account* and the segmentation process is known as *narrow framing*. Because narrow framing tends to overlook interdependencies between mental accounting structures, the segmentation process is often suboptimal. Tversky and Kahneman present examples in which narrow framing leads to the selection of stochastically dominated choices.

Prospect theory is a descriptive framework, not a normative framework. People who choose stochastically dominated alternatives do so because they do not always grasp the complete structure of the decision tasks they confront. The complete structure is typically opaque, not transparent, and people lack the ability to frame complex decision tasks transparently.

Regret and Self-control

In the early development of prospect theory, Kahneman and Tversky focused on the role of regret. Regret is the psychological pain associated with recognizing after the fact that taking a different decision would have produced a better outcome. Kahneman *et al.* [7] eventually built prospect theory using the S-shaped value function, but in their 1982 work they continued to emphasize the importance of regret. They pointed out that regret will be magnified by the ease with which a person can imagine taking a different decision.

Self-control refers to situations when a person is conflicted, and *thinks* he or she should take one decision, but emotionally *feels* like taking a different decision. Studies of self-control in financial economics tend to emphasize the difficulty in delaying gratification. However, self-control applies more broadly,

and in particular applies when the emotion of regret prevents a person from taking a decision that he or she “thinks” is appropriate.

Heuristics and Biases

The weighting functions in *SP/A* theory and prospect theory reflect the way that people process known probabilistic information. In these theories, people weight probabilities as they do because of emotion (as in *SP/A* theory) or psychophysics (as in prospect theory). In contrast, heuristics and biases involve errors in judgments about the probabilities themselves. A person might know the true probability of winning the lottery but because of hope overweights its value psychologically when deciding whether or not to purchase a lottery ticket. On the other hand, a person who is unrealistically optimistic would tend to overestimate the probability of winning the lottery.

The volume edited by Kahneman *et al.* [6] contains the foundation contributions to the heuristics and biases literature. A heuristic is a crude rule of thumb for making judgments about probabilities, statistics, future outcomes, and so on. A bias is a predisposition toward making a particular judgmental error. The heuristics and biases approach studies the heuristics upon which people rely to form judgments, and the associated biases in those judgments.

Some biases are associated with specific heuristics. Examples of these biases relate to such heuristic principles as availability, anchoring, and representativeness.

Availability is the tendency to form judgments based on information that is readily available, but to underweight information because it is not readily available. For example, a person might underestimate the danger from riptides and overestimate the danger from shark attacks because media stories tend to report most shark attacks but rarely report less dramatic incidents involving riptides. Heuristics and biases associated with availability reflect the importance of salience and attention.

Anchoring is the tendency to formulate an estimate by using a process that begins with an initial number (the anchor) and then making adjustments relative to the anchor. Anchoring bias is the tendency for the associated adjustments to be too small.

Representativeness is the tendency to rely on stereotypes to make a judgment. For example, a person who relies on representativeness might be

especially bold in predicting that the future return of a particular stock will be very favorable because its past long-term performance has been very favorable. This is because they form the judgment that favorable past performance is representative of good stocks. However, representativeness leads such predictions to be overly bold, because of insufficient attention to factors that induce regression to the mean.

Although some biases relate directly to specific heuristics, other biases stem from a variety of factors. For example, people tend to be overconfident about both their abilities and their knowledge. People who are overconfident about their abilities overestimate those abilities. People who are overconfident about their knowledge tend to think they know more than they actually do. In particular, people who are overconfident about their knowledge tend to set confidence intervals around their estimates that are too narrow. As a result, they wind up being surprised at their mistakes more often than they anticipate.

Other examples of biases that do not stem directly from specific heuristics are unrealistic optimism and the illusion of control. Unrealistic optimism involves overestimating the probabilities of favorable events and underestimating the probabilities of unfavorable events. The illusion of control is overestimating the role of skill relative to luck in the determination of outcomes.

Implications for Portfolio Selection

SP/A Theory and Portfolio Selection

Shefrin and Statman [15] use the *SP/A* framework as the basis of behavioral portfolio theory. They develop a model with two dates, $t = 0$ and $t = 1$, in which an investor with initial wealth $W = 1$ chooses a portfolio at $t = 0$. The model is structured so that at $t = 1$ one of n possible states will occur, and the subjective probability (density) associated with the occurrence of state i is p_i . The model also features a complete market, meaning that securities are priced in accordance with state prices $v_1, \dots, v_n$, where v_i is the price associated with the delivery of 1 unit of consumption in state i .

A portfolio return configuration is given by $x_1, \dots, x_n$ where x_i denotes the number of units of consumption paid if state i occurs. Notice that

because $W = 1$, $x_1, \dots, x_n$ are indeed gross rates of return, which are assumed to be nonnegative.

The decision task for investor with SP/A preferences is to choose a portfolio return configuration $x_1, \dots, x_n$ to maximize the objective function $V(SP, A)$ subject to the constraint

$$\sum_{i=1}^n v_i x_i = 1 \quad (4)$$

The maximization of $V(SP, A)$ for fixed A is formally equivalent to a constrained expected utility maximization problem, where the decision weights derived from the h -transform are treated like probabilities. The associated constraint is $A = \text{Prob}\{x \geq \alpha\}$. The effect of this constraint, when active, is to introduce a flat region $i_L \leq i \leq i_U$ for which $x_i = \alpha$. Notably, the investor meets the A -constraint by fulfilling this constraint from the favorable states down. This can result in three regions: $x_n > \alpha$, $x_i = \alpha$ for $i_L \leq i \leq i_U$, and $x_i < \alpha$ for $i < i_L$, and when u is linear $x_i = 0$ for $i < i_L$. In effect, an SP/A portfolio can be thought of as the combination of a risky bond and a call option on a neoclassical portfolio with a high exercise price.

There are two key questions associated with SP/A portfolios. First, how do the h -function and curvature of the utility function $u(x)$ impact the choice of portfolio payoff configuration? Second, how is the return configuration impacted by the values of α and A ?

Lopes suggests that the utility function $u(x)$ is mildly concave, although for purposes of exposition she treats it as linear in her discussions. Notably, linearity encourages an investor to concentrate as much wealth as possible on purchasing claims associated with the state featuring the lowest state price per unit probability. This will lead to a lottery property, meaning the small probability of a very large payoff. Concavity in the utility function will dampen the lottery property.

The impact of fear and hope occurs through the SP function, where the probabilities associated with the least favorable states and the most favorable states are overweighted. Such overweighting leads to higher returns in extreme states than would occur otherwise, and therefore lower returns in intermediate states.

It is the impact of the A variable that makes SP/A theory distinct from other psychologically based

theories of risk. It should be kept in mind that the investor tends to fulfill the A -constraint from the most favorable state down. Increasing the value of A leads to a shift in returns from both the most unfavorable states and the most favorable states, where $x_i \neq \alpha$, to expand the middle region where $x_i = \alpha$. If an investor increases the value of α , then she also shifts return from the extremes to the middle, but with the purpose of raising the level of the middle region.

SP/A theory implies that investors will choose portfolios whose return patterns can be generated by combining a risky bond and a call option on a neoclassical portfolio associated with unconstrained SP -maximization.

Prospect Theory and Portfolio Selection

Two features of prospect theory are particularly germane to portfolio selection, and both involve the manner in which the information underlying the selection task is framed. The first feature is the simplification of the selection task through the use of mental accounts. The second feature is the reference point used to define gains and losses.

The use of mental accounting leads investors to evaluate decisions about securities with little or no reference to other securities in the portfolio. This is in sharp contrast to neoclassical theory, where the value of a security to an investor very much depends on the return covariance of that security with other securities in the portfolio. Mental accounting implies that investors make little if any direct use of the return variance–covariance matrix.

Shefrin and Statman [14] suggest that the most natural reference point for the mental account associated with an individual security is original purchase price. As discussed earlier, the location of a reference point is important because people's attitude toward risk depends critically on whether they view the set of possible outcomes as gains, losses, or a mixture of gains and losses. This feature was first highlighted by Kahneman and Tversky [5], and is shared by SP/A theory.

Mental accounting is also associated with investors having a multitude of different goals. In this case, the investor associates a mental account or portfolio layer to a specific goal α and associated probability A . Downside protection is associated with a low α and high probability A . Upside potential is associated

with a high α and associated probability A that is as high as feasibility allows.

Regret, Self-control, and Portfolio Selection

People experience regret when they admit to having made a decision that turned out poorly. When an investor purchases a stock that subsequently performs poorly, the investor is prone to experiencing regret. Shefrin and Statman [14] suggest that the degree of regret is especially high when an investor sells a stock at a loss.

Because there are tax benefits from selling stocks at a loss, many investors who delay tax-loss selling forego potential benefits. In doing so, they pay a price to defer the pain of regret, hoping that the stock will bounce back so that they can avoid selling at a loss. On the flip side, selling a stock for a gain can be a source of pride, even if there is a tax penalty for doing so. In this respect, imperfect self-control can lead investors to sell stocks before their gains become long-term, thereby leading to a higher tax liability. Therefore, regret and imperfect self-control predispose investors to sell winners too early and ride losers too long, a phenomenon that Shefrin and Statman [14] call *the disposition effect*.

The S-shaped utility function in prospect theory implies that decision makers are prone to be risk averse in the domain of gains but risk seeking in the domain of losses. For this reason, prospect theory is the natural starting point for discussing the disposition effect. However, prospect theory does not explain why an investor would knowingly incur an unnecessary tax penalty.

Heuristics, Biases, and Portfolio Selection

Few, if any, investors have objectively correct knowledge of return distributions. Most investors formulate their beliefs by applying heuristics to the information at their disposal. As such, they are vulnerable to forming biased beliefs.

Biases take many forms. Barber and Odean [2] point out that because of reliance on the availability heuristic, investors tend to place undue stress on stocks that have attracted their attention. They call this phenomenon *the attention hypothesis*. De Bondt and Thaler [3] suggest that individual investors who rely on representativeness are prone to extrapolate past performance with undue weight on the

recent past. They call this phenomenon *the overreaction effect*. Shefrin [13] suggests that professional investors who rely on representativeness apply it differently than individual investors and are prone to attach too high a probability to reversals. Odean [11] suggests that overconfidence leads all investors to be overconfident in their beliefs, thereby trading with excessive frequency on unwarranted convictions.

Empirical Evidence

Much of the literature pertaining to behavioral portfolio selection involves the development of hypotheses in theoretical papers followed by other papers which tested these hypotheses.

Polkovnichenko [12] and Kumar [8] provide evidence that supports hypotheses stemming from *SP/A*-based portfolio theory. Kumar's work documents that the portfolios of individual investors overweight high-risk stocks, which he calls *lottery stocks*, while the portfolios of professional investors underweight lottery stocks. Polkovnichenko's work characterizes the degree to which the portfolios of individual investors are driven by fear, as reflected in their unwillingness to hold equities. His work also highlights the lack of diversification in most investors' portfolios. Odean [11] documents the degree to which individual investors are prone to the disposition effect, and Frazzini [4] shows that professional investors are also prone. Barber and Odean [2] provide evidence that individual investors purchase stocks by relying on the availability heuristic more so than professional investors. Barber and Odean [1] provide evidence that excessive trading by individual investors harms performance.

References

- [1] Barber, B. & Odean, T. (2000). Trading is hazardous to your wealth: the common stock investment performance of individual investors with Brad Barber, *Journal of Finance* **LV**(2), 773–806.
- [2] Barber, B. & Odean, T. (2008). All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors, *The Review of Financial Studies* **21**(2), 785–818.
- [3] De Bondt, W. & Thaler, R. (1985). Does the stock market overreact? *Journal of Finance* **40**, 793–805.
- [4] Frazzini, A. (2006). The disposition effect and underreaction to news, *Journal of Finance* **41**(6), 2017–2046.

-
- [5] Kahneman, D. & Tversky, A. (1979). Prospect theory: an analysis of decision making under risk, *Econometrica* **47**(2), 263–291.
 - [6] Kahneman, D., Slovic, P. & Tversky, A. (1982). The psychology of preferences, *Scientific American* **246**, 160–173.
 - [7] Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.
 - [8] Kumar, A. (2007). *Who Gambles in the Stock Market?* Working paper, University of Texas.
 - [9] Lopes, L. (1987). Between hope and fear: the psychology of risk, *Advances in Experimental Social Psychology* **20**, 255–295.
 - [10] Lopes, L.L. & Oden, G.C. (1999). The role of aspiration level in risk choice: a comparison of cumulative prospect theory and SP/A theory, *Journal of Mathematical Psychology* **43**, 286–313.
 - [11] Odean, T. (1998). Are investors reluctant to realize their losses? *Journal of Finance* **53**(5), 1775–1798.
 - [12] Polkovnichenko, V. (2005). Household portfolio diversification: a case for rank-dependent preferences, *The Review of Financial Studies* **18**(4), 1467–1501.
 - [13] Shefrin, H. (2005). *A Behavioral Approach to Asset Pricing*, Elsevier Academic Press, Boston.
 - [14] Shefrin, H. & Statman, M. (1985). The disposition to sell winners too early and ride losers too long: theory and evidence, *Journal of Finance* **40**(3), 777–790.
 - [15] Shefrin, H. & Statman, M. (2000). Behavioral portfolio theory *Journal of Financial and Quantitative Analysis* **35**, 127–151.
 - [16] Tversky, A. & Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty, *Journal of Risk and Uncertainty* **5**, 297–323.

Related Articles

Ambiguity; Expectations Hypothesis; Modern Portfolio Theory; Risk Aversion; Utility Function.

HERSH SHEFRIN

Kelly Problem

Consider a financial market with K assets whose prices $P_i(t), i = 1, \dots, K$ are stochastic, dynamic processes, and a risk-free asset whose price is $P_0(t)$. The vector of prices at time t is

$$\mathbf{P}(t) = (P_0(t), P_1(t), \dots, P_K(t))' \quad (1)$$

If the prices are given at points in time t_1 and t_2 , with $t_1 < t_2$, then the rate of return over that time on a unit of capital invested in asset i is

$$R_i(t_1, t_2) = \frac{P_i(t_2)}{P_i(t_1)} = 1 + r_i(t_1, t_2), i = 0, \dots, K \quad (2)$$

When there are dividends D_i accrued in the time interval, then the return is $R_i(t_1, t_2) = (P_i(t_2) + D_i(t_2 - t_1))/P_i(t_1)$.

Suppose an investor has w_t units of capital at time t , and that capital is fully invested in the assets, with the proportions invested in each asset given by $x_i(t), i = 0, \dots, K$, where $\sum_{i=0}^K x_i(t) = 1$. Then an investment or trading strategy at time t is the vector process

$$\mathbf{X}(t) = (x_0(t), x_1(t), \dots, x_K(t))' \quad (3)$$

Given the investments $w_{t_1} \mathbf{X}(t_1)$ at time t_1 , the accumulated capital at time t_2 is

$$W(t_2) = w_{t_1} R'(t_1, t_2) \mathbf{X}(t_1) = w_{t_1} \sum_{i=0}^K R_i(t_1, t_2) x_i(t_1) \quad (4)$$

The trajectory of returns between time t_1 and time t_2 depends on the asset, and is typically nonlinear. So changing the investment strategy at points in time between t_1 and t_2 will possibly improve capital accumulation. If trades could be timed to correspond to highs and lows in prices, then greater capital would be accumulated. To consider the effect of changes in strategy, partition the time interval into n segments, with $d = \frac{t_2 - t_1}{n}$, so that the accumulated capital is monitored, and the investment strategy is possibly revised at times $t_1, t_1 + d, \dots, t_1 + nd = t_2$. Then

wealth at time t_2 is

$$W_n(t_2) = w_{t_1} \prod_{i=0}^{n-1} R'(t_1 + id, t_1 + (i+1)d) \times X(t_1 + id) \quad (5)$$

Alternatively, wealth is

$$W_n(t_2) = w_{t_1} \left(\exp \left[\frac{1}{n} \sum_{i=0}^{n-1} \ln(R'(t_1 + id, t_1 + (i+1)d) X(t_1 + id)) \right] \right)^n \quad (6)$$

The exponential form highlights the *growth rate* with the strategy $\mathbf{X} = (X(t_1), \dots, X(t_1 + (n-1)d))$,

$$G_n(\mathbf{X}) = \frac{1}{n} \sum_{i=0}^{n-1} \ln(R'(t_1 + id, t_1 + (i+1)d) \times X(t_1 + id)) \quad (7)$$

As the partitioning of the interval gets finer, so that $d \rightarrow 0$, then monitoring and trading are continuous. If d is fixed and the random variables $V_i = \ln(R'(t_1 + id, t_1 + (i+1)d) X(t_1 + id)), i = 0, \dots, n-1$ are independent and identically distributed (i.i.d.) with mean μ_d and variance σ_d^2 , then $S_n(t_2) = (1/(\sigma_d \sqrt{n})) \sum (V_i - \mu_d)$ converges as n increases (i.e., as t_2 increases) to a standard normal variable. The simplest continuous time process with normally distributed accumulations is the Brownian motion model. In the continuous case, therefore, it is usually assumed that the instantaneous returns $dP_i(t)/P_i(t)$ are approximated by Brownian motion.

If the distribution of accumulated capital (wealth) at the horizon is the criterion for deciding on an investment strategy, then the rate of growth of capital becomes the determining factor when the horizon is distant. For fixed d and i.i.d V_i , the growth rate converges to the mean growth rate as n increases, so considering the average growth rate between t_1 and t_2 , for strategy $\mathbf{X} = (X(t_1), \dots, X(t_1 + (n-1)d))$,

$$E[G_n(\mathbf{X})] = \frac{1}{n} \sum_{i=0}^{n-1} E[\ln(R'(t_1 + id, t_1 + (i+1)d) \times X(t_1 + id))] \quad (8)$$

2 Kelly Problem

Table 1 Some good and bad properties of the optimal capital growth strategy

Feature	Property	Reference
Good	Maximizes the asymptotic rate of growth	Algeot and Cover [1] and Brieman [3]
	Maximizes median log wealth	Ethier [5]
	Minimizes expected time to asymptotically large goals	Algeot and Cover [1], Brieman [3] and Browne [4]
	Never risks ruin	Hakansson and Miller [8]
	Kelly is the unique evolutionary strategy	Hens and Schenk-Hoppe [9]
Bad	It takes a long time to outperform other strategies with high probability	Aucamp [2]; Browne [4] and Thorp [13]
	The total amount invested swamps the gains	Ethier and Tavaré [6]; Griffin [7]
	The average return converges to half the return from optimal expected wealth	Ethier and Tavaré [6]; Griffin [7]
	Kelly strategy does not optimize the expected nonlogarithmic utility of wealth. Example: Bernoulli trials with $1/2 < p < 1$, and $u(w) = w$. Then $x = 1$ maximizes $u(w)$, but $x = 2p - 1$ maximizes $E[\ln(w)]$.	Samuelson [11]; Thorp [12]

where E denotes the expected value.

The case usually discussed is the one in which the incremental returns are serially independent. So the maximization of $E[G_n(X)]$ is

$$\max \{E[\ln(R'(t_1 + id, t_1 + (i + 1)d)X(t_1 + id))]\} \quad (9)$$

separately for each i . If the distribution of the returns is the same for each i , a fixed strategy holds over time.

The strategy that solves equation (9) subject to the normalization constraint is called the *Kelly* or *optimal capital growth strategy*. This strategy is the unique *evolutionary stable* strategy, in the sense that it asymptotically overwhelms other portfolio rules that may be used within the population of investors with the accumulated wealth criterion. Strategies that survive in the long run must converge to the optimal growth strategy [9]. In the case of a stationary returns distribution, the Kelly or log optimal portfolio is $X^{*'} = (x_0^*, \tilde{X}^{*'})$, where $x_0^* = 1 - \sum_{i=1}^K x_i^*$. This Kelly strategy is a *fixed mix*. In other words, the fraction of wealth invested in assets is determined by X^* , but rebalancing is required to maintain the fractions as wealth varies.

The optimal growth/Kelly strategy has been studied extensively. A list of some of its properties is given in Table 1.

A variation on the Kelly strategy is the fractional Kelly strategy defined as $\tilde{X}_f = f\tilde{X}^*$, $f \geq 0$. The

fractional Kelly strategy has the same distribution of wealth across risky assets as the Kelly, but varies the fraction of wealth invested in those risky assets. Table 2 gives an example of the gain in security and the loss in return from a half-Kelly strategy. The results are from a simulation [15] assuming initial wealth of \$1000 and 700 decision points for investing in five possible assets, each with expected return of 1.14.

Observe that the Kelly strategy has enormous returns most of the time, but it is possible to make 700 independent bets, all with a 14% advantage at differing odds, and lose 98% of one's fortune. So the Kelly strategy, which is used by many great investors (see [14]) is risky in the short term because the absolute Arrow-Pratt risk aversion index is almost zero; it is also risky in the long term and must be used with care.

Table 2 Performance of Kelly and half-Kelly strategies

Final wealth statistic	Kelly	Half-Kelly
Minimum	18	145
Maximum	483 883	111 770
Mean	48 135	13 069
Median	17 269	8043
Probability of exceeding 500	0.916	0.990
Probability of exceeding 1000	0.870	0.945
Probability of exceeding 10 000	0.598	0.480
Probability of exceeding 50 000	0.302	0.03
Probability of exceeding 100 000	0.166	0.001

References

- [1] Algeot, P. & Cover, T.M. (1988). Asymptotic optimality and asymptotic equipartition properties of log-optimum investment, *Annals of Probability* **16**, 876–898.
- [2] Aucamp, D. (1993). On the extensive number of plays to achieve superior performance with the geometric mean strategy, *Management Science* **39**, 1163–1172.
- [3] Brieman, L. (1961). Investment policies for expanding business optimal in a long-run sense, *Naval Research Logistics Quarterly* **7**, 647–651.
- [4] Browne, S. (1998). The return on investment from proportional portfolio strategies, *Advances in Applied Probability* **30**, 216–238.
- [5] Ethier, S.N. (2004). The Kelly system maximizes median fortune, *Journal of Applied Probability* **41**, 1230–1236.
- [6] Ethier, S.N. & Tavaré, S. (1983). The proportional bettor's return on investment, *Journal of Applied Probability* **20**, 563–573.
- [7] Griffin, P. (1985). Different measures of win rates for optimal proportional betting, *Management Science* **30**, 1540–1547.
- [8] Hakansson, N. & Miller, B. (1975). Compound-return mean-variance efficient portfolios never risk ruin, *Management Science* **22**, 391–400.
- [9] Hens, T. & Schenk-Hoppe, K. (2005). Evolutionary stability of portfolio rules in incomplete markets, *Journal of Mathematical Economics* **41**, 43–66.
- [10] Kelly, J. (1956). A new interpretation of information rate, *Bell System Technology Journal* **35**, 917–926.
- [11] Samuelson, P.A. (1971). The fallacy of maximizing the geometric mean in long sequences of investing or gambling, *Proceedings of National Academy of Science* **68**, 2493–2496.
- [12] Thorp, E.O. (1971). Portfolio choice and the Kelly criterion, in *Proceedings of the Business and Economics Section of the American Statistical Association*, pp. 215–224.
- [13] Thorp, E.O. (2006). The Kelly criterion in blackjack, sports betting, and the stock market, in *Handbook of Asset and Liability Management, Theory and Methods*, S.A. Zenios & W.T. Ziemba, eds, North-Holland, Vol. 1, pp. 406–427.
- [14] Ziemba, W.T. (2005). The symmetric downside-risk Sharpe ratio and the evaluation of great investors and speculators, *Journal of Portfolio Management* Fall, 108–122.
- [15] Ziemba, W.T. & Hausch, D.B. (1986). *Betting at the Racetrack*, Dr. Z Investments, Inc., San Luis Obispo, CA.

Further Reading

- Bell, R.M. & Cover, T.M. (1980). Competitive optimality of logarithmic investment, *Mathematics of Operations Research* **5**, 161–166.
- MacLean, L.C., Ziemba, W.T. & Blazenko, G. (1992). Growth versus security in dynamic investment analysis, *Management Science* **38**, 1562–1585.
- Rotando, L.M. & Thorp, E.O. (1992). The Kelly criterion and the stock market, *American Mathematical Monthly*, December, 922–931.
- Stutzer, M. (2003). Portfolio choice with endogenous utility: a large deviations approach, *Journal of Econometrics* **116**, 365–386.

Related Articles

Expected Utility Maximization; Fixed Mix Strategy; Sharpe Ratio.

LEONARD C. MACLEAN & WILLIAM T.
ZIEMBA

Drawdown Minimization

Drawdown versus Shortfall

One measure of riskiness of an investment is “drawdown”, defined, most often in the asset management space, as the decline in net asset value from a historic high point. Mathematically, if the net asset value is denoted by $V_t, t \geq 0$, then the current “peak-to-trough” drawdown is given by $D_t = V_t - \max_{0 \leq u \leq t} V_u$. The maximum drawdown, $\max_{0 \leq u \leq t} D_u$, is a statistic that the CFTC forces managed futures advisors to disclose, and so many investment advisors and managers implicitly face drawdown constraints in setting their investment strategies. Hedge funds, for example, implicitly face drawdown constraints in that many multiperiod hedge fund contracts reflect investor preferences related to the maximum drop in a fund’s asset value from the previous peak. These often include a high-water-mark provision that sets the strike price of each period’s incentive fee equal to the all-time high of fund value (see **Hedge Funds**).

Another measure of riskiness that is related but often confused terminologically with drawdown is that of “shortfall”, which is simply the gap, or loss level of the current value from the initial or some other given value. This value could be constant but more often is determined by a stochastic exogenous or endogenous benchmark. For example, the shortfall with respect to the endogenous benchmark of the running maximum is the drawdown.

Since drawdown and shortfall are essentially equivalent in single period models, the research on the topic reviewed in this article is focused on multiperiod and, in particular, on continuous-time models pioneered in [16] and [17] where optimal portfolio rules are derived by solving a multiperiod portfolio optimization problem. The minimization of short-fall probability in a single-period model dates back to [18] and [21]. See **Value-at-Risk; Expected Shortfall** for work on portfolio selection with drawdown constraints in single period mean-variance models.

In the continuous time framework, there is an implicit nonnegativity constraint on wealth, which is one form of a shortfall constraint (see **Merton Problem**).

The models reviewed here differ in their assumptions regarding investment horizons (finite or infinite), constraints (fixed or stochastic benchmark), stochastic processes (diffusion with and without jumps), as well as objective function (purely probabilistic or expected utility based). In general, without transactions costs, the incorporation of drawdown constraints induces a portfolio insurance strategy: specifically, in the stationary stochastic model case, the strategy is that of a constant proportions portfolio insurance (CPPI) with different “floor” levels determined by the horizon and the objective (see **Transaction Costs** for portfolio optimization with transaction costs).

In this case, the risky asset price is assumed to follow a geometric Brownian motion with drift $\mu + r$, and diffusion coefficient σ^2 where r is the rate of return on cash. Hence, as is standard (see **Merton Problem**), the dynamics of the investor’s wealth portfolio are given by

$$dW_t = r W_t dt + x_t [\mu dt + \sigma dZ_t] \quad (1)$$

where Z_t is a standard Brownian motion and where x_t denotes the dollar holdings of stock.

For reference, investment strategies that are of the form

$$x_t = k \times W_t \quad (2)$$

are called a *constant proportion rebalanced portfolio rule* (see **Fixed Mix Strategy**), and are optimal in a variety of settings ([7]). Investment strategies that are of the form

$$x_t = k(W_t - F_t) \quad (3)$$

are referred to as *constant proportion portfolio insurance strategies (CPPI)* with scale multiple k and floor F_t (see **Constant Proportion Portfolio Insurance**).

Such CPPI strategies are, in fact, constant proportional strategies on the *surplus* $W_t - F_t$ and in a pure diffusion setting insure that wealth remains above the floor level F_t at all times, (although it is possible that F_t serves as an absorbing barrier). These strategies effectively synthesize an overlay of a put option on top of the wealth generated by a constant proportional

strategy, and are at the core of many of the strategies that have been discussed here.

Infinite Horizon Drawdown Minimization

In a seminal paper on the subject, Grossman and Zhou [15] show how to extend the Merton framework to encompass a more general drawdown constraint of $W_t \geq \alpha M_t$, where W_t is the wealth level at time t , M_t is the running maximum wealth up to that point, that is, $M_t = \max_{0 \leq u \leq t} W_u$, and α is an exogenous number between 0 and 1.

The motivation for this constraint is that, in practice, a fund manager may implicitly be subject to redemptions that depend on whether the manager's portfolio stays above a (possibly discounted) previous high, M_t .

For an investor with constant relative risk aversion utility, Grossman and Zhou [15] show that the optimal policy implies an investment in the risky asset at time t in proportion to the "surplus" $W_t - \alpha M_t$, that is, a CPPI ([3]) strategy with floor given by a multiple of the running maximum wealth, $F_t = \alpha M_t$. The analysis of Grossman and Zhou [15] is extended in [12] by allowing for intermediate consumption.

Infinite Horizon Shortfall Minimization

There are a variety of approaches and objectives related to shortfall minimization. For example, in a stochastic model where rational investment strategies may enable wealth to hit a given shortfall (e.g., liability driven models), one might choose a strategy that minimizes the *probability* of ever hitting this shortfall level directly ([5, 6, 8, 9, 11, 13, 19, 20]), or one might incorporate a (expected) *shortfall* constraint into other objectives such as a standard utility maximization framework ([1, 2, 4, 20] and [22]).

Directly minimizing the probability of hitting a shortfall level is a relevant objective only in economic settings where there is a possibility that a shortfall level is reached under a rational investment strategy. One such setting is the case of external risk, such as an insurance company, or liabilities as treated in [5], where the investor's total wealth evolves now according to a combination of an uncontrolled

Brownian motion (the "risk" part), and a controlled geometric Brownian motion (the investment possibility). Drawdown and shortfall prevention strategies for deterministic liabilities are treated in [6] where it is shown that if the initial wealth or reserve is below the funding level of the perpetual liability, the optimal strategy is linear in the negative surplus, that is, has an *inverse* CPPI structure, namely $k(F - W)$. For initial wealth above the funding level, various CPPI strategies using the funding level as the floor are optimal for a variety of utility and probabilistic objectives.

Other settings of interest include cases where there is an exogenous and uncontrollable benchmark relative to which the shortfall is measured. Infinite horizon probabilistic objectives are treated in [8] in an incomplete market setting, and connects these results with those obtained from standard utility maximization problems. A risk-constrained problem that yields a CPPI related strategy is treated in [11]. Stutzer [19] treats long-horizon shortfalls and deviations from benchmarks in the context of a large-deviation approach.

Finite Horizon Shortfall Minimization

The structure of the optimal strategy changes significantly when the horizon is finite in that the optimal strategies become *replicating strategies* for various structures of finite term options. Specifically, the strategy that minimizes the *shortfall probability* starting from a wealth process below the target level is given by the replicating strategy for a digital or binary put option on the shortfall level. This is discussed in [9, 10] and in an equivalent hedging framework in [13]. The optimal dynamic policy in the case of multiple risky assets has a time-dependent component (determined by the risk premium and remaining time) and a state-dependent component, which is a function of the current percentage of the distance to the target. The minimization is treated in [13] in the context of determining a partial hedging strategy on a contingent claim that minimizes the hedging cost for a given shortfall probability. This strategy may be considered a dynamic version of the static Value-at-Risk (VaR) concept and the authors label it *quantile hedging*. The potential riskiness of such a strategy is illustrated in [9, 10] via the fact that since it replicates a digital or

binary option, the strategy effectively acts as the delta of the digital option with all the instability of that delta as the term decays if the strike remains unachieved. Strategies that mitigate this fact and therefore minimize the *expected shortfall* are constructed in [14]. These strategies effectively replicate options with standard put payoffs as opposed to digitals or binaries.

Utility Maximization Approaches to the Expected Shortfall

Utility maximization approaches to the expected shortfall problem also lead to optimal strategies that have optionlike features. Basak and Shapiro [1] consider an agent's utility maximization problem in a model with a VaR constraint, which states that the probability of his wealth falling below some "floor" $\underline{W}$ is not allowed to exceed some prespecified level α :

$$\Pr(W(T) \geq \underline{W}) \geq 1 - \alpha \quad (4)$$

and is clearly related to the objectives treated earlier by Browne, Föllmer, and Leukert [9, 13, 14] (*see Value-at-Risk*).

In this framework, the case $\alpha = 1$ corresponds to the standard benchmark agent that does not limit losses and $\alpha = 0$ corresponds to the portfolio insurer (or put option purchaser) who maintains his wealth above the floor in all states [3].

Basak and Shapiro [1] show that the VaR constrained agent's wealth can be expressed as either (i) the portfolio insurer solution plus a short position in binary options or (ii) the benchmark agent's solution plus an appropriate position in "corridor" options (*see Corridor Options*). Similar to the analysis and earlier findings mentioned above they observe that since the VaR constrained agent is only concerned with the probability (and not the magnitude) of a loss, he or she chooses to leave the worst states uninsured because they are the most expensive ones to insure against. Thus, as in [14], Basak and Shapiro [1] examine a so-called LEL-RM (limited-expected-losses-based risk management) strategy, which remedies some of the shortcomings of the VaR constrained solution. Other related papers considering variants of these results are found in [2, 22]

References

- [1] Basak, S. & Shapiro, A. (2001). Value-at-risk-based risk management: optimal policies and asset prices, *Review of Financial Studies* **14**, 371–405.
- [2] Basak, S., Shapiro, A. & Tepla, L. (2004). Risk management with benchmarking, *Management Science* **52**, 542–557.
- [3] Black, F. & Perold, A.F. (2004). Theory of constant proportion portfolio insurance, *Journal of Economic Dynamics and Control* **16**, 403–426.
- [4] Boyle, P. & Tian, W. (2007). Portfolio management with constraints, *Mathematical Finance* **17**(3), 319–343.
- [5] Browne, S. (1995). Optimal investment policies for a firm with a random risk process: exponential utility and minimizing the probability of ruin, *Mathematics of Operations Research* **20**, 937–958.
- [6] Browne, S. (1997). Survival and growth with a liability: optimal portfolio strategies in continuous time, *Mathematics of Operations Research* **22**, 468–493.
- [7] Browne, S. (1998). The return on investment from proportional portfolio strategies, *Advances in Applied Probability* **30**(1), 216–238.
- [8] Browne, S. (1999). Beating a moving target: optimal portfolio strategies for outperforming a stochastic benchmark, *Finance and Stochastics* **3**, 275–294.
- [9] Browne, S. (1999). Reaching goals by a deadline: digital options and continuous-time active portfolio management, *Advances in Applied Probability* **31**, 551–577.
- [10] Browne, S. (1999). The risks and rewards of minimizing shortfall probability, *Journal of Portfolio Management* **25**(4), 76–85. Summer 1999.
- [11] Browne, S. (2000). Risk constrained dynamic active portfolio management, *Management Science* **46**(9), 1188–1199.
- [12] Dybvig, P.H. (1995). Duesenberry's ratcheting of consumption: optimal dynamic consumption and investment given intolerance for any decline in standard of living, *Review of Economic Studies* **62**, 287–313.
- [13] Föllmer, H. & Leukert, P. (1999). Quantile hedging, *Finance Stochastics* **3**, 251–273.
- [14] Föllmer, H. & Leukert, P. (2000). Efficient hedging: cost versus shortfall risk, *Finance and Stochastics* **4**, 117–146.
- [15] Grossman, S. & Zhou, Z. (1993). Optimal investment strategies for controlling drawdowns, *Mathematical Finance* **3**, 241–276.
- [16] Merton, R.C. (1971). Optimum consumption and portfolio rules in a continuous-time model, *Journal of Economic Theory* **3**, 373–413.
- [17] Merton, R.C. (1990). *Continuous Time Finance*, Blackwell: Cambridge.
- [18] Roy, A.D. (1952). Safety first and the holding of assets, *Econometrica* **20**(3), 431–449.

- [19] Stutzer, M.J. (2000). A portfolio performance index, *Financial Analysts Journal* **56**, 52–61.
- [20] Stutzer, M.J. (2004). Asset allocation without unobservable parameters, *Financial Analysts Journal* **60**(5), 38–51.
- [21] Telser, L.G. (1955). Safety first and Hedging, *Review of Economics and Statistics* **23**, 1–16.
- [22] Tepla, L. (2001). Optimal investment with minimum performance constraints, *Journal of Economic Dynamics and Control* **25**, 1629–1645.

Related Articles

Constant Proportion Portfolio Insurance; Corridor Options; Expected Shortfall; Fixed Mix Strategy; Hedge Funds; Merton Problem; Transaction Costs; Value-at-Risk.

SID BROWNE & ROBERT KOSOWSKI

Universal Portfolios

The mainstream mathematical approach to investment portfolio selection involves modeling asset prices as a stochastic process, and deriving asset allocation strategies that are optimal with respect to some statistical criteria, such as the mean and variance of returns, or an expected utility. The application of this approach to real-world markets is complicated by the lack of a fully specified stochastic process. It is, in fact, debatable whether an underlying stochastic process exists at all.

The theory of universal portfolios offers an alternative approach that forgoes stochastic models and associated optimality criteria and instead finds universal portfolio selection strategies that perform well relative to *all* strategies in a target class of allocation strategies, for *all* possible sequences of asset price changes. The term *universal portfolio* was coined in [3], which first put forth this alternative paradigm for the specific target class of constant rebalanced portfolios. This particular target class, which has also been the focus of much of the subsequent related work, seems to strike a favorable balance between the richness of the class and the degree to which the class performance can be tracked. The theory, in its greatest generality, however, can involve any target class, and would be concerned with finding the degree to which it can be tracked universally. The article focuses on universal portfolios for constant rebalanced portfolios, with only a brief mention of what is known beyond this.

Constant rebalanced portfolios [14] are a well-known (though under several aliases) class of allocation strategies that, at designated times, buy and sell assets to restore the fractions of wealth invested in each asset to an initial, fixed allocation. For a given sequence of rebalancing times, constant rebalanced portfolios can thus be parameterized by a vector with real valued, nonnegative components summing to one, specifying the fractions of wealth to which asset allocations are rebalanced. For example, given a pair of assets, the constant rebalanced portfolio corresponding to the vector $(1/3, 2/3)$ initially invests $1/3$ and $2/3$ of starting capital, respectively, in each of the assets. Thereafter, as the asset prices change, it buys and sells the assets at each rebalancing time to restore the fractions of cumulative wealth invested in each asset back to the initial $(1/3, 2/3)$.

One well-known motivating property of constant rebalanced portfolios is that they maximize the expectation of a variety of utility functions when asset price returns between rebalancing times are statistically independent and identically distributed from one interval to the next. In general, the optimal allocation depends on both the underlying probability distribution and the utility function, making the optimal choice, from this point of view, a challenging one for any given real-world setting. The theory of universal portfolios for constant rebalanced portfolios, which seeks to perform well relative to all portfolios, offers a principled way around this difficulty. It is worth reemphasizing, however, that beyond this motivational aspect, the theory makes *no* stochastic assumptions on asset prices.

Given a fixed number of assets and a sequence of asset price changes between n -designated rebalancing times, let S_n^* denote the largest return attained by any constant rebalanced portfolio on that sequence of asset price changes. This will be the benchmark for universal portfolio performance. Note that S_n^* and the (best) constant rebalanced portfolio that achieves it depend on the entire sequence of asset price changes, whereas the allocations of any realizable portfolio selection strategy must be causal, or nonanticipating, depending only on the previously observed price data. As such, the return S_n^* is not directly realizable. Furthermore, depending on the actual asset returns, S_n^* may exceed the return of any fixed constant rebalanced portfolio by an exponentially increasing factor. This means that naively using a seemingly safe choice like the uniform allocation might be exponentially suboptimal.

Nonetheless, nonanticipating universal portfolios have been found with returns $\hat{S}_n$ that approach S_n^* , in the sense that the exponential rates of return per rebalancing interval, $\hat{W}_n = (1/n) \log \hat{S}_n$ and $W_n^* = (1/n) \log S_n^*$, are guaranteed to satisfy $\hat{W}_n - W_n^* \geq -((m-1)/(2n)) \log n - O(1/n)$, uniformly for *all* sequences of asset prices. Since this relative exponential rate of return tends to 0 in n , it follows that if S_n^* is increasing at some exponential rate of growth (as would be hoped for and expected in practice), $\hat{S}_n$ will increase at the same exponential rate, asymptotically. The remainder of the article explains the universal portfolios achieving this benchmark performance.

Formal Definitions and Notation

Investments in m assets are to be adjusted at designated rebalancing times, with the gap between two consecutive such times constituting an investment period. The multiplicative factors by which the respective asset prices change over an investment period (gross returns) are denoted by a price relative vector $\mathbf{x} = (x_1, \dots, x_m)'$, with $x_j \geq 0$ specifying the return factor of asset j , and $\mathbf{v}'$ denoting the transpose of vector $\mathbf{v}$. The entire cumulative wealth is assumed to be reinvested at each rebalancing time and a non-negative portfolio vector $\mathbf{b} = (b_1, \dots, b_m)'$, satisfying $\sum_j b_j = 1$, denotes the fractions of cumulative wealth to invest in each asset (fraction b_j in asset j). If a rebalancing according to the portfolio vector $\mathbf{b}$ is followed by asset returns corresponding to the price relative vector $\mathbf{x}$, the cumulative wealth is multiplied by a factor of $\mathbf{b}'\mathbf{x} = \sum_{j=1}^m b_j x_j$, over that investment period. Consequently, a sequence of n rebalancings according to the portfolios $\mathbf{b}^n = \mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ and a corresponding sequence of ensuing price relatives $\mathbf{x}^n = \mathbf{x}_1, \dots, \mathbf{x}_n$ result in an overall return factor (cumulative gross return) of

$$S_n(\mathbf{x}^n, \mathbf{b}^n) = \prod_{i=1}^n \mathbf{b}_i' \mathbf{x}_i \quad (1)$$

A constant rebalanced portfolio reinvests according to a fixed $\mathbf{b}$ at each rebalancing time, resulting in a return factor of $S_n(\mathbf{x}^n, \mathbf{b}) = \prod_{i=1}^n \mathbf{b}' \mathbf{x}_i$. The best constant rebalanced portfolio $\mathbf{b}^*(\mathbf{x}^n)$ for a price relative sequence achieves $S_n^*(\mathbf{x}^n) = \max_{\mathbf{b} \in \mathcal{B}} S_n(\mathbf{x}^n, \mathbf{b})$, where $\mathcal{B}$ denotes the set of all valid portfolio vectors. A nonanticipating portfolio selection strategy is a sequence of functions $\hat{\mathbf{b}}_i(\cdot)$ that map previously occurring price relative vectors into portfolio vectors. The corresponding portfolio used at rebalancing time i can thus be expressed as $\hat{\mathbf{b}}_i(\mathbf{x}_1, \dots, \mathbf{x}_{i-1})$. As mentioned, universal portfolio theory for the target class of constant rebalanced portfolios seeks nonanticipating portfolio selection strategies that perform well relative to $S_n^*(\mathbf{x}^n)$ for all $\mathbf{x}^n$.

μ -Weighted Universal Portfolios

For a target class consisting of only two portfolio allocation strategies, a natural corresponding universal portfolio would split the starting capital into two pools and invest each pool according to the respective target portfolios, thereby achieving the average of

the target returns. The class of constant rebalanced portfolios is considerably more complex, since it is a continuum of portfolio selection strategies, but an extension of this idea still works well. Given a measure or distribution μ on the space of valid portfolio vectors $\mathcal{B}$ satisfying $\mu(\mathcal{B}) = 1$, the μ -weighted universal portfolio can be thought of as investing an infinitesimal fraction $d\mu(\mathbf{b})$ of initial capital according to each constant rebalanced portfolio $\mathbf{b}$, where $d\mu(\mathbf{b})$, informally, is the differential measure at $\mathbf{b}$. The return factor $\hat{S}_{n,\mu}(\mathbf{x}^n)$ of such a strategy for a sequence of price relatives $\mathbf{x}^n$, is the average, with respect to μ , of the returns of the underlying constant rebalanced portfolios, or more precisely,

$$\hat{S}_{n,\mu}(\mathbf{x}^n) = \int_{\mathcal{B}} S_n(\mathbf{x}^n, \mathbf{b}) d\mu(\mathbf{b}) \quad (2)$$

A more rigorous formulation of the μ -weighted universal portfolio is that at time i it uses the portfolio function,

$$\hat{\mathbf{b}}_i(\mathbf{x}^{i-1}) = \frac{\int_{\mathcal{B}} S_{i-1}(\mathbf{x}^{i-1}, \mathbf{b}) \mathbf{b} d\mu(\mathbf{b})}{\int_{\mathcal{B}} S_{i-1}(\mathbf{x}^{i-1}, \mathbf{b}) d\mu(\mathbf{b})} \quad (3)$$

which can be interpreted as a past performance weighted average of constant rebalanced portfolios. Setting $S_0(\mathbf{x}^0, \mathbf{b}) = 1$ and incorporating these portfolio vectors into equation (1), results in a telescoping product, with the overall return factor simplifying to equation (2).

The μ -weighted universal portfolio was first proposed in [3] for the special case of μ equal to the uniform distribution, or Lebesgue measure, on the set of valid portfolio vectors $\mathcal{B}$. Later, Cover and Ordentlich [5] refined the performance bounds for the uniform case and found that a different μ corresponding to the Dirichlet($1/2, \dots, 1/2$) distribution with density proportional to $1/\sqrt{b_1 b_2 \dots b_m}$ exhibits a better worst case performance against the benchmark $S_n^*(\mathbf{x}^n)$. Letting $\hat{S}_{n,U}(\mathbf{x}^n)$ and $\hat{S}_{n,D}(\mathbf{x}^n)$, respectively, denote the return factors of the uniform and Dirichlet-weighted universal portfolios, the following performance bounds were derived in [5]:

$$\begin{aligned} \min_{\mathbf{x}^n} \frac{\hat{S}_{n,U}(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} &= \left[\binom{n+m-1}{m-1} \right]^{-1}, \\ \min_{\mathbf{x}^n} \frac{\hat{S}_{n,D}(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} &= \frac{\Gamma(m/2)\Gamma(n+1/2)}{\Gamma(1/2)\Gamma(n+m/2)} \end{aligned} \quad (4)$$

where the minimizations are over all sequences of price relative vectors. Stirling's approximation of the Gamma function applied to the performance bound for the Dirichlet-weighted universal portfolio shows that it decreases in n as $c_m/n^{(m-1)/2}$, with c_m depending only on the number of assets m (and not n). After taking logarithms and normalizing by n , this gives the worst-case exponential rate of return of $\hat{S}_{n,D}$ relative to S_n^* mentioned in the introduction section of the article. Exact algorithms for computing (3) for the uniform and Dirichlet-weighted universal portfolios are given in [5]. Lower complexity approximate algorithms based on quantization [3] and randomization [2, 11] have also been proposed.

Max-min Universal Portfolios

It was shown in [16] that the maximum over all nonanticipating portfolio strategies of the minimum return relative to the benchmark best constant rebalanced portfolio is given by

$$\begin{aligned} & \max_{\{\hat{\mathbf{b}}_i(\cdot)\}} \min_{\mathbf{x}^n} \frac{\hat{S}_n(\mathbf{x}^n, \{\hat{\mathbf{b}}_i(\cdot)\})}{S_n^*(\mathbf{x}^n)} \\ &= \left[\sum_{n_1+\dots+n_m=n} \left(\frac{n!}{n^m} \prod_{j=1}^m \frac{n_j}{n_j!} \right) \right]^{-1} \end{aligned} \quad (5)$$

where the maximization is over all nonanticipating portfolio strategies, with $\hat{S}_n(\mathbf{x}^n, \{\hat{\mathbf{b}}_i(\cdot)\})$ denoting the return factor of such a strategy, and the summation on the right is over m -tuples of nonnegative integers summing to n . The max-min universal portfolio for horizon n achieves the game theoretic optimum of equation (5). Letting $\hat{S}_{n,\max-\min}(\mathbf{x}^n)$ denote its return factor, the ratio $\hat{S}_{n,\max-\min}(\mathbf{x}^n)/S_n^*(\mathbf{x}^n)$ at time n is, therefore, guaranteed to not fall below the right-hand side of equation (5). As for the Dirichlet-weighted universal portfolio, this relative return factor behaves like $\tilde{c}_m/n^{(m-1)/2}$, where $\tilde{c}_m$ improves on (is larger than) the corresponding constant for the Dirichlet case.

The max-min universal portfolio for horizon n can also be expressed as splitting the pool of starting capital among a collection of constituent strategies. In this case, there are m^n such strategies, identified with the sequences of length n of elements from the set $\{1, \dots, m\}$. The strategy corresponding

to the sequence $j^n = j_1, \dots, j_n$ invests the entire cumulative wealth at time i in asset j_i . The fraction of starting capital allocated to the constituent strategy corresponding to j^n is $V \prod_{j=1}^m (n_j/n)^{n_j}$ where V ensures that the fractions add up to one. It turns out that V equals the right-hand side of equation (5). The portfolio function $\hat{\mathbf{b}}_i(\mathbf{x}^{i-1})$ to use at each time i can be computed efficiently (polynomially in n) using an algorithm similar to that proposed in [5] (see also [16]) for the uniform and Dirichlet-weighted universal portfolios. Note that unlike these universal portfolios, the max-min universal portfolio entails a different sequence of portfolio functions for each horizon n .

The worst case sequences of price relatives that pin the ratio $\hat{S}_{n,\max-\min}(\mathbf{x}^n)/S_n^*(\mathbf{x}^n)$ to the right-hand side of equation (5) are those in which each price relative vector $\mathbf{x}_i$ has exactly one nonzero component.^a These are characteristic of horse race or gambling markets [5]. A conditionally max-min universal portfolio is proposed in [15] that “locks in” potential improvements in the relative performance ratio if $\mathbf{x}^n$ is more benign. The max-min universal portfolio and its performance bound is extended to a constrained set of price relative vectors in [4].

Extensions

Other universal portfolios for the target class of constant rebalanced portfolios that trade-off the above performance bounds for lower computational complexity are proposed in [1, 8, 13]. The uniform-weighted universal portfolio and the performance bound of [3] are extended to continuous time in [10]. In [16], a different continuous time universal portfolio is proposed that corresponds to the dynamic hedging strategy of a derivative security, termed the *hindsight-allocation option*, that pays the return of the best constant rebalanced portfolio computed in hindsight. Universal portfolios under transactions costs are pursued in [2, 9]. The former extends the performance bounds of the uniform-weighted universal portfolio to incorporate linear transactions costs and the latter considers an alternative target class (for $m = 2$ assets) of no-trade-region strategies that are known to be optimal under linear transactions costs in certain probabilistic settings.

Universal portfolios for alternative target classes are also proposed in [5] for side-information-dependent constant rebalanced portfolios, in [12, 17]

for piecewise constant rebalanced portfolios, and in [7] for a class of portfolio strategies that can depend on past asset prices. Finally, universal portfolios incorporating short sales and margin are considered in [6].

End Notes

^aThe max–min relative performance (5) continues to hold if the minimization over nonnegative sequences of price relatives is replaced by an infimum over strictly positive sequences. This also applies to (4).

References

- [1] Agarwal, A., Hazan, E., Kale, S. & Schapire, R.E. (2006). Algorithms for portfolio management based on the Newton method, in *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, Pittsburgh, PA, pp. 9–14.
- [2] Blum, A. & Kalai, A. (1999). Universal portfolios with and without transaction costs, *Machine Learning* **35**(3), 193–205.
- [3] Cover, T.M. (1991). Universal portfolios, *Mathematical Finance* **1**(1), 1–29.
- [4] Cover, T.M. (2004). Minimax regret portfolios for restricted stock sequences, *Proceedings of the IEEE Symposium on Information Theory*, Jun 2004, Chicago, IL, p. 140.
- [5] Cover, T.M. & Ordentlich, E. (1996). Universal portfolios with side information, *IEEE Transactions on Information Theory* **42**(2), 348–363.
- [6] Cover, T.M. & Ordentlich, E. (1998). Universal portfolios with short sales and margin, *Proceedings of the IEEE Symposium on Information Theory*, Aug 1998, Cambridge, MA, p. 174.
- [7] Cross, J.E. & Barron, A.R. (2003). Efficient universal portfolios for past-dependent target classes, *Mathematical Finance* **13**(2), 245–276.
- [8] Helmbold, D.P., Schapire, R.E., Singer, Y. & Warmuth, M.K. (1998). On-line portfolio selection using multiplicative updates, *Mathematical Finance* **8**(4), 325–347.
- [9] Iyengar, G. (2005). Universal investment in markets with transaction costs, *Mathematical Finance* **15**(2), 359–371.
- [10] Jamshidian, F. (1992). Asymptotically optimal portfolios, *Mathematical Finance* **2**(2), 131–150.
- [11] Kalai, A.T. & Vempala, S. (2002). Efficient algorithms for universal portfolios, *The Journal of Machine Learning Research* **3**(3), 423–440.
- [12] Kozat, S.S. & Singer, A.C. (2007). Universal constant rebalanced portfolios with switching, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, Vol. 3, pp. 1129–1132.
- [13] Merhav, N. & Feder, M. (1993). Universal schemes for sequential decision from individual data sequences, *IEEE Transactions on Information Theory* **39**(4), 1280–1292.
- [14] Mulvey, J. (2009). Constantly rebalanced portfolios, *Encyclopedia of Quantitative Finance* **1**, 346.
- [15] Ordentlich, E. (1996). *Universal Investment and Universal Data Compression*, Ph.D. Thesis, Stanford University.
- [16] Ordentlich, E. & Cover, T.M. (1998). The cost of achieving the best portfolio in hindsight, *Mathematics of Operations Research* **23**(4), 960–982.
- [17] Singer, Y. (1997). Switching portfolios, *International Journal of Neural Systems* **8**(4), 445–455.

ERIK ORDENTLICH

Risk-sensitive Asset Management

In risk-sensitive asset management, maximization of the criterion

$$J(v_0; T) = \frac{1}{\gamma} \log E[e^{\gamma \log V_T}] = \frac{1}{\gamma} \log E[V_T^\gamma] \quad (1)$$

for $\gamma < 1, \neq 0$ is considered, where V_T is the total wealth an investor possesses, defined by $V_T = \sum_i N_T^i S_T^i$ with N_T^i the number of shares invested into i th security S_T^i at time T and v_0 the initial wealth. It is equivalent to expected power utility maximization with the criterion $\frac{1}{\gamma} E[V_T^\gamma]$. Looking at asymptotics as $\gamma \rightarrow 0$,

$$\begin{aligned} \frac{1}{\gamma} \log E[e^{\gamma \log V_T}] &\sim E[\log V_T] \\ &+ \gamma \text{Var}[\log V_T] + O(\gamma^2) \end{aligned} \quad (2)$$

we see that maximizing the criterion amounts asymptotically to maximizing expected log utility while minimizing its variance if $\gamma < 0$. In that sense, $\gamma < 0$ means risk averse. On the other hand, $\gamma > 0$ means risk seeking since it comes to maximizing log utility as well as its variance.

The infinite time horizon problem of maximizing

$$\overline{\lim}_{T \rightarrow \infty} \frac{1}{T} J(v_0, x; T) = \overline{\lim}_{T \rightarrow \infty} \frac{1}{\gamma T} \log E[e^{\gamma \log V_T}] \quad (3)$$

is often considered in an incomplete market model, where security prices are defined by

$$dS^0(t) = r(X_t) S^0(t) dt \quad (4)$$

$$dS^i(t) = S^i(t) \left\{ \alpha^i(X_t) dt + \sum_{k=1}^{n+m} \sigma_k^i(X_t) dW_t^k \right\} \quad (5)$$

$i = 1, \dots, m$, with an $(n+m)$ -dimensional Brownian motion process $W_t = (W_t^1, W_t^2, \dots, W_t^{n+m})$ defined on a filtered probability space $(\Omega, \mathcal{F}, P; \mathcal{F}_t)$. Its volatilities σ , instantaneous mean returns α , and interest rate r , are affected by economic factors

$(X_t^1, \dots, X_t^n)$ defined as the solution of the stochastic differential equation

$$dX_t = \beta(X_t) dt + \lambda(X_t) dW_t, \quad X(0) = x \in R^n \quad (6)$$

Introducing portfolio proportion h_t^i invested into i th security defined by $h^i(t) = \frac{N^i(t) S^i(t)}{V(t)}$ for each $i = 0, \dots, m$ and setting $h(t) = (h^1(t), h^2(t), \dots, h^m(t))$, the total wealth V_t turns out to satisfy

$$\begin{aligned} \frac{dV(t)}{V(t)} &= \{r(X_t) + h(t)^* \hat{\alpha}(X_t)\} dt \\ &+ h(t)^* \sigma(X_t) W_t \end{aligned} \quad (7)$$

under the self-financing condition, where $\hat{\alpha}(x) = \alpha(x) - r(x)\mathbf{1}$. In these maximization problems, the portfolio proportion h_t is considered an investment strategy and assumed to be $\mathcal{G}_t^{S, X} := \sigma(S(u), X(u), u \leq t)$ progressively measurable in the case of full information. The problem is often considered under partial information where h_t is assumed to be $\mathcal{G}_t^S := \sigma(S(u), u \leq t)$ measurable. Here we discuss the case of full information and the set of admissible strategies $\mathcal{A}(T)$ (or $\mathcal{A}$) is determined as the totality of $\mathcal{G}_t^{S, X}$ progressively measurable investment strategies satisfying some suitably defined integrability conditions.

If $\gamma < 0$, introducing the value function

$$\hat{v}(t, x) = \inf_{h \in \mathcal{A}(T-t)} \log E[e^{\gamma \log V_{T-t}(h)}] \quad (8)$$

we see that

$$\sup_h J(v_0, x; T) = \frac{1}{\gamma} \hat{v}(0, x) \quad (9)$$

Under the change of measure,

$$P^h(A) = E \left[e^{\gamma \int_0^T h_s^* \sigma(X_s) dW_s - \frac{\gamma^2}{2} \int_0^T h_s^* \sigma \sigma^*(X_s) h_s ds} : A \right] \quad (10)$$

the value function is expressed as

$$\hat{v}(t, x) = \gamma \log v_0 + \inf_{h \in \mathcal{A}(T)} \log E^h \left[e^{\gamma \int_0^{T-t} \eta(X_s, h_s) ds} \right] \quad (11)$$

with the initial wealth v_0 , where

$$\eta(x, h) = h^* \hat{\alpha}(x) - \frac{1-\gamma}{2} h^* \sigma \sigma^*(x) h + r(x) \quad (12)$$

By using the Brownian motion $W_t^h := W_t - \gamma \int_0^t \sigma^*(X_s) h_s ds$ under the new probability measure P^h the dynamics of economic factor X_t is written as

$$dX_t = \{\beta(X_t) + \gamma \lambda \sigma^*(X_t) h_t\} dt + \lambda(X_t) dW_t^h \quad (13)$$

Thus the Hamilton–Jacobi–Bellman (H–J–B) equation for the value function is deduced as

$$\begin{cases} \frac{\partial v}{\partial t} + \frac{1}{2} \text{tr}[\lambda \lambda^* D^2 v] + \frac{1}{2} (Dv)^* \lambda \lambda^* Dv \\ \quad + \inf_h \{[\beta + \gamma \lambda \sigma^* h]^* Dv + \gamma \eta(x, h)\} = 0 \\ v(T, x) = \gamma \log v_0 \end{cases} \quad (14)$$

which can be rewritten as

$$\begin{cases} \frac{\partial v}{\partial t} + \frac{1}{2} \text{tr}[\lambda \lambda^* D^2 v] + \beta_\gamma^* Dv \\ \quad + \frac{1}{2} (Dv)^* \lambda N_\gamma^{-1} \lambda^* Dv - U_\gamma = 0 \\ v(t, x) = \gamma \log v_0 = 0 \end{cases} \quad (15)$$

where

$$\beta_\gamma = \beta + \frac{\gamma}{1-\gamma} \lambda \sigma^* (\sigma \sigma^*)^{-1} \hat{\alpha} \quad (16)$$

$$N_\gamma^{-1} = I + \frac{\gamma}{1-\gamma} \sigma^* (\sigma \sigma^*)^{-1} \sigma \quad (17)$$

$$U_\gamma = -\frac{\gamma}{2(1-\gamma)} \hat{\alpha}^* (\sigma \sigma^*)^{-1} \hat{\alpha} + r(x) \quad (18)$$

Under suitable conditions the H–J–B equation (15) has a solution with sufficient regularity [1, 13]. Moreover, if the condition

$$\hat{\alpha}^* (\sigma \sigma^*)^{-1} \hat{\alpha} \rightarrow \infty, |x| \rightarrow \infty \quad (19)$$

is assumed, then the solution such that $v(t, x) \rightarrow -\infty, \forall t < T, \text{ as } |x| \rightarrow \infty$ is unique and the identification

$$v(0, x; T) \equiv v(0, x) = \hat{v}(0, x) \equiv \inf_{h \in \mathcal{A}(T)} J^0(v, x; h; T) \quad (20)$$

can be verified.

The value for the problem on infinite time horizon is defined as

$$\hat{\chi}(\gamma) = \inf_{h \in \mathcal{A}} \chi(h; \gamma) \quad (21)$$

$$\chi(h; \gamma) = \lim_{T \rightarrow \infty} \frac{1}{T} \log E[V_T(h)^\gamma] \quad (22)$$

by suitably setting the set $\mathcal{A}$ of admissible strategies. The corresponding H–J–B equation of ergodic type for the problem is considered

$$\begin{aligned} \chi(\gamma) &= \frac{1}{2} \text{tr}[\lambda \lambda^* D^2 w] + \frac{1}{2} (Dw)^* \lambda \lambda^* Dw \\ &\quad + \inf_h \{[\beta + \gamma \lambda \sigma^* h]^* Dw + \gamma \eta(x, h)\}, \\ &= \frac{1}{2} \text{tr}[\lambda \lambda^* D^2 w] + \beta_\gamma^* Dw \\ &\quad + \frac{1}{2} (Dw)^* \lambda N_\gamma^{-1} \lambda^* Dw - U_\gamma \end{aligned} \quad (23)$$

However, even if we set as $\mathcal{A} = \{h_{\cdot|[0, T]} \in \mathcal{A}(T), \forall T\}$ identification of $\hat{\chi}(\gamma)$ with the solution $\chi(\gamma)$ to the H–J–B equation (23) cannot be shown in general. Indeed, even in the case of a linear Gaussian model (see below), such identification cannot be seen always to hold [5, 11, 13]. Instead, we introduce the asymptotic value

$$\tilde{\chi}(\gamma) = \lim_{T \rightarrow \infty} \frac{1}{T} v(0, x; T) \quad (24)$$

Then, a general discussion is possible for linear Gaussian models. Since the verification $v(0, x) = \hat{v}(0, x)$ holds in general for the problem on a finite time horizon, in the case of linear Gaussian models,

$$\lim_{T \rightarrow \infty} \frac{1}{T} \inf_{h \in \mathcal{A}} J^0(v, x; h; T) = \tilde{\chi}(\gamma) \quad (25)$$

is verified. Assume that $r(x) = r$, $\alpha(x) = Ax + a$, $\sigma(x) = \Sigma$, $\beta(x) = Bx + b$, $\lambda(x) = \Lambda$, where A, B, Σ, Λ are constant matrices, a, b are constant vectors and r is a constant. Then, the solution to equation (15) has an explicit expression as $v(t, x) = \frac{1}{2} x^* P(t) x + q(t)^* x + k(t)$, where $P(t)$ is a nonpositive definite solution to the Riccati equation

$$\begin{aligned} \dot{P}(t) + P(t) \Lambda N^{-1} \Lambda^* P(t) + K_1^* P(t) \\ + P(t) K_1 - C^* C = 0, \quad P(T) = 0 \end{aligned} \quad (26)$$

and $q(t)$, $K(t)$ are respectively solutions to

$$\begin{aligned} \dot{q}(t) + (K_1 + \Lambda N^{-1} \Lambda P(t))^* q(t) + P(t)b \\ + \frac{\gamma}{1-\gamma} (A^* + P(t) \Lambda \Sigma^*) \\ \times (\Sigma \Sigma^*)^{-1} \hat{a} = 0, \quad q(T) = 0 \end{aligned} \quad (27)$$

and

$$\begin{aligned} \dot{k}(t) + \frac{1}{2} \text{tr}[\Lambda \Lambda^* P(t)] + \frac{1}{2} q(t)^* \Lambda \Lambda^* q(t) \\ + \frac{\gamma}{2(1-\gamma)} (\hat{a} + \Sigma \Lambda^* q(t))^* (\Sigma \Sigma^*)^{-1} \\ \times (\hat{a} + \Sigma \Lambda^* q(t)) = 0 \end{aligned} \quad (28)$$

$$k(T) = \gamma \log v_0 \quad (29)$$

where

$$K_1 := B + \frac{\gamma}{1-\gamma} \Lambda \Sigma^* (\Sigma \Sigma^*)^{-1} A \quad (30)$$

$$C := \sqrt{-\frac{\gamma}{1-\gamma}} \Sigma^* (\Sigma \Sigma^*)^{-1} A \quad (31)$$

$$N^{-1} := I + \frac{\gamma}{1-\gamma} \Sigma^* (\Sigma \Sigma^*)^{-1} \Sigma \quad (32)$$

If $G := B - \Lambda \Sigma (\Sigma \Sigma^*)^{-1} A$ is stable, then $P(t) = P(t; T)$, $q(t) = q(t; T)$ converge as $T \rightarrow \infty$ respectively to $\bar{P}$, $\bar{q}$, which are respectively solutions to

$$K_1^* \bar{P} + \bar{P} K_1 + \bar{P} \Lambda N^{-1} \Lambda^* \bar{P} - C^* C = 0 \quad (33)$$

$$\begin{aligned} (K_1 + \Lambda N^{-1} \Lambda^* \bar{P})^* \bar{q} + \bar{P} b \\ + \frac{\gamma}{1-\gamma} (A^* + \bar{P} \Lambda \Sigma^*) (\Sigma \Sigma^*)^{-1} \hat{a} = 0 \end{aligned} \quad (34)$$

and $-\dot{k}(t) = -\dot{k}(t; T)$ converges to $\chi(\gamma)$ determined by

$$\begin{aligned} \chi(\gamma) = \frac{1}{2} \text{tr}[\Lambda \Lambda^* \bar{P}] + \frac{1}{2} \bar{q}^* \Lambda \Lambda^* \bar{q} \\ + \frac{\gamma}{2(1-\gamma)} (\hat{a} + \Sigma \Lambda^* \bar{q})^* (\Sigma \Sigma^*)^{-1} \\ \times (\hat{a} + \Sigma \Lambda^* \bar{q}) \end{aligned} \quad (35)$$

The nonpositive definite solutions to equations (33) and (34) are unique and $(K_1 + \Lambda N^{-1} \Lambda^* \bar{P})$ is stable under the present assumptions. Thus $\frac{1}{T} v(0; x; T)$

converges to $\tilde{\chi}(\gamma) = \chi(\gamma)$ and $(\chi(\gamma), w)$ defined by $w(x) = \frac{1}{2} x^* \bar{P} x + \bar{q}^* x$ turns out to be a solution to equation (23). If, furthermore,

$$\bar{P} \Lambda \Sigma^* (\Sigma \Sigma^*)^{-1} \Sigma \Lambda^* \bar{P} < A^* (\Sigma \Sigma^*)^{-1} A \quad (36)$$

holds, then one can show that $\chi(\gamma) = \hat{\chi}(\gamma)$ [11, 13]. The infimum in equation (23) is attained by $\hat{h}(x) = \frac{1}{1-\gamma} (\sigma \sigma^*)^{-1} (\hat{a} + \sigma \lambda^* D w)(x)$ and in the linear Gaussian model $\hat{h}(x) = \frac{1}{1-\gamma} (\Sigma \Sigma^*)^{-1} [\hat{a} + \Sigma \Lambda^* \bar{q} + (A + \Sigma \Lambda^* \bar{P})x]$. Thus, under condition (37) optimal strategy for $\hat{\chi}(\gamma)$ is defined as $\hat{h}_t = \hat{h}(X_t)$, $t < \infty$ [11]. Decomposing as $\hat{h}_t = \frac{1}{1-\gamma} \hat{h}_t^1 + \frac{1}{1-\gamma} \hat{h}_t^2 := \frac{1}{1-\gamma} (\Sigma \Sigma^*)^{-1} [\hat{a} + A X_t] + \frac{1}{1-\gamma} (\Sigma \Sigma^*)^{-1} [\Sigma \Lambda^* \bar{q} + \Sigma \Lambda^* \bar{P} X_t]$, Davis–Lleo [4] regard this decomposition as a generalization of Merton’s mutual funds theorem (see **Merton Problem**). Here $\hat{h}_t^1$ is a log utility portfolio (Kelly portfolio, see below and in **Kelly Problem**).

When $0 < \gamma < 1$ maximizing the criterion

$$\check{\chi}(h; \gamma) = \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \log E[V_T(h)]^\gamma \quad (37)$$

is considered. As a generic structure it can be seen that there exists γ^f such that $\check{\chi}(\gamma) = \sup_h \check{\chi}(h; \gamma)$ diverges for $\gamma^f < \gamma < 1$. However, it is only in one-dimensional linear Gaussian models that one can find the infimum of such γ^f explicitly [6].

The problems under bench-marked setting can be considered similarly (cf [2, 4]).

Noting that

$$\begin{aligned} \frac{1}{T} \log V_T(h) &= \frac{-1}{2T} \int_0^T \{h_t - \rho^{-1} \hat{a}(X_t)\}^* \\ &\quad \times \rho \{h_t - \rho^{-1} \hat{a}(X_t)\} dt \\ &\quad + \frac{1}{2T} \int_0^T \{r(X_t) + \hat{a}^* \rho^{-1} \hat{a}(X_t)\} dt \\ &\quad + \frac{1}{T} \int_0^T h_t^* \sigma(X_t) dW_t \end{aligned} \quad (38)$$

where $\rho = \sigma \sigma^*$, $h_t^K := (\sigma \sigma^*)^{-1} \hat{a}(X_t)$ turns out to maximize pathwise the growth rate of $V_T(h)$ on the long run and it is called *Kelly portfolio* (log-utility portfolio) [10] or *numéraire portfolio*. This is a control problem at the level of the law of large numbers.

The problem of maximizing the criterion

$$\overline{J}(\kappa, h) = \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \log P(\log V_T(h) \geq \kappa T) \quad (39)$$

is a kind of large deviation control problem and it is considered as the dual to risk-sensitive asset management in the risk-seeking case $0 < \gamma < 1$ [7, 9, 17, 18, 20]. On the other hand, the problem of minimizing the criterion

$$\underline{J}(\kappa, h) = \underline{\lim}_{T \rightarrow \infty} \frac{1}{T} \log P(\log V_T(h) \leq \kappa T) \quad (40)$$

is also a kind of large deviation control problem and it is considered as the dual to risk-sensitive asset management in the risk-averse case $\gamma < 0$ [3, 8, 15, 20]. Studies of these problems are still in progress.

Choosing as admissible strategies the set of all $\mathcal{G}_t^S := \sigma(S(u), u \leq t)$ progressively measurable processes satisfying some integrability conditions, the problems under partial information are considered as well [7, 12, 14–16, 19].

References

- [1] Bensoussan, A., Frehse, J. & Nagai, H. (1998). Some results on risk-sensitive control with full information, *Applied Mathematics and Optimization* **37**, 1–41.
- [2] Browne, S. (1999). Beating a moving target : optimal portfolio strategies for outperforming a stochastic benchmark, *Finance and Stochastics* **3**, 275–294.
- [3] Browne, S. (1999). The risk and rewards of minimizing shortfall probability, *Journal of Portfolio Management* **25**(4), 76–85.
- [4] Davis, M. & Lleo, S. (2008). Risk-sensitive benchmarked asset management, *Quantitative Finance* **8**, 415–426.
- [5] Fleming, W.H. & Sheu, S.J. (1999). Optimal long term growth rate of expected utility of wealth, *Annals of Applied Probability* **9**(3), 871–903.
- [6] Fleming, W.H. & Sheu, S.J. (2002). Risk-sensitive control and an optimal investment model. II, *Annals of Applied Probability* **12**(2), 730–767.
- [7] Hata, H. & Iida, Y. (2006). A risk-sensitive stochastic control approach to an optimal investment problem with partial information, *Finance and Stochastics* **10**, 395–426.
- [8] Hata, H., Nagai, H. & Sheu, S.J. Asymptotics of the probability minimizing a “down-side” risk, to appear in *Annals of Applied Probability*.
- [9] Hata, H. & Sekine, J. (2005). Solving long term optimal investment problems with Cox-Ingersoll-Ross interest rates, *Advances in Mathematical Economics* **8**, 231–255.
- [10] Kelly, J. (1956). A new interpretation of information rate, *Bell System Technical Journal* **35**, 917–926.
- [11] Kuroda, K. & Nagai, H. (2002). Risk sensitive portfolio optimization infinite time horizon, *Stochastics and Stochastics Reports* **73**, 309–331.
- [12] Nagai, H. (1999). Risk-sensitive dynamic asset management with partial information, in *Stochastics in Finite and Infinite Dimensions, a volume in honor of G. Kallianpur*, J. Xiong ed, Birkhäuser, pp. 321–340.
- [13] Nagai, H. (2003). Optimal strategies for risk-sensitive portfolio optimization problems for general factor models, *SIAM Journal of Control and Optimization* **41**, 1779–1800.
- [14] Nagai, H. (2004). Risk-sensitive portfolio optimization with full and partial information, *Stochastic Analysis and Related Topics, Advanced Studies in Pure Mathematics* **41**, 257–278.
- [15] Nagai, H. *Asymptotics of the probability minimizing a “down-side” risk under partial information*, Preprint.
- [16] Nagai, H. & Peng, S. (2002). Risk-sensitive dynamic portfolio optimization with partial information on infinite time horizon, *Annals of Applied Probability* **12**(1), 173–195.
- [17] Pham, H. (2003). A large deviations approach to optimal long term investment, *Finance and Stochastics* **7**, 169–195.
- [18] Pham, H. (2003). A risk-sensitive control dual approach to a large deviations control problem, *Systems and Control Letters* **49**, 295–309.
- [19] Rishel, R. (1999). Optimal portfolio management with partial observation and power utility function, *Stochastic Analysis, Control, Optimization and Applications*, a volume in honor of W.H. Fleming. 605–620.
- [20] Stutzer, M. (2003). Portfolio choice with endogeneous utility: a large deviations approach, *Journal of Econometrics* **116**, 365–386.

Further Reading

- Bielecki, T.R. & Pliska, S.R. (1999). *Risk sensitive dynamic asset management*, Applied Mathematics and Optimization **39**, 337–360.
- Merton, R.C. (1990). *Continuous Time Finance*, Blackwell, Malden.

Related Articles

Expected Utility Maximization: Duality Methods; Expected Utility Maximization; Kelly Problem; Merton Problem; Stochastic Control.

HIDEO NAGAI

Robust Portfolio Optimization

Portfolio selection is the problem of allocating capital over a number of available assets in order to maximize the “return” on the investment while minimizing the “risk”. Although the benefits of diversification in reducing risk have been appreciated since the inception of financial markets, Markowitz [25, 26] formulated the first mathematical model for portfolio selection. In the Markowitz portfolio selection model, the “return” on a portfolio is measured by the expected value of the random portfolio return, and the associated “risk” is quantified by the variance of the portfolio return. Markowitz showed that, given either an upper bound on the risk that the investor is willing to take or a lower bound on the return the investor is willing to accept, the optimal portfolio can be obtained by solving a convex quadratic programming problem. This mean–variance model has had a profound impact on the economic modeling of financial markets and the pricing of assets—the capital asset pricing model (CAPM) developed primarily by Sharpe [30], Lintner [22], and Mossin [28] was an immediate logical consequence of the Markowitz theory. In 1990, Sharpe and Markowitz shared the Nobel Memorial Prize in Economic Sciences for their work on portfolio allocation and asset pricing.

In spite of the theoretical success of the mean–variance model, practitioners have shied away from this model. The following quote from Michaud [27] summarizes the problem: “Although Markowitz efficiency is a convenient and useful theoretical framework for portfolio optimality, in practice it is an error prone procedure that often results in *error-maximized* and *investment-irrelevant* portfolios.” This behavior is a reflection of the fact that solutions of optimization problems are often very sensitive to perturbations in the parameters of the problem; since the estimates of the market parameters are subject to statistical errors, the results of the subsequent optimization are not very reliable. Various aspects of this phenomenon have been extensively studied in the literature on portfolio selection. Chopra and Ziemba [8] study the cash-equivalent loss from the use of estimated parameters instead of the true parameters. Broadie [5] investigates the influence of errors on the efficient frontier

and Chopra [7] investigates the turnover in the composition of the optimal portfolio as a function of the estimation error (see also Part II of [31] for a summary of this research). Several studies have shown that imposing constraints on the portfolio weights in a mean–variance optimization problem leads to better out-of-sample performance [15, 16]. Practitioners have always imposed no short-sale constraints and/or bounds for each security to improve diversity. It is suggested that constraining portfolio weights may reduce volatility, increase realized efficiency, and decrease downside risk or shortfall probability. Jagannathan and Ma [19] provide theoretical justification for these observations. Michaud [27] suggests resampling the mean returns μ and the covariance matrix Σ of the assets from a confidence region around a nominal set of parameters, and then aggregating the portfolios obtained by solving a Markowitz problem for each sample. Recently, scenario-based stochastic programming models have also been proposed for handling the uncertainty in parameters (see Part V of [31] for a survey of this research). Neither of the above two scenario based approaches provide any hard guarantees on the portfolio performance, and both become very inefficient as the number of assets grows.

Robust optimization is a deterministic optimization framework in which one explicitly models the parameter uncertainty, and the optimization problems are solved assuming worst-case behavior of these perturbations. This robust optimization framework was introduced in [3] for linear programming and in [2] for general convex programming [4]. There is also a parallel literature on robust formulations of optimization problems originating from robust control [10, 12] and [11].

In order to clearly understand the main ideas underlying the robust portfolio selection approach, consider the following very simple model. Suppose the true (unknown) covariance matrix Σ and the true (unknown) mean return vector μ are known to lie in *uncertainty sets*

$$S_m = \{\mu : (\mu - \mu_0)^T \Sigma_0^{-1} (\mu - \mu_0) \leq \beta_1^2\} \quad (1)$$

and

$$S_v = \{\Sigma : \|\Sigma - \Sigma_0\|_F \leq \beta_2\} \quad (2)$$

where $\|\mathbf{A}\|_F = \sqrt{\text{Tr}(\mathbf{A}^T \mathbf{A})}$. The confidence regions associated with maximum likelihood estimates of

the parameters (μ, Σ) have precisely the ellipsoidal structure described above and these confidence regions may be used as the uncertainty sets. The robust portfolio selection problem corresponding to the uncertainty sets S_m and S_v is given by

$$\max_{\{\phi: \mathbf{1}^T \phi = 1\}} \min_{\{\mu \in S_m, \Sigma \in S_v\}} \{\mu^T \phi - \tau \phi^T \Sigma \phi\} \quad (3)$$

that is, the utility of holding a portfolio ϕ is the worst-case utility when the parameters are allowed to vary in their uncertainty sets. Thus, robust optimization implicitly assumes Knightian uncertainty, that is, the market parameters are assumed to be ambiguous.

For fixed ϕ , the solution of the inner minimization problem

$$\begin{aligned} \min_{\{\mu \in S_m, \Sigma \in S_v\}} \{\mu^T \phi - \tau \phi^T \Sigma \phi\} \\ = \mu_0^T \phi - \beta_1 \sqrt{\phi^T \Sigma_0 \phi} - \tau (\phi^T (\Sigma_0 + \beta_2 \mathbf{I}) \phi) \end{aligned} \quad (4)$$

Thus, the robust portfolio selection problem is equivalent to

$$\max_{\{\phi: \mathbf{1}^T \phi = 1\}} \left\{ \mu_0^T \phi - \beta_1 \sqrt{\phi^T \Sigma_0 \phi} - \tau (\phi^T (\Sigma_0 + \beta_2 \mathbf{I}) \phi) \right\} \quad (5)$$

The objective function of this optimization problem can be reinterpreted in the following manner. The optimal portfolio ϕ^* is the optimal solution of the classical mean–variance optimization problem with

1. a perturbed mean vector:

$$\mu = \hat{\mu} - \left(\frac{\beta_1}{\sqrt{(\phi^*)^T \Sigma_0 \phi^*}} \right) \Sigma_0 \phi^* \quad (6)$$

that is, each component of the mean vector is adjusted to reduce the return on the portfolio ϕ^* , and

2. a perturbed covariance matrix

$$\Sigma = \Sigma_0 + \beta_2 \mathbf{I} \quad (7)$$

that is, the volatility of each of the assets is increased by an amount β_2 .

Thus, the robust portfolio selection problem can be interpreted as a modification of the classical mean–variance optimization problem where the

parameter values are dynamically adjusted to account for the uncertainty.

The optimization problem (5) can be reformulated as a second-order cone program (SOCP) (see [23] or [29] for details). This fact has important theoretical and practical implications. Since the computational complexity of an SOCP is comparable to that of a convex quadratic program, it follows that robust active portfolio selection is able to provide protection against parameter fluctuations at very moderate computational cost. Moreover, a number of commercial solvers such as MOSEK, CPLEX, and Frontline System (supplier of EXCEL SOLVER) provide the capability for solving SOCPs in a numerically robust manner.

The simple model described in the preceding text does not scale as the number of assets grows. At the very minimum, the data required to calculate the maximum likelihood estimate Σ_0 grows as $\mathcal{O}(n^2)$, where n is the number of assets. Goldfarb and Iyengar [17] work with a robust factor model wherein the single period return $\mathbf{r}$ is assumed to be a random variable given by

$$\mathbf{r} = \mu + \mathbf{V}^T \mathbf{f} + \boldsymbol{\varepsilon} \quad (8)$$

where $\mu \in \mathbf{R}^n$ is the vector of mean returns, $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{F}) \in \mathbf{R}^m$ is the vector of returns of the factors that drive the market, $\mathbf{V} \in \mathbf{R}^{m \times n}$ is the matrix of factor loadings of the n assets, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ is the vector of residual returns. For a detailed discussion of appropriate uncertainty sets for the parameters μ , $\mathbf{V}$, $\mathbf{F}$, and $\mathbf{D}$ and methods used to parameterize these sets from data, see Section 6 in [17].

Halldórsson and Tütüncü [18] show that if the uncertain mean return vector μ and the uncertain covariance matrix Σ of the asset returns $\mathbf{r}$ belong to the component-wise uncertainty sets $S_m = \{\mu : \mu^L \leq \mu \leq \mu^U\}$ and $S_v = \{\Sigma : \Sigma \succeq 0, \Sigma^L \leq \Sigma \leq \Sigma^U\}$, respectively, the robust problem reduces to a nonlinear saddle-point problem that involves semidefinite constraints. Here $\mathbf{A} \succeq 0$ (respectively > 0) denotes that the matrix $\mathbf{A}$ is symmetric and positive semidefinite (respectively definite). This approach has several shortcomings when applied to practical problems—the model is not a factor model (in applied work, factor models are popular because of the econometric relevance of the factors), no procedure is provided for specifying the extreme values (μ^L, μ^U) , and (Σ^L, Σ^U) defining the uncertainty structure and,

moreover, the solution algorithm, although polynomial, is not practicable when the number of assets is large. A multiperiod robust model, where the uncertainty sets are finite sets, was proposed in [1].

Recently, Delage and Ye [9] have proposed a distributionally robust model for portfolio selection. They assume that the distribution f of the random return ξ is uncertain and is assumed to belong to uncertainty sets of the form:

$$\mathcal{D}_1(\mathcal{S}, \mu_0, \Sigma_0, \gamma_1, \gamma_2) = \left\{ f : \begin{array}{l} \mathbb{P}_f(\xi \in \mathcal{S}) = 1 \\ (\mathbb{E}_f[\xi] - \mu_0)' \Sigma_0^{-1} (\mathbb{E}_f[\xi] - \mu_0) \leq \gamma_1 \\ \mathbb{E}_f[(\xi - \mu_0)(\xi - \mu_0)'] \leq \gamma_2 \Sigma_0 \end{array} \right\} \quad (9)$$

Notice that the uncertainty set for the covariance matrix $\mathbb{E}_f[(\xi - \mu_0)(\xi - \mu_0)']$ has an *upper* bound and the uncertainty set for the mean vector $\mathbb{E}_f[\xi]$ is centered around the nominal covariance matrix Σ_0 instead of the true covariance matrix. Delage and Ye consider robust portfolio selection problems of the form

$$\max_{\phi \in \Phi} \min_{f \in \mathcal{D}_1} \mathbb{E}_f[u(\phi, \xi)] \quad (10)$$

where $u(\phi, \xi)$ is a piece-wise linear concave utility function of the form $u(\phi, \xi) = \min_k \{a_k \xi' \phi + b_k\}$. Fix $\phi \in \Phi$. Then convex duality implies that the inner optimization problem $\min_{f \in \mathcal{D}_1} \mathbb{E}_f[u(\phi, \xi)]$ is equivalent to

$$\begin{aligned} \max \quad & r - t, \\ \text{s. t.} \quad & r \leq a_k \xi' \phi + b_k + \xi' Q \xi + \xi' q \quad \forall k, \forall \xi \in \mathcal{S} \\ & t \geq (\gamma_2 \Sigma_0 + \mu_0 \mu_0') \bullet Q + \mu_0' q \\ & \quad + \sqrt{\gamma_1} \|\Sigma_0^{\frac{1}{2}}(q + 2Q\mu_0)\| \\ & Q \succeq 0 \end{aligned} \quad (11)$$

where $A \bullet B = \text{Tr}(AB)$ denotes the Frobenius inner product of matrices. Note that equation (11) is a *semiinfinite* optimization problem since the first constraint has to hold for all $\xi \in \mathcal{S}$. Moreover, it is a semidefinite program, and is, therefore, a much harder problem compared to SOCPs or convex quadratic programs. Using results from convex optimization theory, Delage and Yu show that an ϵ -approximate

solution in $\mathcal{O}(n^{6.5} \log(\frac{1}{\epsilon}))$ iterations. This complexity is prohibitive for a portfolio selection problem of any reasonable size; however, it does provide a new perspective, namely, uncertainty sets for random returns. Lim *et al.* have studied a minimax regret formulation for portfolio selection in continuous time [21].

The robust optimization methodology has also been extended to *active* portfolio management problems where the goal is to beat a given benchmark by using information that is not broadly available in the market. Since errors in estimating the returns of assets are expected to have serious consequences for an active strategy, robust models are likely to result in portfolios with significantly superior performance. Erdoğan *et al.* [14] show how the basic robust portfolio selection model extends to robust active portfolio management. Since active portfolio strategies tend to execute many trades, properly modeling and managing trading costs are essential for the success of any practical active portfolio management model [20, 24]. Erdoğan *et al.* [14] show that a very large class of piecewise convex trading cost functions can be incorporated into an active portfolio selection problem in a tractable manner. Ceria and Stubbs [6] show how to impose a variety of side constraints on the exceptional return α and still recast the portfolio selection problem as an SOCP. Erdogan *et al.* [13] show how to incorporate analysts' views about α and nonparametric loss functions into robust active portfolio management.

References

- [1] Ben-Tal, A., Margalit, T. & Nemirovski, A. (2000). Robust modeling of multi-stage portfolio problems, in *High Performance Optimization*, Kluwer Academic Publishers, Dordrecht, pp. 303–328.
- [2] Ben-Tal, A. & Nemirovski, A. (1998). Robust convex optimization, *Mathematics of Operations Research* **23**(4), 769–805.
- [3] Ben-Tal, A. & Nemirovski, A. (1999). Robust solutions of uncertain linear programs, *Operations Research Letters* **25**(1), 1–13.
- [4] Ben-Tal, A. & Nemirovski, A. (2001). *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.
- [5] Broadie, M. (1993). Computing efficient frontiers using estimated parameters, *Annals of Operations Research* **45**, 21–58.

- [6] Ceria, S. & Stubbs, R.A. (2006). Incorporating estimation errors into portfolio selection: robust portfolio construction, *Journal of Asset Management* **7**, 109–127.
- [7] Chopra, V.K. (1993). Improving optimization, *Journal of Investing* **2**, (Fall), 51–59.
- [8] Chopra, V.K. & Ziemba, W.T. (1993). The effect of errors in means, variances and covariances on optimal portfolio choice, *Journal of Portfolio Management* **19**, (Winter), 6–11.
- [9] Delage, E. & Ye, Y. Distributionally robust optimization under moment uncertainty with applications to data-driven problem, Under review in *Operations Research*.
- [10] El Ghaoui, L. & Lebret, H. (1997). Robust solutions to least-squares problems with uncertain data, *SIAM Journal on Matrix Analysis and Applications* **18**(4), 1035–1064.
- [11] El Ghaoui, L. & Niculescu, N. (eds) (1999). *Recent Advances on LMI Methods in Control*, SIAM.
- [12] El Ghaoui, L., Oustry, F. & Lebret, H. (1998). Robust solutions to uncertain semidefinite programs, *SIAM Journal on Optimization* **9**(1), 33–52.
- [13] Erdoğan, E., Goldfarb, D. & Iyengar, G. (2006). *Robust Active Portfolio Management*, Technical Report TR-2004-11, Computational Optimization Research Center (CORC), IEOR Department, Columbia University, Available at <http://www.corc.ieor.columbia.edu/reports/techreports/tr-2004-11.pdf>
- [14] Erdoğan, E., Goldfarb, D. & Iyengar, G. (2008). Robust active portfolio management, *Journal of Computational Finance* **11**(4), 71–98.
- [15] Frost, P.A. & Savarino, J.E. (1986). An empirical Bayes approach to efficient portfolio selection, *Journal of Financial and Quantitative Analysis* **21**, 293–305.
- [16] Frost, P.A. & Savarino, J.E. (1988). For better performance: constrain portfolio weights, *Journal of Portfolio Management* **15**, 29–34.
- [17] Goldfarb, D. & Iyengar, G. (2003). Robust portfolio selection problems, *Mathematics of Operations Research* **28**(1), 1–38.
- [18] Halldórsson, B.V. & Tütüncü, R.H. (2000). *An Interior-point Method for a Class of Saddle Point Problems*. Technical report, Carnegie Mellon University, April 2000.
- [19] Jagannathan, R. & Ma, T. (2003). Risk reduction in large portfolios: why imposing the wrong constraints helps, *Journal of Finance* **58**, 1651–1683.
- [20] Kissell, R. & Glantz, M. (2003). *Optimal Trading Strategies: Quantitative Approaches for Managing Market Impact and Trading Risk*, AMACOM.
- [21] Lim, A.E.B., Shanthikumar, J.G. & Watwai, T. Robust asset allocation with benchmarked objectives, Under review in *Mathematical Finance*.
- [22] Lintner, J. (1965). Valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets, *Review of Economics and Statistics* **47**, 13–37.
- [23] Lobo, M.S., Vandenberghe, L., Boyd, S. & Lebret, H. (1998). Applications of second-order cone programming, *Linear Algebra and its Applications* **284**(1–3), 193–228.
- [24] Loeb, T.F. (1983). Trading cost: the critical link between investment information and results, *Financial Analysts Journal*, 39–44.
- [25] Markowitz, H.M. (1952). Portfolio selection, *Journal of Finance* **7**, 77–91.
- [26] Markowitz, H.M. (1959). *Portfolio Selection*, Wiley, New York.
- [27] Michaud, R.O. (1998). *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*, Harvard Business School Press, Boston.
- [28] Mossin, J. (1966). Equilibrium in capital asset markets, *Econometrica* **34**(4), 768–783.
- [29] Nesterov, Y. & Nemirovski, A. (1993). *Interior-point polynomial algorithms in convex programming*, SIAM, Philadelphia.
- [30] Sharpe, W. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk, *Journal of Finance* **19**(3), 425–442.
- [31] Ziemba, W.T. & Mulvey, J.M. (eds) (1998). *Worldwide Asset and Liability Modeling*, Cambridge University Press, Cambridge, UK.

GARUD IYENGAR

Diversification

Diversification involves spreading investments among various assets in order to improve portfolio performance in some manner. Mathematical analysis of portfolio diversification was introduced in 1952 by Markowitz [6] with his concept of mean/variance portfolio efficiency. A portfolio was called *efficient* if, among all portfolios of the same assets, the portfolio had minimum variance for a given expected rate of return. Hence, in the setting of expected portfolio return and variance, portfolio diversification served to control the risk of a portfolio. In 1982, Fernholz and Shay [3] showed that if the expected compound growth rate of a portfolio is considered rather than the expected rate of return, then portfolio diversification can increase the expected growth rate as well as control risk.

Mean/Variance Portfolio Diversification

Suppose that a portfolio $\mathbf{P}$ holds n assets $\mathbf{X}_1, \dots, \mathbf{X}_n$, and let $p_1, \dots, p_n$ be the weights, or proportions of each corresponding asset in the portfolio. In this case, the weights need not be all positive, but they must add up to 1, $p_1 + \dots + p_n = 1$. A negative value of p_i indicates a short sale of $\mathbf{X}_i$.

Suppose that α_i is the expected rate of return of $\mathbf{X}_i$ and that σ_{ij} is the covariance of return between $\mathbf{X}_i$ and $\mathbf{X}_j$, with the variance of return of $\mathbf{X}_i$ written as $\sigma_i^2 = \sigma_{ii}$. With this notation, the expected rate of return of $\mathbf{P}$ is given by

$$\alpha_P = \sum_{i=1}^n p_i \alpha_i \quad (1)$$

and the variance of $\mathbf{P}$ is

$$\sigma_P^2 = \sum_{i,j=1}^n p_i p_j \sigma_{ij} \quad (2)$$

In mean/variance theory [6, 7], a portfolio is optimally diversified if the portfolio variance σ_P^2 is minimal under the constraints

$$p_1 + \dots + p_n = 1 \quad \text{and} \quad \alpha_P \geq A \quad (3)$$

where A is a given constant. For a *long-only* portfolio, the additional constraints $p_1 \geq 0, \dots, p_n \geq 0$ are imposed.

In mean/variance theory, there is no specific measure of portfolio diversification, but it is understood that diversification can reduce the portfolio variance without lowering the expected return of the portfolio. Portfolios with minimum variance for a given value of portfolio return are called *efficient* portfolios, and we say that efficient portfolios lie on the *efficient frontier* in mean/variance space.

Diversification and Expected Growth Rate

The expected rate of return of a financial asset is more precisely called the expected *arithmetic* rate of return of the asset. Another measure of portfolio performance is the expected *logarithmic* rate of return of an asset, and this logarithmic rate is frequently called the expected (compound) *growth rate* of the asset. It was shown in [1, 3] that the expected growth rate of a financial asset is a better indicator of long-term performance than the expected return, and, for this reason, it is likely to be preferable for multiperiod performance analysis (see [4]). The relation between the expected rate of return α of an asset and its expected growth rate γ is

$$\alpha = \gamma + \frac{1}{2}\sigma^2 \quad (4)$$

where σ^2 is the variance of the asset. Equation (4) is an application of *Itô's rule* for stochastic integration [5]; the relation is exact for continuous-time analysis and approximate for multiperiod discrete-time analysis (see [1]).

For the portfolio $\mathbf{P}$, the portfolio variance σ_P^2 is the same whether it is measured with regard to the portfolio return or the portfolio growth rate. However, the relationship between the portfolio growth rate and the growth rates of the individual assets is more complicated than the corresponding relationship for arithmetic returns given by equation (1). If the assets $\mathbf{X}_1, \dots, \mathbf{X}_n$ have growth rates $\gamma_1, \dots, \gamma_n$, then the growth rate of $\mathbf{P}$ is given by

$$\gamma_P = \sum_{i=1}^n p_i \gamma_i + \gamma_P^* \quad (5)$$

where

$$\gamma_P^* = \frac{1}{2} \left(\sum_{i=1}^n p_i \sigma_i^2 - \sum_{i,j=1}^n p_i p_j \sigma_{ij} \right) \quad (6)$$

is called the *excess growth rate* of $\mathbf{P}$ (see [1–3]). From equation (2), we see that the excess growth rate is half the difference of the weighted average of the variances of the assets in the portfolio minus the variance of the portfolio itself. We see from equations (5) and (6) that diversification affects both the variance and the expected growth rate of the portfolio.

Excess Growth and the Efficacy of Diversification

The excess growth rate γ_P^* of the portfolio $\mathbf{P}$ can be used as a measure of the *efficacy of diversification* of the portfolio. The more effective the diversification, the greater the difference between the average variance of the assets and the portfolio variance, and the greater the contribution to the portfolio growth rate.

Equation (6) remains valid if the covariances σ_{ij} are replaced by covariances measured relative to some *numeraire* asset (see [1]), so in this sense the excess growth rate is *numeraire invariant*. Suppose $\mathbf{Z}$ is a financial asset; then the covariance of $\mathbf{X}_i$ and $\mathbf{X}_j$ relative to $\mathbf{Z}$ is given by

$$\sigma_{ij/Z} = \sigma_{ij} - \sigma_{iZ} - \sigma_{jZ} + \sigma_Z^2 \quad (7)$$

where σ_{iZ} is the covariance of $\mathbf{X}_i$ with $\mathbf{Z}$, σ_{jZ} is the covariance of $\mathbf{X}_j$ with $\mathbf{Z}$, and σ_Z^2 is the variance of $\mathbf{Z}$. With the notation $\sigma_{i/Z}^2 = \sigma_{ii/Z}$, equation (6) becomes

$$\gamma_P^* = \frac{1}{2} \left(\sum_{i=1}^n p_i \sigma_{i/Z}^2 - \sum_{i,j=1}^n p_i p_j \sigma_{ij/Z} \right) \quad (8)$$

In particular, if the asset $\mathbf{Z}$ is the portfolio $\mathbf{P}$ itself, then the variance of $\mathbf{P}$ relative to itself vanishes:

$$\sum_{i,j=1}^n p_i p_j \sigma_{ij/P} = \sigma_{P/P}^2 = 0 \quad (9)$$

so we have simply

$$\gamma_P^* = \frac{1}{2} \sum_{i=1}^n p_i \sigma_{i/P}^2 \quad (10)$$

If $\mathbf{P}$ is a long-only portfolio, then all the terms $p_i \sigma_{i/P}^2$ are nonnegative, and it follows from equation (10) that γ_P^* is also nonnegative. Hence, for such a portfolio, diversification will not decrease the portfolio growth rate, and is likely to increase it.

The Efficacy of Diversification in the US Stock Market

Here, we shall consider an example of the excess growth rate as a measure of efficacy of diversification. Figure 1 shows the smoothed, annualized excess growth rate for the US stock market over most of the twentieth century. The data used to construct Figure 1 come from the monthly stock database of the Center for Research in Securities Prices (CRSP) at the University of Chicago. The stocks included in this market are those traded on the New York Stock Exchange (NYSE), the American Stock Exchange (AMEX), and the NASDAQ Stock Market, with adjustments made for real estate investment trusts (REITs), closed-end funds, and American depository receipts (ADRs). Until 1962, the data included only NYSE stocks, after July 1962 AMEX stocks were included, and at the beginning of 1973 NASDAQ stocks were included. The number of stocks in this market varies from a few hundred in 1927 to about 7500 after 1999.

We see in Figure 1 that the excess growth rate of the market has varied considerably over time, from an estimated minimum excess growth rate

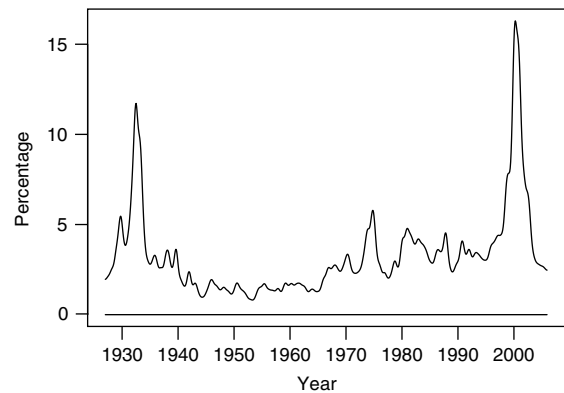


Figure 1 Efficacy of diversification of the US stock market, as measured by the market excess growth rate, 1927–2005

of about 1% a year in the 1950s to a maximum of about 16% a year near 2000. The volatility of the stocks has a significant effect on the efficacy of diversification, with higher excess growth rates appearing in the bubble years of the 1930s, the 1970s, and around 2000, even though during these periods there was concentration of capital into the larger stocks. In contrast, the excess growth rate increased only modestly with the increase in the number of stocks in the market from under 1000 in the early years to over 5000 in the later years. Hence, in the absence of higher volatility, the addition of new assets did not significantly increase the efficacy of diversification over the observed period.

References

- [1] Fernholz, R. (2002). *Stochastic Portfolio Theory*, Springer-Verlag, New York.
- [2] Fernholz, R. & Karatzas, I. (2008). Stochastic portfolio theory: an overview, in *Mathematical Modelling and Numerical Methods in Finance*, A. Bensoussan, Q. Zhang & P. Ciarlet, eds, Elsevier, Amsterdam.
- [3] Fernholz, R. & Shay, B. (1982). Stochastic portfolio theory and stock market equilibrium, *Journal of Finance* **37**, 615–624.
- [4] Hughson, E., Stutzer, M. & Yung, C. (2006). The misuse of expected returns, *Financial Analysts Journal* **62**, 88–96.
- [5] Itô, K. (1951). On stochastic differential equations, *Memiors of the American Mathematical Society* **4**, 1–51.
- [6] Markowitz, H. (1952). Portfolio selection, *Journal of Finance* **7**, 77–91.
- [7] Markowitz, H. (1959). *Portfolio Selection*, John Wiley & Sons, New York.

ROBERT FERNHOLZ

Merton Problem

How does an individual decide on where to invest his wealth and how much of it to use during his lifetime? This is a basic question that needs to be answered in order to understand and predict individual economic behavior and also in order to derive the aggregate demands for securities that, together with their supply schedules, determine their prices in equilibrium.

In two path-breaking papers, Merton [21, 22] has formulated and derived the individual's optimal consumption–investment behavior in a continuous-time framework that allows the introduction of a useful structure that can be appropriately molded to model different interesting situations and to yield concrete results. This formulation has come to be known as the *Merton problem*.

Formulation of the Merton Problem

Let $P(t) = (P_0(t), \dots, P_n(t))$, $0 \leq t \leq T$, denote the prices at time t of $n + 1$ limited liability assets paying no dividends and traded continuously in a perfect market. The price dynamics of $P(t)$ is assumed to follow a correlated vector Ito process whose i th component obeys the general stochastic differential equation

$$\frac{dP_i}{P_i} = \alpha_i(P, t) dt + \sigma_i(P, t) dz_i \quad (1)$$

For $i, j = 0, \dots, n$, the dz_i s are correlated Wiener processes satisfying $dz_i dz_j = \rho_{ij}(P, t) dt$ for given real functions $\rho_{ij}(\cdot, \cdot)$; $dz_i dt = 0$; the expectations $E(dz_i)$ equal zero; and denote $\sigma_{ij}(\cdot, \cdot) := \sigma_i \sigma_j \rho_{ij}$.

At time 0, the individual is endowed with an initial number of units of wealth W_0 and he then selects a complete plan of consumption and investment spanning his lifetime $[0, T]$ so as to maximize his expected utility of consumption and the bequest he will leave at time T .

Formally, at any future time t in $(0, T]$, based on the history of the prices, his previous consumption, and previous investment choices, the individual plans to consume at a rate $c(t)$ and trade so as to hold his remaining wealth $W(t)$ in a portfolio that is invested in $N_i(t)$ shares of asset i . The consumption and investment choices are said to be the individual's controls.

The individual plans to control his consumption and investment so as to

$$\max E_0 \left[\int_0^T u(c(t), t) dt + B(W(T)) \right] \quad (2)$$

subject to his wealth constraint

$$W(0) = W_0, W(t) = \sum_{i=0}^n N_i(t) P_i(t), \quad 0 \leq t \leq T \quad (3)$$

where u is the individual's instantaneous utility, and $B(\cdot)$ is the bequest function.

Solution by Dynamic Programming

To solve the Merton problem by dynamic programming, the wealth constraint (3) is needed in differential form. Taking the total stochastic differential in equation (3), noting that both the number of shares and their prices are Ito processes, results in

$$\begin{aligned} dW(t) = & \sum_{i=0}^n N_i(t) dP_i(t) + \sum_{i=0}^n dN_i(t) dP_i(t) \\ & + \sum_{i=0}^n dN_i(t) P_i(t) \end{aligned} \quad (4)$$

Clearly, the first term on the right-hand side (RHS) of equation (4) is associated with the capital gains to the portfolio over the interval dt resulting from the change in the asset prices. Equally clearly, the third term on the RHS of that equation is associated with the inflow of wealth from external sources that is used to buy additional shares for the portfolio (negative inflow would mean outflow, as when shares are sold to finance consumption.). It is not clear, though, how to associate the middle term on the RHS—with the capital gain or with the cash inflow to the portfolio.

Taking care that choices made at any given time do not anticipate the future, Merton [22] shows that the middle term on the RHS of equation (4), along with the third term, comprise together the total inflow of funds to the portfolio. Therefore, in the absence of other income, the incremental inflow to the portfolio at time t in this problem is given by $-c(t) dt = \sum_{i=0}^n dN_i(t) dP_i(t)$

2 Merton Problem

+ $\sum_{i=0}^n dN_i(t)P_i(t)$, and equation (4) becomes the Merton self-financing condition

$$dW(t) = \sum_{i=0}^n N_i(t) dP_i(t) - c(t) dt \quad (5)$$

It is remarkable that a special case of Merton's self-financing condition in equation (5) is equivalent to the Black–Scholes partial differential equation (PDE), which is prominent in derivative pricing.^a

It is convenient to express the Merton self-financing condition, equation (4), in terms of portfolio weights, $w_i(t) := N_i(t)P_i(t)/W(t)$, which would serve as the portfolio controls from now on instead of the number of shares. Substituting the weights and the asset returns from equation (1) in equation (5) yields the differential wealth dynamics,

$$\begin{aligned} dW(t) = & \left(-c + W \sum_{i=0}^n w_i \alpha_i \right) dt \\ & + W \sum_{i=0}^n w_i \sigma_i dz_i, \text{ with } \sum_{i=0}^n w_i = 1 \end{aligned} \quad (6)$$

which the individual obeys when solving for the consumption and investment controls in the expected utility maximization problem (2).

Since the utility functional in equation (2) is time-additive, the Merton problem can be solved by dynamic programming. To that end, define the value function (also called the *indirect utility of wealth*) by

$$\begin{aligned} J(W(t), t) := & \max_{c(\tau), \{w_i(\tau)\}, t \leq \tau \leq T} \\ & E_t \left[\int_t^T u(c(\tau), \tau) d\tau + B(W(T)) \right] \end{aligned} \quad (7)$$

subject to the wealth dynamics in equation (6), and where E_t denotes the expectation operator given the information at time t , that is, the knowledge at time t of the prices, wealth, and consumption rate that determine the conditional probabilities of the future prices.

Then, $J(W(t), t) = \max_{c(\tau), \{w_i(\tau)\}, t \leq \tau \leq T}$

$$\begin{aligned} & E_t \left[\int_t^{t+\Delta t} u(c(\tau), \tau) d\tau \right. \\ & \left. + \int_{t+\Delta t}^T u(c(\tau), \tau) d\tau + B(W(T)) \right] \\ = & \max_{c(\tau), \{w_i(\tau)\}, t \leq \tau \leq t+\Delta t} E_t \left\{ \int_t^{t+\Delta t} u(c(\tau), \tau) d\tau \right. \\ & \left. + \max_{c(\tau), \{w_i(\tau)\}, t+\Delta t \leq \tau \leq T} E_{t+\Delta t} \left[\int_{t+\Delta t}^T u(c(\tau), \tau) d\tau \right. \right. \\ & \left. \left. + B(W(T)) \right] \right\} \end{aligned}$$

(the law of iterated expectations was used above: $E_t E_{t+\Delta t}[\cdot] = E_t[\cdot]$)

$$\begin{aligned} = & \max_{c(\tau), \{w_i(\tau)\}, t \leq \tau \leq t+\Delta t} E_t \left\{ \int_t^{t+\Delta t} u(c(\tau), \tau) d\tau \right. \\ & \left. + J(W(t+\Delta t), t+\Delta t) \right\} \\ = & \max_{c(t), \{w_i(t)\}} E_t [u(c(t), t) \Delta t + o(\Delta t) \\ & + J(W(t), t) + \Delta J] \\ = & J(W(t), t) + \max_{c(t), \{w_i(t)\}} \{u(c(t), t) \Delta t + o(\Delta t) \\ & + E_t[\Delta J]\} \end{aligned} \quad (8)$$

where $o(x)$ means a quantity that tends to zero faster than does x .

By Ito's lemma and by the wealth equation (5),

$$\begin{aligned} E_t[\Delta J] = & E_t \left[J_t \Delta t + J_W \Delta W + \frac{1}{2} J_{WW} (\Delta W)^2 \right. \\ & \left. + o(\Delta t) \right] = J_t \Delta t + J_W \left(-c + W \sum_{i=0}^n w_i \alpha_i \right) \Delta t \\ & + \frac{1}{2} J_{WW} W^2 \left(\sum_{i=0}^n \sum_{j=0}^n w_i w_j \sigma_{ij} \right) \Delta t + o(\Delta t) \end{aligned} \quad (9)$$

Substituting the last expression in equation (8), subtracting $J(W(t), t)$ from both sides, dividing by Δt , and taking the limit $\Delta t \rightarrow 0$, yields

$$0 = \max_{c(t), \{w_i(t)\}_0^n} \left\{ u(c(t), t) + J_t \right.$$

$$\begin{aligned}
& + J_W \left(-c + W \sum_{i=0}^n w_i \alpha_i \right) + \frac{1}{2} J_{WW} W^2 \\
& \times \left(\sum_{i=0}^n \sum_{j=0}^n w_i w_j \sigma_{ij} \right) \Big\} \\
& \text{subject to } \sum_{i=0}^n w_i = 1
\end{aligned} \tag{10}$$

Equation (10) is the Hamilton–Jacobi–Bellman (HJB) equation for the Merton problem.

A Solved Example

The HJB equation yields a nonlinear partial differential equation in the unknown function of two variables $J(W, t)$ with the end condition $J(W, T) = B(W(T))$, the utility from bequest. A solution example is illustrated next for the case with one riskless asset yielding a constant rate of return $\alpha_0 = r$ and with n risky assets following correlated geometric Brownian motions, that is, α_i, σ_i , and σ_{ij} are all constants, and $\sigma_{0j} = \sigma_{j0} = 0$, for $i, j = 0, \dots, n$.

Substituting $w_0 = 1 - \sum_{i=1}^n w_i$ and $\sum_{i=0}^n w_i \alpha_i = r + \sum_{i=1}^n w_i (\alpha_i - r)$ in the HJB equation (10) yields

$$0 = \max_{c(t), \{w_i(t)\}_1^n} G(c(t), w_1(t), \dots, w_n(t)), \tag{11}$$

where $G(c, w_1, \dots, w_n) := u(c, t) + J_t + J_W [-c + rW + W \sum_{i=1}^n w_i (\alpha_i - r)] + (1/2) J_{WW} W^2 (\sum_{i=1}^n \sum_{j=1}^n w_i w_j \sigma_{ij})$ is a real function of $n+1$ free real variables, and the maximization problem exhibits no constraints.

To locate the point $(c^*, w_1^*, \dots, w_n^*)$ that maximizes G , requires the $n+1$ first-order conditions,

$$\frac{\partial G}{\partial c} = 0 = \frac{\partial u}{\partial c}(c, t) - J_w \tag{12}$$

$$\begin{aligned}
\frac{\partial G}{\partial w_i} &= 0 = J_w (\alpha_i - r) + J_{WW} W \sum_{j=1}^n w_j \sigma_{ij}, \\
&(i = 1, \dots, n)
\end{aligned} \tag{13}$$

If for every time t , $u(c, t)$ is strictly concave and twice continuously differentiable in c , then equation (12) can be inverted to yield $c^* = f(J_w, t)$. The system of linear equations (13) can be solved for

the weights on the risky assets w_i^* as functions of J_w, J_{WW} , and W . Then these $(c^*, w_1^*, \dots, w_n^*)$ are substituted back in equation (11) to yield

$$0 = G(c^*, w_1^*, \dots, w_n^*) \tag{14}$$

which then becomes a nonlinear partial differential equation of the second order in the unknown function $J(W, t)$ with the end condition $J(W, T) = B(W(T))$.

Merton [21] demonstrates closed-form solutions to equation (14) for some special cases. For example, assume that the instantaneous utility of consumption is the isoelastic, $u(c, t) = \exp(-\rho t) c^\gamma / \gamma$; utility from bequest is 0; there are two available assets, one is risk-free returning at a rate r and the other risky, following a Geometric Brownian motion with a constant drift, α , and a constant variance per unit time, σ^2 . Then the optimal weight that the portfolio puts on the risky asset is the constant $w^*(t) = (\alpha - r) / [(1 - \gamma)\sigma^2]$, $0 \leq t \leq T$, and the optimal consumption rate is given by $c^*(t) = W(t)a / [1 - \exp(-a(T - t))]$, $0 \leq t \leq T$, where $W(t)$ is the value of the portfolio at time t , and where

$$a := \frac{\gamma}{1 - \gamma} \left[\frac{\rho}{\gamma} - r - \frac{1}{2} \left(\frac{\alpha - r}{\sigma} \right)^2 \frac{1}{1 - \gamma} \right] \tag{15}$$

Generally, however, equation (14) must be solved numerically.

The Merton Problem with State-dependent Price Process Parameters

Merton [23] extends his problem to include parameters of the price processes that depend on a state variable x which itself follows an Ito process, that is,

$$\begin{aligned}
\frac{dP_i}{P_i} &= \alpha_i(x, t) dt + \sigma_i(x, t) dz_i, \\
(0 \leq t \leq T, i = 0, \dots, n)
\end{aligned} \tag{16}$$

$$dx = a(x, t) dt + s(x, t) dz \tag{17}$$

where the dz_i and dz are correlated Wiener processes satisfying $dz_i dz_i = \rho_{ij}(x, t) dt$; $dz_i dz = \rho_{ix} dt$; $dz_i dt = dz dt = 0$, $E(dz_i) = 0$; $(i, j = 0, \dots, n)$. Denote $\alpha := (\alpha_1, \dots, \alpha_n)$, $r := (r, \dots, r)$, $\sigma_x := (\rho_{1x}\sigma_1, \dots, \rho_{nx}\sigma_n)$, and σ denote by the matrix with $\sigma_{ij} := \sigma_i \sigma_j \rho_{ij}$ in the ij place.

Defining the indirect utility $J(W, x, t)$ as before, but recognizing that it now depends also on the state variable, and following the same derivation as before with the obvious modifications, Merton [23] shows that in the presence of a riskless asset returning at the constant rate r , the optimal investment plan described by the control vector process of weights on the risky assets in the portfolio that maximizes lifetime utility of consumption is given by

$$\begin{aligned} \mathbf{w}_t^* &= \left(-\frac{J_W}{W J_{WW}} \right) \sigma^{-1} (\boldsymbol{\alpha} - \mathbf{r}) + \left(-\frac{J_{xW}}{W J_{WW}} \right) \sigma^{-1} \sigma_x \\ &=: D\mathbf{d} + H\mathbf{h}, \quad 0 \leq t \leq T \end{aligned} \quad (18)$$

The scalars D and H are agent specific, but the vectors $\mathbf{d}$ and $\mathbf{h}$ are not. It follows that every investor behaves as if the risky part of his portfolio is split between two mutual funds holding total portfolios with weights that are proportional to $\mathbf{d}$ and $\mathbf{h}$, respectively; then there is also the part that is invested in the risk-free asset. The result is a three-fund separation theorem. Merton [23] shows that while the first risky mutual fund is used to diversify, that is, to obtain the largest expected return for a given amount of risk borne, the second risky mutual fund is used to hedge unfavorable shifts in the state variable x , in the sense that if an increase in x diminishes planned consumption, then the investor compensates himself by shifting wealth to the asset with returns that increase with x .

The Merton problem was first formulated and solved using stochastic control in [21, 22]. It provides the basis for the intertemporal capital asset pricing model in [23]. Extensions of the problem include [19, 20, 26]. In references [3, 4, 12–15], the problem is treated in incomplete markets and under other market constraints. Transaction costs are introduced into the Merton problem in [1, 5, 11, 29]. The problem is extended to incomplete information settings in [6, 8, 27]; to settings with habit formation utilities in [7, 17, 28]; and to settings with recursive utilities in [2, 10]. Textbooks that provide a detailed treatment of the Merton problem include [9, 16, 25].

End Notes

^aSpecifically, suppose the portfolio comprises two assets, one risky and one riskless, and that there are no inflows to or outflows from the portfolio, that is, $c(t) = 0$ for all t . Then, if only Markov controls of the portfolio are considered,

namely, number of units of the two assets that depend only on time and the concurrent price, it follows that the value of the portfolio which then obviously depends only on time and the concurrent price of the risky asset, can be shown to necessarily satisfy the Black–Scholes PDE. Moreover, the number of shares of the risky asset in the portfolio is equal to the partial derivative of the said portfolio value function with respect to the price of the risky asset (see [18, 24]). The converse is also true.

References

- [1] Akian, M., Menaldi, J.-L. & Sulem, A. (1996). On an investment-consumption model with transaction costs, *SIAM Journal on Control and Optimization* **34**, 329–364.
- [2] Bergman, Y.Z. (1985). Time preference and capital asset pricing models, *Journal of Financial Economics* **14**, 145–159.
- [3] Brennan, M., Schwartz, E. & Lagnado, R. (1997). Strategic asset allocation, *Journal of Economic Dynamics and Control* **21**, 1377–1403.
- [4] Cvitanic, J. & Karatzas, I. (1996). Hedging and portfolio optimization under transaction costs: a martingale approach, *Mathematical Finance* **6**, 133–165.
- [5] Davis, M. & Norman, A. (1990). Portfolio selection with transaction costs, *Mathematics of Operations Research* **15**, 676–713.
- [6] Detemple, J. (1986). Asset pricing in a production economy with incomplete information, *Journal of Finance* **41**, 383–391.
- [7] Detemple, J. & Zapatero, F. (1992). Optimal consumption-portfolio policies with habit formation, *Mathematical Finance* **2**, 251–274.
- [8] Dotan, M. & Feldman, D. (1986). Equilibrium interest rates and multiperiod bonds in a partially observable economy, *Journal of Finance* **41**, 369–382.
- [9] Duffie, D. (2001). *Dynamic Asset Pricing Theory*, Princeton University Press, Princeton.
- [10] Duffie, D. & Epstein, L. (1992). Asset pricing with stochastic differential utility, *Review of Financial Studies* **5**, 411–436.
- [11] Duffie, D. & Sun, T.S. (1990). Transactions costs and portfolio choice in a discrete-continuous time setting, *Journal of Economic Dynamics and Control* **14**, 35–51.
- [12] Dybvig, P. (1995). Duesenberry's ratcheting of consumption: optimal dynamic consumption and investment given intolerance for any decline in standard of living, *Review of Economic Studies* **62**, 287–313.
- [13] Fleming, A. & Zariphopoulou, T. (1991). An optimal Investment-consumption model with borrowing constraints, *Mathematics of Operations Research* **16**, 802–822.
- [14] He, H. & Pagès, H. (1993). Labor income, borrowing constraints, and equilibrium asset prices, *Economic Theory* **3**, 663–696.

- [15] Hindy, A. (1995). Viable prices in financial markets with solvency constraints, *Journal of Mathematical Economics* **24**, 105–136.
- [16] Ingersoll, J. (1987). *Theory of Financial Decision Making*, Rowman and Littlefield, Totowa.
- [17] Ingersoll, J. (1992). Optimal consumption and portfolio rules with intertemporally dependent utility of consumption, *Journal of Economic Dynamics and Control* **16**, 681–712.
- [18] Jarrow, R.A. & Rudd, A. (1983). *Option Pricing*, Irwin, 100–105.
- [19] Karatzas, I., Lehoczky, J., Sethi, S. & Shreve, S. (1986). Explicit solution of a general consumption-investment problem, *Mathematics of Operations Research* **11**, 261–264.
- [20] Lehoczky, J., Sethi, S. & Shreve, S. (1983). Optimal consumption and investment policies allowing consumption constraints and bankruptcy, *Mathematics of Operations Research* **8**, 613–636.
- [21] Merton, R.C. (1969). Lifetime portfolio selection under uncertainty—continuous-time case, *Review of Economics and Statistics* **51**, 247–257.
- [22] Merton, R.C. (1971). Optimum consumption and portfolio rules in a continuous-time model, *Journal of Economic Theory* **3**, 373–413.
- [23] Merton, R.C. (1973). Intertemporal capital asset pricing model, *Econometrica* **41**, 867–887.
- [24] Merton, R.C. (1977). On the pricing of contingent claims and the Modigliani–Miller theorem, *Journal of Financial Economics* **5**, 241–250.
- [25] Merton, R.C. (1990). *Continuous-Time Finance*, Basil Blackwell.
- [26] Richard, S. (1975). Optimal consumption, portfolio, and life insurance rules for an uncertain lived individual in a continuous time model, *Journal of Financial Economics* **2**, 187–203.
- [27] Schweizer, M. (1994). Risk-minimizing hedging strategies under restricted information, *Mathematical Finance* **4**, 327–342.
- [28] Sundaresan, S. (1989). Intertemporally dependent preferences in the theories of consumption, portfolio choices and equilibrium asset pricing, *Review of Financial Studies* **2**, 73–89.
- [29] Vayanos, D. (1998). Transaction costs and asset pricing: a dynamic equilibrium model, *Review of Financial Studies* **11**, 1–58.

YAACOV Z. BERGMAN

Mean–Variance Hedging

In a nutshell, *mean–variance hedging (MVH)* is the problem of approximating, with minimal mean-squared error, a given payoff by the final value of a self-financing trading strategy in a financial market. *Mean–variance portfolio selection (MVPS)*, on the other hand, consists of finding a self-financing strategy whose final value has maximal mean and minimal variance.

More precisely, let $S = (S_t)_{0 \leq t \leq T}$ be an $(\mathbb{R}^d$ -valued) stochastic process on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, P)$ and think of S_t as discounted time t prices of d underlying risky assets. Assume S is a semimartingale and denote by Θ a class of $(\mathbb{R}^d$ -valued) predictable S -integrable processes $\vartheta = (\vartheta_t)_{0 \leq t \leq T}$ satisfying suitable technical conditions. Together with an initial capital x , each ϑ describes, via its time t holdings ϑ_t in S , a self-financing strategy whose *value* at time t is given by the stochastic integral (see **Stochastic Integrals**)

$$V_t(x, \vartheta) = x + \int_0^t \vartheta_u dS_u =: x + G_t(\vartheta) \quad (1)$$

Mean–variance portfolio selection, for some risk aversion parameter $\gamma > 0$, then amounts to

$$\begin{aligned} & \text{maximize } E[V_T(x, \vartheta)] - \gamma \text{Var}[V_T(x, \vartheta)] \\ & \text{over all } \vartheta \in \Theta \end{aligned} \quad (2)$$

and MVH, for a final time T payoff given by a square-integrable $\mathcal{F}_T$ -measurable random variable H , amounts to (see **Hedging**)

$$\begin{aligned} & \text{minimize } E[|V_T(x, \vartheta) - H|^2] \\ & \text{over all } \vartheta \in \Theta \end{aligned} \quad (3)$$

By writing the objective of equation (2) as $m(\vartheta) - \gamma E[|V_T(x, \vartheta) - m(\vartheta)|^2]$ and adding the constraint $m(\vartheta) := E[V_T(x, \vartheta)] = m$, we can solve problem (2) by first solving problem (3) for a constant payoff $H \equiv m$ and then optimizing over m . So we first focus on mean–variance hedging.

Remark 1 A Google Scholar search quickly reveals that the literature on “mean–variance hedging” and “mean–variance portfolio selection” is vast; it cannot be properly surveyed here. Hence we have chosen

references partly for historical interest, partly for novelty, and partly for other subjective reasons. Any omissions may be blamed on this and lack of space.

In mathematical terms, MVH as in equation (3) is simply the problem of finding the best approximation in $L^2 = L^2(P)$ of H by an element of $\mathcal{G} := G_T(\Theta)$. Existence (for arbitrary H) is thus tantamount to closedness of $\mathcal{G}$ in L^2 , which depends on the precise choice of Θ ; results in that direction can be found in [15, 16, 18, 24, 44, 50]. Since the optimal approximand is given by the projection in L^2 of H onto $\mathcal{G}$, MVH (without constraints on strategies) has the pleasant feature that its solution is linear as a function of H . The main challenge, however, is to find more explicit descriptions of the optimal strategy $\hat{\vartheta}^H$, that is, the minimizer for problem (3). The key difficulty there stems from the fact that S is, in general, a P -semimartingale, but not a P -martingale.

Remark 2 If S is a P -martingale, MVH of H is solved by projecting the P -martingale V^H associated with H onto the stable subspace of all stochastic integrals of S , and the optimal strategy is the integrand in the Galtchouk–Kunita–Watanabe decomposition of V^H with respect to S under P . This is also the (first component of the) strategy which is *risk-minimizing* for H in the sense of [22]; see also [11]. However, in this mathematically classical case, MVH is of minor interest for finance since a martingale stock price process has zero excess return.

Historically, mean–variance portfolio selection is much older than mean–variance hedging. It is traditionally credited to Harry Markowitz (1952), although closely related work by Bruno de Finetti (1940) has been discovered recently; [2] provides an overview (see **Markowitz, Harry; Modern Portfolio Theory**). For the static one-period case where $G_T(\vartheta) = \vartheta^{\text{tr}}(S_T - S_0)$ and ϑ is a nonrandom vector, [40, 41] contain a general formulation and [43] an explicit solution (see also **Risk–Return Analysis**). A multi-period treatment, whether in discrete or in continuous time, is considerably more delicate; this was already noticed in [45] and is explained more carefully a bit later.

Mean–variance hedging in the general formulation (3) seems to have been introduced only around 1990. It first appeared in a specific framework in [49], which generalizes a particular example from

[21], and was subsequently extended to very general settings; see [47, 55] for surveys of the literature up to around 2000. Most of these papers use martingale techniques, and an important quantity in that context is the *variance-optimal martingale measure* $\tilde{P}$, obtained as the solution to the dual problem of minimizing over all (signed) local martingale measures Q for S the $L^2(P)$ -norm of the density dQ/dP (see **Equivalent Martingale Measures**). It turns out [53] that if one modifies problem (3) to

$$\begin{aligned} & \text{minimize } E[|V_T(x, \vartheta) - H|^2] \\ & \text{over all } x \in \mathbb{R} \text{ and } \vartheta \in \Theta \end{aligned} \quad (4)$$

the optimal initial capital is given by $\tilde{x} = E_{\tilde{P}}[H]$, and $\tilde{P}$ also plays a key role in finding the optimal strategy $\tilde{\vartheta}^H$. If S is continuous, then $\tilde{P}$ is equivalent to P (see **Equivalence of Probability Measures**) so that its density process $Z^{\tilde{P}}$ is strictly positive [19]. This can then be exploited to give a more explicit description of $\tilde{\vartheta}^H$, either *via* an elegant change of numeraire ([24], see also **Change of Numeraire**), or *via* a change of measure and a recursive formula [48]; see also [1] for an overview of partial extensions to discontinuous settings. For general discontinuous S , [15] have shown that the optimal strategy can be found in the same way as the locally risk-minimizing strategy [55] provided one first makes a change from P to a new (their so-called opportunity-neutral) probability measure P^* .

One common feature of all the above results is that they require for a more explicit description of $\tilde{\vartheta}^H$ the density process ($Z^{\tilde{P}}$ or Z^{P^*}) of some measure, and that this process is very difficult to find in general. Things become much simpler under the (frequently made but restrictive) assumption that S has a *deterministic mean–variance trade-off* (also called *nonstochastic opportunity set*), because $\tilde{P}$ then coincides with the *minimal martingale measure* $\hat{P}$ (see **Minimal Martingale Measure**), which can always be written down directly from the semimartingale decomposition of S ; see [52]. The process S typically has a deterministic mean–variance trade-off if it has independent returns or is a Lévy process (see **Lévy Processes**); this explains why MVH can be used so easily in such settings.

The original MVH problem (3) is a static problem in the sense that one tries at time 0 to find an optimal strategy for the entire interval $[0, T]$. For an

intertemporally dynamic formulation, one would at any time t

$$\begin{aligned} & \text{minimize } E[|V_T(x, \vartheta) - H|^2 | \mathcal{F}_t] \\ & \text{over all } \vartheta \in \Theta_t(\psi) \end{aligned} \quad (5)$$

where $\Theta_t(\psi)$ denotes all strategies $\vartheta \in \Theta$ that agree up to time t with a given $\psi \in \Theta$. In view of equation (1), one recognizes in equation (5) a *linear–quadratic stochastic control (LQSC)* problem, and this point of view allows to exploit additional theory (see **Stochastic Control**) and to obtain, in some situations, more explicit results about the optimal strategy as well. The idea to tackle MVH *via* LQ control techniques and backward stochastic differential equations (BSDEs; see **Backward Stochastic Differential Equations**) seems to originate with M. Kohlmann and X. Y. Zhou. Together with various coauthors, they developed this approach through several papers in an Itô diffusion setting for S ; references [29, 31, 37, 60, 61] provide an overview. A key contribution was made a little earlier in [36] in a discrete-time model by embedding the MVPS problem into a class of auxiliary LQSC problems. Extensions beyond the Brownian setting are given in [10, 38, 39] among others; approaches in discrete time can be found in [14, 25] or [51].

As already stated, MVH is very popular and has been used and studied in many examples and contexts. A few of these are mentioned below:

- stochastic volatility models ([5, 34]; see also **Stochastic Volatility Models**);
- insurance and ALM applications [17, 20, 57];
- weather derivatives or electricity loads ([12, 46]; see also **Weather Derivatives; Commodity Risk**);
- uncertain horizon models [42];
- insider trading [6, 13, 30];
- robustness and model uncertainty ([23, 58]; see also **Robust Portfolio Optimization**);
- default risk and credit derivatives ([4, 8, 28]; see also Section 10, “Credit Derivatives”, of this encyclopedia).

Perhaps the main difference between mean–variance hedging and mean–variance portfolio selection is that MVPS is not consistent over time, in the following sense. If, in analogy to equation (5), we

consider for each t the problem to

$$\begin{aligned} & \text{maximize } E[V_T(x, \vartheta)|\mathcal{F}_t] \\ & - \gamma \text{Var}[V_T(x, \vartheta)|\mathcal{F}_t] \text{ over } \vartheta \end{aligned} \quad (6)$$

this is no longer a standard stochastic control problem because of the variance term. In particular, the crucial *dynamic programming* property fails: if ϑ^* solves problem (2) on $[0, T]$ and we consider problem (6) where we optimize over all $\vartheta \in \Theta_t(\vartheta^*)$, that is, strategies ϑ that agree with ϑ^* up to time t , the solution of this conditional problem over $[t, T]$ will differ from ϑ^* , in general. This makes things surprisingly difficult and explains why MVPS in a general multiperiod setting has still not been solved in a satisfactorily explicit manner.

From the purely geometric structure of the problem, one can derive by elementary arguments the optimal final value

$$G_T(\vartheta^*) = \frac{1}{2\gamma} \left(\alpha - \frac{d\tilde{P}}{dP} \right) \quad (7)$$

with $\alpha = E \left[\left(d\tilde{P}/dP \right)^2 \right]$; this can be seen from [53, 54] or also found in [59]. However, (7) mainly shows that finding the optimal strategy ϑ^* is inextricably linked to a precise knowledge of the variance-optimal martingale measure $\tilde{P}$, which is very difficult to obtain in general. For the case of a *deterministic mean-variance trade-off* (nonstochastic opportunity set), we have already seen that $\tilde{P}$ equals the minimal martingale measure $\hat{P}$ so that equation (7) readily gives the solution to the MVPS problem (2) in explicit form. This includes for instance the results obtained by Li and Ng [36] in finite discrete time or by Zhou and Li [61] who used BSDE techniques in continuous time. Other work in various settings includes [7, 35, 56].

One major area of recent developments in MVPS is the inclusion of constraints (for instance, [9, 26, 27, 32, 33]). Another challenging open problem is to find a time-consistent formulation ([3] provides a first attempt).

References

- [1] Arai, T. (2005). Some remarks on mean-variance hedging for discontinuous asset price processes, *International Journal of Theoretical and Applied Finance* **8**, 425–443.
- [2] Barone, L. (2008). Bruno de Finetti and the case of the critical line's last segment, *Insurance: Mathematics and Economics* **42**, 359–377.
- [3] Basak, S. & Chabakauri, G. (2008). *Dynamic mean-variance asset allocation*, London Business School, available at SSRN: <http://ssrn.com/abstract=965926>, forthcoming in Review of Financial Studies.
- [4] Biagini, F. & Cretarola, A. (2007). Quadratic hedging methods for defaultable claims, *Applied Mathematics and Optimization* **56**, 425–443.
- [5] Biagini, F., Guasoni, P. & Pratelli, M. (2000). Mean-variance hedging for stochastic volatility models, *Mathematical Finance* **10**, 109–123.
- [6] Biagini, F. & Oksendal, B. (2006). Minimal variance hedging for insider trading, *International Journal of Theoretical and Applied Finance* **9**, 1351–1375.
- [7] Bick, A. (2004). The mathematics of the portfolio frontier: a geometry-based approach, *Quarterly Review of Economics and Finance* **44**, 337–361.
- [8] Bielecki, T.R., Jeanblanc, M. & Rutkowski, M. (2004). Hedging of defaultable claims, in *Paris-Princeton Lecture Notes on Mathematical Finance*, Lecture Notes in Mathematics, Springer, Vol. 1847, pp. 1–132.
- [9] Bielecki, T.R., Jin, H., Pliska, S.R. & Zhou, X.Y. (2005). Continuous-time mean-variance portfolio selection with bankruptcy prohibition, *Mathematical Finance* **15**, 213–244.
- [10] Bobrovnytska, O. & Schweizer, M. (2004). Mean-variance hedging and stochastic control: beyond the Brownian setting, *IEEE Transactions on Automatic Control* **49**, 396–408.
- [11] Bouleau, N. & Lamberton, D. (1989). Residual risks and hedging strategies in Markovian markets, *Stochastic Processes and their Applications* **33**, 131–150.
- [12] Brockett, P.L., Wang, M., Yang, C. & Zou, H. (2006). Portfolio effects and valuation of weather derivatives, *Financial Review* **41/1**, 55–76.
- [13] Campi, L. (2005). Some results on quadratic hedging with insider trading, *Stochastics* **77**, 327–348.
- [14] Černý, A. (2004). Dynamic programming and mean-variance hedging in discrete time, *Applied Mathematical Finance* **11**, 1–25.
- [15] Černý, A. & Kallsen, J. (2007). On the structure of general mean-variance hedging strategies, *Annals of Probability* **35**, 1479–1531.
- [16] Choulli, T., Krawczyk, L. & Stricker, C. (1998). $\mathcal{E}$ -martingales and their applications in mathematical finance, *Annals of Probability* **26**, 853–876.
- [17] Dahl, M. & Möller, T. (2006). Valuation and hedging of life insurance liabilities with systematic mortality risk, *Insurance: Mathematics and Economics* **39**, 193–217.
- [18] Delbaen, F., Monat, P., Schachermayer, W., Schweizer, M. & Stricker, C. (1997). Weighted norm inequalities and hedging in incomplete markets, *Finance and Stochastics* **1**, 181–227.
- [19] Delbaen, F. & Schachermayer, W. (1996). The variance-optimal martingale measure for continuous processes,

- Bernoulli* **2**, 81–105. Amendments and corrections (1996). *Bernoulli* **2**, 379–380.
- [20] Delong, L. & Gerrard, R. (2007). Mean-variance portfolio selection for a non-life insurance company, *Mathematical Methods of Operations Research* **66**, 339–367.
- [21] Duffie, D. & Richardson, H.R. (1991). Mean-variance hedging in continuous time, *Annals of Applied Probability* **1**, 1–15.
- [22] Föllmer, H. & Sondermann, D. (1986). Hedging of non-redundant contingent claims, in *Contributions to Mathematical Economics*, W. Hildenbrand & A. Mas-Colell, eds, North-Holland, pp. 205–223.
- [23] Goldfarb, D. & Iyengar, G. (2003). Robust portfolio selection problems, *Mathematics of Operations Research* **28**, 1–38.
- [24] Gouriéroux, C., Laurent, J.P. & Pham, H. (1998). Mean-variance hedging and numéraire, *Mathematical Finance* **8**, 179–200.
- [25] Gugushvili, S. (2003). Dynamic programming and mean-variance hedging in discrete time, *Georgian Mathematical Journal* **10**, 237–246.
- [26] Hu, Y. & Zhou, X.Y. (2006). Constrained stochastic LQ control with random coefficients, and application to portfolio selection, *SIAM Journal on Control and Optimization* **44**, 444–466.
- [27] Jin, H. & Zhou, X.Y. (2007). Continuous-time Markowitz's problems in an incomplete market, with no-shorting portfolios, in *Stochastic Analysis and Applications. Proceedings of the Second Abel Symposium, Oslo, 2005*, F.E. Benth, G. Di Nunno, T. Lindstrøm, B. Øksendal & T. Zhang eds, Springer, pp. 125–151.
- [28] Kohlmann, M. (2007). The mean-variance hedging of a defaultable option with partial information, *Stochastic Analysis and Applications* **25**, 869–893.
- [29] Kohlmann, M. & Tang, S. (2002). Global adapted solution of one-dimensional backward stochastic Riccati equations, with application to the mean-variance hedging, *Stochastic Processes and their Applications* **97**, 255–288.
- [30] Kohlmann, M., Xiong, D. & Ye, Z. (2007). Change of filtrations and mean-variance hedging, *Stochastics* **79**, 539–562.
- [31] Kohlmann, M. & Zhou, X.Y. (2000). Relationship between backward stochastic differential equations and stochastic controls: a linear-quadratic approach, *SIAM Journal on Control and Optimization* **38**, 1392–1407.
- [32] Korn, R. & Trautmann, S. (1995). Continuous-time portfolio optimization under terminal wealth constraints, *Mathematical Methods of Operations Research* **42**, 69–92.
- [33] Labbé, C. & Heunis, A.J. (2007). Convex duality in constrained mean-variance portfolio optimization, *Advances in Applied Probability* **39**, 77–104.
- [34] Laurent, J.P. & Pham, H. (1999). Dynamic programming and mean-variance hedging, *Finance and Stochastics* **3**, 83–110.
- [35] Leippold, M., Trojani, F. & Vanini, P. (2004). A geometric approach to multiperiod mean variance optimization of assets and liabilities, *Journal of Economic Dynamics and Control* **28**, 1079–1113.
- [36] Li, D. & Ng, W.-L. (2000). Optimal dynamic portfolio selection: Multi-period mean-variance formulation, *Mathematical Finance* **10**, 387–406.
- [37] Lim, A.E.B. (2004). Quadratic hedging and mean-variance portfolio selection with random parameters in an incomplete market, *Mathematics of Operations Research* **29**, 132–161.
- [38] Lim, A.E.B. (2006). Mean-variance hedging when there are jumps, *SIAM Journal on Control and Optimization* **44**, 1893–1922.
- [39] Mania, M. & Tevzadze, R. (2003). Backward stochastic PDE and imperfect hedging, *International Journal of Theoretical and Applied Finance* **6**, 663–692.
- [40] Markowitz, H.M. (1952). Portfolio selection, *Journal of Finance* **7**, 77–91.
- [41] Markowitz, H.M., Lacey, R., Plymen, J., Dempster, M.A.H. & Tompkins, R.G. (1994). The general mean-variance portfolio selection problem [and discussion], *Philosophical Transactions: Physical Sciences and Engineering, Mathematical Models in Finance* **347**(1684), 543–549.
- [42] Martellini, L. & Urošević, B. (2006). Static mean-variance analysis with uncertain time horizon, *Management Science* **52**, 955–964.
- [43] Merton, R.C. (1972). An analytic derivation of the efficient portfolio frontier, *Journal of Financial and Quantitative Analysis* **7**, 1851–1872.
- [44] Monat, P. & Stricker, C. (1995). Föllmer-Schweizer decomposition and mean-variance hedging of general claims, *Annals of Probability* **23**, 605–628.
- [45] Mossin, J. (1968). Optimal multiperiod portfolio policies, *Journal of Business* **41**, 215.
- [46] Näsäkkälä, E. & Keppo, J. (2005). Electricity load pattern hedging with static forward strategies, *Managerial Finance* **31/6**, 116–137.
- [47] Pham, H. (2000). On quadratic hedging in continuous time, *Mathematical Methods of Operations Research* **51**, 315–339.
- [48] Rheinländer, T. & Schweizer, M. (1997). On L^2 -projections on a space of stochastic integrals, *Annals of Probability* **25**, 1810–1831.
- [49] Schweizer, M. (1992). Mean-variance hedging for general claims, *Annals of Applied Probability* **2**, 171–179.
- [50] Schweizer, M. (1994). Approximating random variables by stochastic integrals, *Annals of Probability* **22**, 1536–1575.
- [51] Schweizer, M. (1995). Variance-optimal hedging in discrete time, *Mathematics of Operations Research* **20**, 1–32.
- [52] Schweizer, M. (1995). On the minimal martingale measure and the Föllmer-Schweizer decomposition, *Stochastic Analysis and Applications* **13**, 573–599.

-
- [53] Schweizer, M. (1996). Approximation pricing and the variance-optimal martingale measure, *Annals of Probability* **24**, 206–236.
 - [54] Schweizer, M. (2001). From actuarial to financial valuation principles, *Insurance: Mathematics and Economics* **28**, 31–47.
 - [55] Schweizer, M. (2001). A guided tour through quadratic hedging approaches, in *Option Pricing, Interest Rates and Risk Management*, E. Jouini, J. Cvitanić & M. Musiela, eds, Cambridge University Press, pp. 538–574.
 - [56] Steinbach, M.C. (2001). Markowitz revisited: mean-variance models in financial portfolio analysis, *SIAM Review* **43**, 31–85.
 - [57] Thomson, R.J. (2005). The pricing of liabilities in an incomplete market using dynamic mean-variance hedging, *Insurance: Mathematics and Economics* **36**, 441–455.
 - [58] Toronjadze, T. (2001). Optimal mean-variance robust hedging under asset price model misspecification, *Georgian Mathematical Journal* **8**, 189–199.
 - [59] Xia, J. & Yan, J.A. (2006). Markowitz's portfolio optimization in an incomplete market, *Mathematical Finance* **16**, 203–216.
 - [60] Zhou, X.Y. (2003). Markowitz's world in continuous time, and beyond, in *Stochastic Modeling and Optimization: With Applications in Queues, Finance, and Supply Chains*, D.D. Yao, H. Zhang & D.Y. Zhou, eds, Springer, pp. 279–309.
 - [61] Zhou, X.Y. & Li, D. (2000). Continuous-time mean-variance portfolio selection: a stochastic LQ framework, *Applied Mathematics and Optimization* **42**, 19–33.

Related Articles

Expected Utility Maximization; Hedging; Minimal Martingale Measure; Risk-Return Analysis; Stochastic Integrals.

MARTIN SCHWEIZER

Risk Exposures

Broadly speaking, risk can be defined as the volatility of returns leading to “unexpected losses” (see **Value-at-Risk**), with higher volatility indicating higher risk.^a The volatility of returns is directly or indirectly influenced by numerous variables, which we called *risk factors*, and by the interaction between these risk factors.

Risk factors can be broadly grouped together into the following categories: market risk, credit risk, liquidity risk, operational risk, legal and regulatory risk, business risk, strategic risk, and reputation risk [2]. These categories can then be further decomposed into more specific categories, as we show in Figure 1 for market risk and credit risk.

In this figure, we have subdivided market risk into equity risk, interest rate risk, currency risk, and commodity price risk in a manner that is in line with our detailed discussion below. Then we have divided interest rate risk into trading risk and the special case of gap risk: the latter relates to the risk that arises in the balance sheet of an institution from the different sensitivities of assets and liabilities to changes in interest rates.

In theory, the more all-encompassing the categorization and the more detailed the decomposition, the more closely the company’s risk will be captured. In practice, this process is limited by the level of model complexity that can be handled by the available technology and by the cost and availability of internal and market data.

Market Risk

Market risk is the risk that changes in financial market prices and rates will reduce the dollar value of a security or a portfolio. Price risk for fixed-income products can be decomposed into a “general market” risk component—the risk that the market as a whole will fall in value—and a “specific” risk component, unique to the particular financial transaction under consideration, that also reflects the credit risk hidden in the instrument. In trading activities, risk arises both from open (unhedged) positions and from imperfect correlations between market positions that are intended to offset one another.

Market risk is given many different names in different contexts. For example, in the case of a fund, the fund may be marketed as tracking the performance of a certain benchmark. In this case, market risk is important to the extent that it creates a “risk of tracking error”. *Basis risk* is a term used in the risk management industry to describe the chance of a breakdown in the relationship between the price of a product, on the one hand, and the price of the instrument used to hedge that price exposure, on the other hand. Again, it is really just a context-specific form of market risk.

There are four major types of market risk:

1. Interest rate risk

The simplest form of interest rate risk is the risk that the value of a fixed-income security will fall as a result of an increase in market interest rates. However, in complex portfolios of interest-rate-sensitive assets, many different kinds of exposure can arise from differences in the maturities, nominal values, and reset dates of instruments and cashflows that are asset like (i.e., “longs”) and those that are liability like (i.e., “shorts”).

In particular, “curve” risk can arise in portfolios in which long and short positions of different maturities are effectively hedged against a *parallel shift* in yields but not against a change in the *shape of the yield curve*. Meanwhile, even where offsetting positions have the same maturity, basis risk can arise if the rates of the positions are imperfectly correlated. For example, three-month Eurodollar instruments and three-month Treasury bills both naturally pay three-month interest rates. However, these rates are not perfectly correlated with each other, and spreads between their yields may vary over time. As a result, a three-month Treasury bill funded by three-month Eurodollar deposits represents an imperfect offset or hedged position.

2. Equity price risk

This is the risk associated with volatility in stock prices. The general market risk of equity refers to the sensitivity of an instrument or portfolio value to a change in the level of broad stock market indices. The “specific” or “idiosyncratic” risk of equity refers to that portion of a stock’s price volatility determined by characteristics specific to the firm, such as its line of business, the quality of its management, or a breakdown in its production process. According

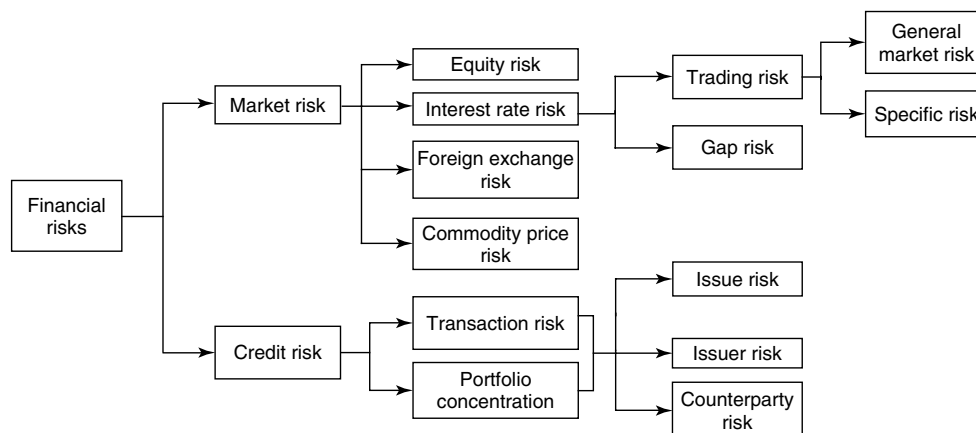


Figure 1 Schematic presentation, by categories, of financial risks

to portfolio theory, general market risk cannot be eliminated through portfolio diversification, whereas specific risk can be diversified away.

3. Foreign exchange risk

Foreign exchange risk arises from open or imperfectly hedged positions in a particular currency. These positions may arise as a natural consequence of business operations, rather than from any conscious desire to take a trading position in a currency. Foreign exchange volatility can sweep away the return from expensive cross-border investments, and at the same time place a firm at a competitive disadvantage in relation to its foreign competitors. It may also generate huge operating losses and, through the uncertainty it causes, inhibit investment. The major drivers of foreign exchange risk are imperfect correlations in the movement of currency prices and fluctuations in international interest rates. Although it is important to acknowledge exchange rates as a distinct market risk factor, the valuation of foreign exchange transactions requires knowledge of the behavior of domestic and foreign interest rates, as well as of spot exchange rates.

4. Commodity price risk

The price risk of commodities differs considerably from interest rate and foreign exchange risk, since most commodities are traded in markets in which the concentration of supply in the hands of a few suppliers can magnify price volatility. Fluctuations in the depth of trading in the market (i.e., market

liquidity) often accompany and exacerbate high levels of price volatility. Other fundamentals affecting a commodity price include the ease and cost of storage, which varies considerably across the commodity markets (e.g., from gold, to electricity, to wheat). As a result of these factors, commodity prices generally have higher volatilities and larger price discontinuities—that is, moments when prices leap from one level to another—than most traded financial securities. Commodities can be classified according to their characteristics as follows: hard commodities are non-perishable commodities, the markets for which are further divided into precious metals (e.g., gold, silver, and platinum), which have a high price/weight value and base metals (e.g., copper, zinc, and tin); soft commodities or commodities with a short shelf life and that are hard to store, mainly agricultural products (e.g., grains, coffee, and sugar); and energy commodities that consist of oil, gas, electricity, and other energy products.

Credit Risk

Credit risk is the risk that a change in the credit quality of a counterparty will affect the value of a security or a portfolio. Default, whereby a counterparty is unwilling or unable to fulfill its contractual obligations, is the extreme case; however, institutions are also exposed to the risk that a counterparty might be downgraded by a rating agency.

Credit risk is only an issue when the position is an asset, that is, when it exhibits a positive replacement

value. In that situation, if the counterparty defaults, the firm loses either all of the market value of the position or, more commonly, the part of the value that it cannot recover following the credit event. (The value it is likely to recover is called the *recovery value* or *recovery rate* when expressed as a percentage; the amount it is expected to lose is called the *loss given default*.)

Unlike the potential loss given default on coupon bonds or loans, the one on derivative positions is usually much lower than the nominal amount of the deal, and in many cases is only a fraction of this amount. This is because the economic value of a derivative instrument is related to its replacement or market value rather than its nominal or face value. However, the credit exposures induced by the replacement values of derivative instruments are dynamic: they can be negative at one point in time, and yet become positive at a later point in time after market conditions have changed. Therefore, firms must examine not only the current exposure measured by the current replacement value but also the profile of potential future exposures up to the termination of the deal.

Liquidity Risk

Liquidity risk comprises both “funding liquidity risk” and “asset liquidity risk”, though these two dimensions of liquidity risk are closely related. Funding liquidity risk relates to a firm’s ability to raise the necessary cash to roll over its debt, to meet the cash, margin and collateral requirements of counterparties, and (in the case of funds) to satisfy capital withdrawals. Funding liquidity risk can be managed through holding cash and cash equivalents, setting credit lines in place, and through monitoring “buying power”. (Buying powers refers to the amount a trading counterparty can borrow against assets under stressed market conditions.)

Asset liquidity risk, often simply called *liquidity risk*, is the risk that an institution will not be able to execute a transaction at the prevailing market price because there is, temporarily, no appetite for the deal on the “other side” of the market. If the transaction cannot be postponed, its execution may lead to substantial loss on the position. This risk is generally very hard to quantify.

Operational Risk

Operational risk (*see* **Operational Risk**) refers to *potential losses* resulting from inadequate systems, management failure, faulty controls, fraud, and human error.^b Many of the large losses from derivative trading over the last decade are the direct consequence of operational failures. Derivative trading is more prone to operational risk than cash transactions because derivatives are, by their nature, leveraged transactions. The valuation of complex derivatives also creates considerable operational risk. Very tight controls are an absolute necessity if a firm is to avoid large losses.

Operational risk includes “fraud”, for example, when a trader or other employee intentionally falsifies and misrepresents the risks incurred in a transaction. Technology risk, and principally computer systems risk, also falls into the operational risk category.

Human factor risk is a special form of operational risk. It relates to the losses that may result from human errors such as pushing the wrong button on a computer, inadvertently destroying a file, or entering the wrong value for the parameter input of a model.

Legal and Regulatory Risk

Legal and regulatory risk arises for a whole variety of reasons and is closely related to reputation risk (see below). For example, a counterparty might lack the legal or regulatory authority to engage in a risky transaction. In the derivative markets, legal risks often only become apparent when a counterparty, or an investor, loses money on a transaction and decides to sue the provider firm to avoid meeting its obligations. Another aspect of regulatory risk is the potential impact of a change in tax law on the market value of a position. For example, when the British government changed the tax code to remove a particular tax benefit during the summer of 1997, one major investment bank suffered huge losses.

Business Risk

Business risk refers to the classic risks of the world of business such as uncertainty about the demand for products, or the price that can be charged for those products, or the cost of producing and delivering products.

In the world of manufacturing, business risk is largely managed through core tasks of management, for example, choices about channel, products, suppliers, how products are marketed, and so on.

However, there remains the question of how business risk should be addressed within formal risk management frameworks and that have become prevalent in the financial industries.

Although business risks should surely be assessed and monitored, it is not obvious how to do this in a way that complements the banking industry's treatment of classic credit and market risks. There is also room for debate over whether business risks need to be supported by capital in the same explicit way (*see* **Economic Capital**). In the new Basel Capital Accord (*see* **Regulatory Capital**), "business risk" is excluded from the regulators' definition of operational risk, even though some researchers believe it to be a greater source of volatility in bank revenue than the operational event/failure risk that the regulators *have* included within bank minimum capital requirements.

Business risk is affected by such factors as the quality of the strategy and/ or the reputation of the firm as well as other factors. Therefore, it is common practice to view strategic and reputation risks as components of business risk and the risk literature sometimes refers to a complex of business/strategic/reputation risk. In this typology, we differentiate these three components.

Strategic Risk

Strategic risk refers to the risk of significant investments for which there is a high uncertainty about success and profitability. If the venture is not successful, then the firm will usually suffer a major write-off and its reputation among investors will be damaged.

Reputation Risk

Reputation risk is taking on a new dimension after the accounting scandals that defrauded the shareholders, bondholders, and employees of many major corporations during the boom in the equity markets in the late 1990s. Investigations into the mutual funds and insurance industry by New York Attorney General Eliot Spitzer have also made clear just how important a reputation for fair dealing is with both customers and

regulators. In a survey released in August 2004 by PricewaterhouseCoopers (PwC) and the Economist Intelligence Unit (EIU), 34% of the 134 international bank respondents believed that reputation risk is the biggest risk to market and shareholder value faced by banks, while market and credit risk scored only 25% each.

No doubt this was partly because corporate scandals like Enron, Worldcom, and others were still fresh in bankers' minds. Some experts, however, believe that reputation is a genuine emerging issue, and that the new Basel Capital Accord will help to shift the attention of regulators and investors from quantifiable risks like market and credit risk toward strategic and business risk.

Reputation risk poses a special threat to financial institutions because the nature of their business requires the confidence of customers, creditors, regulators, and the general market place. The development of a wide array of structured finance products, including financial derivatives for market and credit risk, asset-backed securities with customized cash flows, and specialized financial conduits that manage pools of purchased assets, has placed pressure on the interpretation of accounting and tax rules, and, in turn, has given rise to significant concerns about the legality and appropriateness of certain transactions. Involvement in such transactions may damage an institution's reputation and franchise value.

Financial institutions are also under increasing pressure to demonstrate their ethical, social, and environmental responsibility. As a defensive mechanism, 10 international banks from seven countries announced in June 2003 the adoption of the "Equator Principles", a voluntary set of guidelines developed by the banks for managing social and environmental issues related to the financing of projects in emerging countries. The Equator Principles are based on the policy and guidelines of the World Bank and International Finance Corporation (IFC) and require the borrower to conduct an environmental assessment for high-risk projects to address issues such as sustainable development and use of renewable natural resources, protection of human health, pollution prevention and waste minimization, socioeconomic impact, and so on.

Banks should clearly monitor and manage risks to their reputation. However, as for business risk, there is no clear industrial view on whether these risks can be meaningfully *measured* or whether they should

be supported by risk capital. In the case of reputation risk, the problem is compounded: a bank that reserves money against the danger of reputation damage may be tacitly admitting in advance that its conduct is reprehensible.

End Notes

^aFor a more detailed discussion of financial and nonfinancial risks, see Crouhy *et al.* [3]

^bThe Basel Committee on Banking Supervision [1] has identified the following event types as having the potential to result in substantial operational risk losses: internal fraud, for example, intentional misreporting of positions, employee theft, and insider trading on an employee's own account; external fraud, for example, robbery, forgery, check kiting and damage from computer hacking; employment practices, and workplace safety, for example, workers compensation claims, violation of employee health and safety rules, organized labor activities, discrimination claims and general liabilities; clients, products and business practices, for example, fiduciary breaches, misuse of confidential customer information, improper trading activities on the bank's account, money laundering, and sales of unauthorized products; damage to physical assets, for example, terrorism, vandalism, earthquakes, fires, and floods; business disruption and system failures, for example, hardware

and software failures, telecommunication problems, and utility outages; and execution, delivery and process management, for example, data entry errors, collateral management failures, incomplete legal documentation, unapproved access given to client accounts, nonclient counterparty misperformance, and vendor disputes (*see Regulatory Capital*).

References

- [1] Basel Committee on Banking Supervision. (2003). *Sound Practices for the Management and Supervision of Operational Risk*.
- [2] Board of Governors of the Federal Reserve System. (2007). *Trading and Capital-markets Activities Manual*, Washington, DC.
- [3] Crouhy, M., Galai, D. & Mark, R. (2006). *The Essentials of Risk Management*, McGraw-Hill.

Related Articles

Commodity Risk; Counterparty Credit Risk; Credit Risk; Risk Management: Historical Perspectives; Liquidity.

MICHEL CROUHY, DAN GALAI &
ROBERT MARK

Risk Management: Historical Perspectives

Financial risk management refers to the design and implementation of procedures for identifying, measuring, and managing financial risks. At the heart of any activity in financial markets lies a trade-off between expected profit and risk. Naturally, deciding on a course of action should be based on the best possible evaluation of risk. After all, ignoring risk is a sure recipe for financial disaster.

Financial risk management evolved as a distinct discipline in the early 1990s. This was made possible by advances in risk methodologies that, coupled with vast increases in computer processing capabilities, could be applied at the top level of financial institutions. This led to a new type of function, the financial risk manager, which is distinct from the usual business lines. By now, most financial institutions have a “Chief Risk Officer” reporting directly to the top management.

The 1970s and Earlier: Foundations of Risk Management

Modern risk management builds on a century of advances in analytical risk-management tools, as described in Table 1. To simplify, start with a position in a single financial instrument. Conceptually, risk measurement can be separated into two drivers. The first is exposure or change in the value of the position in response to changes in the risk factor(s). The second is the range of possible movements in the risk factors, which can be described by a probability distribution function. Risk measurement brings together these two drivers by describing the range of possible losses on the position.

As an example, consider the most basic financial instrument: a bond, which represents a promise of future cash flows. The value of the bond depends on the yield to maturity, which is the main risk factor. The linear part of this exposure is called *duration*. Duration is a measure of the sensitivity of bond prices to small movements in yields (*see Bond*). This concept was first developed by Macaulay [7] in 1938 and has become the cornerstone of fixed-income portfolio management. Similar concepts were

developed later for other instruments. For example, “delta” measures the linear exposure of an option relative to the underlying risk factor. For a long time, risk was evaluated by instrument or for separate business lines.

In 1952, Markowitz [8] developed the mean–variance framework (*see Risk–Return Analysis*), which emphasizes the importance of measuring risk in a total portfolio context. This view was the foundation for modern risk management and led to the award of the 1990 Nobel Prize in Economics.

By its nature, centralized risk measurement is a large-scale aggregation problem, which involves a large number of parameters. Consider, for example, the problem of allocating assets among the thousands of stocks available for investment. Generally, the joint movements in the stock prices can be described by a covariance matrix that lists all pairwise correlations. The problem, however, is that the number of coefficients in this matrix rises with the square of the number of stocks. Too many coefficients create several problems, ranging from numerical instability in the covariance matrix to loss of intuitive understanding of the economic drivers of risk. As a result, simplifications are useful.

In 1963, Sharpe [10] simplified the covariance structure of stocks with a one-factor model (*see Capital Asset Pricing Model*). This reduced the dimensionality of the covariance matrix to a number of coefficients that increased linearly with the number of stocks. Such simplification is at the heart of mapping methods in modern risk measurement. In addition, this decomposition leads to a neat decomposition of risk into a systematic factor and an idiosyncratic effect. The first component is common to all the stocks and cannot be diversified away. In contrast, idiosyncratic effects become less important in a large portfolio because they tend to cancel out each other. Sharpe then showed that the only risk that should be priced in equilibrium is systematic risk. This theory of asset pricing, known as the *capital asset pricing model (CAPM)*, led to the award of the 1990 Nobel Prize in Economics. In 1966, following Sharpe’s single-factor model, multiple-factor models were developed. These provide a more realistic description of common movements among risk factors that it still parsimonious.

In parallel, the financial industry was creating new financial instruments to manage, hedge, or take, financial risks. Derivatives are private contracts

2 Risk Management: Historical Perspectives

Table 1 The evolution of analytical risk-management tools

1938	Macaulay's bond duration
1952	Markowitz mean–variance framework
1963	Sharpe's single-factor beta model
1966	Multiple factor models
1973	Black–Scholes option pricing model and “Greeks”
1982	ARCH models
1993	Value at Risk (VAR)
1994	RiskMetrics
1997	CreditMetrics
2000	Enterprisewide risk management

whose value derives from some underlying price. Early instruments such as forwards or futures had linear exposures to risk factors. For instance, a farmer could hedge the price risk of a wheat harvest by shorting wheat futures. Because a forward or futures contract implies an obligation to buy or sell, the payoff is a linear function of movements in the underlying risk factor, which is the price of wheat in this case. A drop in this price would cause a loss in the value of the harvest, but this should be largely hedged by a gain on the futures position. Commodity contracts exist since 1874; foreign currency futures were introduced in 1972 and interest rate futures in 1975. Over time, the industry started to offer more flexible contracts such as options, which involve the right, as opposed to the obligation, to buy or sell. This right creates a payoff function that is nonlinear in the underlying asset. Such instruments are more complicated to price and hedge.

In 1973, Black and Scholes [3] developed a breakthrough option pricing model (*see* **Black–Scholes Formula**), which led to the award of the 1997 Nobel Prize in Economics, along with Merton. The model provides a particularly elegant formula for the pricing of options, based on a remarkably small set of inputs. It also generates measures of exposure, or sensitivity, which are called the *Greeks*. The timing of this discovery was fortuitous, as equity options started trading in 1973.

The 1980s: Start of Expansion of Derivatives Markets

The need for efficient risk-management tools led to rapid growth in the market for derivatives. These markets can be described in terms of their notional

Table 2 Global markets for derivatives: outstanding contracts (\$ billion)

Instruments, by location and type	1986	2007
Exchange-traded instruments	583	80 624
Interest rate	516	71 095
Currency	18	9237
Stock index	49	292
OTC instruments	500	596 004
Interest rate swaps	400	393 138
Currency swaps	100	56 238
Credit default swaps	0	57 894
Equity, commodity	0	17 509
Total	1083	676 628

Source: ISDA and Bank for International Settlements

outstanding amounts, because the face value represents the amount of exposure exchanged. The year 1986 was the first one for which we started having reliable market statistics. Before that, markets were smaller or even nonexistent. Table 2 shows that the derivatives markets have grown from \$1 trillion to \$677 trillion in 21 years from 1986 to 2007.

The 1980s witnessed the introduction of major new instruments: currency swaps (1981), interest rate swaps (1982), Eurodollar futures (1981), equity index futures (1982), and swaptions (1985), just to name a few.

As a result of the expansion of these markets, risk-management methods started to take hold in the financial industry. As exposures grew, financial institutions started to impose position limits for fixed-income exposures by duration buckets. Later, positions were imposed on options using the “Greeks”. Banks also began to evaluate performance in terms of risk-adjusted profits. Economic capital, which is essentially a Value at Risk (VAR) measure, was assigned to business lines (*see* **Economic Capital**). This led to risk-adjusted return on capital (RAROC) measures, which deflate expected profits by economic capital, thus providing a consistent view of risk-adjusted profitability across businesses (*see* **Risk-adjusted Return on Capital (RAROC)**).

The expansion of these markets was followed by a number of spectacular losses in the early 1990s, however. Headline cases such as Orange County (1994), Metallgesellschaft (1994), and Barings (1995) threatened to create a backlash against derivatives whose risks were apparently not fully appreciated by

their users. In response, the industry devised more comprehensive measures of risk.

The 1990s: Modern Risk Management

Prodded by regulators, the Group of Thirty [5] endorsed VAR in 1993 as part of “best practices” for dealing with derivatives. VAR had been developed at Morgan, in response to the request of its chairman, “At the close of each business day, tell me what the market risks are across all businesses and locations.” VAR summarizes the worst loss over a target horizon that will not be exceeded with a given level of confidence. For instance, at that time, Morgan’s daily VAR was approximately \$15 million at the 95% level of confidence. In other words, the bank should not expect to lose more than \$15 million in 95 days out of 100. This single number describes the risk of the entire trading portfolio of the bank and therefore involves a large-scale aggregation of positions (*see* Jorion [6] and Figure 1 in **Market Risk**).

Morgan [9] then unveiled its “RiskMetrics” system in October 1994. RiskMetrics, available free on the Internet, provided a datafeed for computing VAR, as well as a detailed technical manual. The widespread availability of data immediately engaged the industry and spurred academic and practitioner research into risk management. The methodology was partly based on new models of time-variation in risk, such as the autoregressive conditional heteroskedasticity (ARCH) model, developed by Engle [4] in 1982 and for which he received the 2003 Nobel Prize in Economics.

The basic idea behind VAR is not new, however. It can be traced back to the mean–variance framework developed by Markowitz in 1952. What is new is the integration of all market risks into a centralized common “metric”, combined with the use of current positions. Consider again the two conceptual drivers of risk, exposure, and range of possible movements in the risk factors. This information should be combined into a distribution of dollar gains and losses, which can be summarized by one number, the quantile at some confidence level, or VAR. An alternative measure of risk is the conditional VAR (CVAR), which is the average of losses beyond VAR (*see Expected Shortfall*). The same analysis can be performed for a portfolio of financial instruments exposed to multiple sources of risk.

The new aspect of modern risk management is that risk measures are now position based. Traditionally, risk was measured from historical return information. Nowadays, however, proprietary trading portfolios turn over very quickly. As a result, returns-based risk measures give incomplete information. This is because today’s risk profile may be very different from the average risk profile over a recent history.

Position-based risk measures, however, are rather complex to build. They involve large-scale structural bottom-up models that aggregate risks from individual position data. The choice of the model reflects a trade-off between speed and accuracy. A portfolio may contain millions of positions, which cannot be modeled individually. Instead, the manager has to choose a set of risk factors that spans the risk space effectively. The risk measurement system then has to “map” all the positions on these risk factors, replacing the value of all positions by exposures. Next, these exposures are aggregated across the entire portfolio. The last step then combines the exposures with the statistical distribution of risk factors, which gives a distribution of profits and losses at the top level.

This modern risk architecture has many more uses than for reporting a single VAR number, however. Once this system is in place, the risk manager can easily examine the effect of extreme or unusual situations. In fact, stress testing is an important complement to VAR because VAR only provides an estimate of losses under normal market conditions, that is, at a prespecified confidence level. Once a risk-management system is in place, however, stress scenarios are just hypothetical realizations of the risk factors (*see Stress Testing*). More generally, risk manager should be keenly aware of weaknesses in their risk models.

By now, VAR is the most widely used measure of market risk. Its use has been widely sanctioned by regulators. In 1996, for example, the Basel Committee on Banking Supervision [1] passed a new rule allowing commercial banks to compute their capital charge for market risk based on their own internal VAR measures. This methodology has spread to most financial institutions. Risk systems are also increasingly used to control risk. Position limits are now routinely based on VAR numbers and stress-test results.

The 2000s: Extensions to Other Risks

VAR methods were initially applied to market risk, which is the risk of losses due to movements in the level or volatility of market prices. Later, modern risk-management techniques were extended to credit risk and operational risk. This led to another round of fundamental developments.

Credit risk is the risk of losses due to the fact that counterparties may be unwilling or unable to fulfill their contractual obligations. This is a major source of risk for the economy and has proved much more damaging than episodic losses due to market risk. Time and again, countries have experienced crises in their banking sector due to credit losses that required expensive recapitalizations.

Toward the end of the 1990s, several portfolio credit risk models were developed. Examples are KMV and CreditMetrics. These models integrate the various components of credit risk, which include probability of default, exposures, and loss given default, as well as all their correlations, in large-scale models.

The banking industry dutifully started to measure their credit risk and soon realized that they needed new financial instruments to manage their credit exposures. This led to the rapid expansion of credit default swaps (*see Credit Default Swaps*), which were created in 1994. These instruments created a new market for the exchange of pure credit risk. The market expanded quickly because of its low transaction costs and also because it allowed short positions in credit, which was not feasible before. Buying a credit default swap creates a profit if the credit event happens. This is akin to shorting a corporate bond but much more practical. By 2007, the outstanding notional amount for credit default swaps was close to \$60 trillion, several times the size of the corporate bond markets.

Credit risk models led to fundamental changes in the banking industry. The industry soon realized that holding loans on their books immobilized large amounts of capital and was not very profitable. In response, banks moved to a model where they could reap fees from the origination of loans, which were then repackaged and distributed to other investors. This led to the creation of the structured credit industry, where pools of debt are put together and sold as securities with various levels of priority on the collateral.

Structured credit had its roots in the collateralized mortgage obligation (CMO) market developed in the early 1980s. Mortgage loans were placed in pools and sold to investors in the form of tranches with different priority claims. The same principle applies to collateralized debt obligations (CDOs, *see Collateralized Debt Obligations (CDO)*). By 2001, the methodology of credit risk portfolio measurement was applied to CDOs, allowing rapid pricing of the various tranches. This spurred an exponential growth of this market, until 2007.

This growth revealed flaws, however. On the one hand, these new instruments dispersed risk to investors and financial institutions across the globe, which is a positive development. In the case of the subprime crisis of 2007, the subprime losses were considerable, amounting to several hundred billion dollars. In the past, such losses could have bankrupted the US financial system. This has not happened, fortunately. On the other hand, this securitization process led to laxer credit standards that aggravated the scale of losses. Lenders were more interested in generating fee income than about the ultimate creditworthiness of borrowers because losses would be borne by somebody else, after all.

In addition to credit risk, risk measurement techniques are now being applied to operational risk, which is the risk of losses resulting from inadequate or failed internal processes, people, and systems or from external events (*see Operational Risk*). Operational risk has caused large losses to financial institutions. In 1995, a single rogue trader caused a loss of \$1.3 billion at Barings, which led to the failure of the bank. As a result, the Basel Committee on Banking Supervision [2] mandated a new capital charge against operational risk, which can be based on a VAR measure. Admittedly, however, operational risk measures are much less precise than those covering market and credit risk. Data on operational risk losses are scarcer than for other types of risk. In addition, losses may not be applicable to banks with different control environments.

In the 2000s, many institutions started to develop an integrated approach to enterprisewide risk management (ERM). This is an extension of the VAR concept, whose essence is centralization, to firmwide risks. ERM consists of identification, assessment, and measurement of all relevant risks, including market, credit, operational, and business risk. In the past, risks were considered separately from each other and

hedged one at a time. In contrast, an integrated view accounts for diversification effects and allows more efficient use of capital. It can also help the firm manage the trade-off between risk and expected profit across and within business lines.

Conclusions

It is fair to conclude by stating that modern risk-management techniques have truly revolutionized the financial industry. Institutions are now busily trying to measure and manage their aggregate risks.

The elegance of these risk models should not create a sense of false security, however. VAR risk measures are not designed to provide worst-loss estimates. We should expect losses to exceed VAR numbers with some regularity. In addition, these risk models can be sensitive to the underlying assumptions. This is why stress tests should be used on a regular basis, including scenario analyses and model tests. A good risk manager should understand the weak spots of the models, combined with an economic intuition of the fundamental driver of the risk factors. The subprime crisis that started in 2007, for example, illustrated severe shortcomings in risk models, as institutions suffered losses that were much worse than anticipated, both in terms of frequency and size. A widespread reaction was for these institutions to strengthen their risk-management structures.

More fundamentally, quantitative risk measurement tools do not apply well to some important sources of risk, such as liquidity risk. This is the risk of loss due to the inability to meet payments obligations, which may force early liquidation of assets at fire-sale prices. Creditors may refuse to roll over their funding, creating “bank runs” such as those that befell Northern Rock and Bear Stearns. At an even broader level, no model can fully account for systemic risk, which arises when by default one institution has a cascading effect on other firms, thus posing a threat to

the viability of the entire financial system. Systemic risk should be handled by the central bank, which is effectively becoming the risk manager of last resort.

References

- [1] Basel Committee on Banking Supervision. (1996). *Amendment to the Basel Capital Accord to Incorporate Market Risk*, BIS, Basel.
- [2] Basel Committee on Banking Supervision. (2005). *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*, BIS, Basel.
- [3] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- [4] Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation, *Econometrica* **50**, 987–1007.
- [5] Group of Thirty. (1993). *Derivatives: Practices and Principles*, Group of Thirty, New York.
- [6] Jorion, P. (2005). *Value at Risk: The New Benchmark to Manage Financial Risk*, McGraw-Hill, New York.
- [7] Macaulay, F. (1938). *The Movements of Interest Rates. Bond Yields and Stock Prices in the United States since 1856*, National Bureau of Economic Research, New York.
- [8] Markowitz, H. (1952). Portfolio selection, *Journal of Finance* **7**, 77–91.
- [9] Morgan, J.P. (1995). *RiskMetrics Technical Manual*, J.P. Morgan, New York.
- [10] Sharpe, W. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk, *Journal of Finance* **19**, 425–442.

Related Articles

Bond; Capital Asset Pricing Model; Collateralized Debt Obligations (CDO); Credit Default Swaps; Markowitz, Harry; Operational Risk; Regulatory Capital; Securitization; Sharpe, William F.; Sharpe Ratio; Spectral Measures of Risk; Value-at-Risk.

PHILIPPE JORION

Convex Risk Measures

Quantifying the risk of the uncertainty in the future value of a portfolio is one of the key tasks of risk management. This quantification is usually achieved by modeling the uncertain payoff as a random variable, to which then a certain functional is applied. Such functionals are usually called *risk measures*. The corresponding industry standard, Value at Risk, is often criticized for encouraging the accumulation of shortfall risk in particular scenarios. This deficiency has led to a search for more appropriate alternatives. The first step of this search consists in specifying certain desirable axioms for risk measures. In the second step, one then tries to characterize those risk measures that satisfy these axioms and to identify suitable examples. In the section Monetary, Convex, and Coherent Risk Measures, we first provide the various axiom sets for *monetary*, *convex*, and *coherent* risk measures. In the section Acceptance Sets, the representation of monetary risk measures in terms of their *acceptance sets* is briefly discussed. The general *dual representation* for convex and coherent risk measures is given in the section Dual Representation. Various examples are provided in the section Examples and Applications. In many situations, it is reasonable to assume that a risk measure depends on the randomness of the portfolio value only through its probability law. Such risk measures are usually called *law invariant*. They are discussed in the section Law-invariant Risk Measures. Finally, section Conditional Convex Risk Measures and Time Consistency analyzes various notions of *dynamic consistency* that naturally arise in a multiperiod setting.

Monetary, Convex, and Coherent Risk Measures

The uncertainty in the future value of a portfolio is usually described by a function $X : \Omega \rightarrow \mathbb{R}$, where Ω is a fixed set of scenarios. For instance, X can be the (discounted) value of the portfolio or the sum of its P&L and some economic capital. The goal is to determine a number $\rho(X)$ that quantifies the risk and can serve as a capital requirement, that is, as the minimal amount of capital that, if added to the position and invested in a risk-free manner, makes the position acceptable. The following axiomatic approach

to such risk measures was initiated in the coherent case by [1] and later extended to the class of convex risk measures in [26, 30, 33]. In the sequel, $\mathcal{X}$ denotes a given linear space of functions $X : \Omega \rightarrow \mathbb{R}$ containing the constants.

Definition 1 A mapping $\rho : \mathcal{X} \rightarrow \mathbb{R} \cup \{+\infty\}$ is called a *monetary risk measure* if $\rho(0)$ is finite and if ρ satisfies the following conditions for all $X, Y \in \mathcal{X}$.

- *Monotonicity*: if $X \leq Y$, then $\rho(X) \geq \rho(Y)$.
- *Cash invariance*: if $m \in \mathbb{R}$, then $\rho(X + m) = \rho(X) - m$.

The financial meaning of monotonicity is clear: the downside risk of a position is reduced if the payoff profile is increased. Cash invariance is also called *translation property*; in the normalized case $\rho(0) = 0$ and $\rho(1) = -1$, it is equivalent to *cash additivity*, that is, $\rho(X + m) = \rho(X) + \rho(m)$. This is motivated by the interpretation of $\rho(X)$ as a capital requirement, that is, $\rho(X)$ is the amount that should be raised in order to make X acceptable from the point of view of a supervising agency. Thus, if the risk-free amount m is appropriately added to the position or to the economic capital, then the capital requirement is reduced by the same amount. Note that we work with discounted quantities (cf. [22] for a discussion of forward risk measures and interest rate ambiguity).

Definition 2 A monetary risk measure ρ is called a *convex risk measure* if it satisfies the following:

- *Convexity*: $\rho(\lambda X + (1 - \lambda)Y) \leq \lambda \rho(X) + (1 - \lambda)\rho(Y)$, for $0 \leq \lambda \leq 1$.

Consider the collection of possible future outcomes that can be generated with the resources available to an investor: one investment strategy leads to X , while a second strategy leads to Y . If one *diversifies*, spending only the fraction λ of the resources on the first possibility and using the remaining part for the second alternative, one obtains $\lambda X + (1 - \lambda)Y$. Thus, the axiom of convexity gives a precise meaning to the idea that diversification should not increase the risk. This idea becomes even clearer when we note that, for a monetary risk measure, convexity is in fact equivalent to the weaker requirement of

- *Quasi convexity*: $\rho(\lambda X + (1 - \lambda)Y) \leq \max(\rho(X), \rho(Y))$, for $0 \leq \lambda \leq 1$.

Definition 3 A convex measure of risk ρ is called a *coherent risk measure* if it satisfies the following:

- *Positive homogeneity*: if $\lambda \geq 0$, then $\rho(\lambda X) = \lambda \rho(X)$.

Under the assumption of positive homogeneity, the convexity of a monetary risk measure is equivalent to the following:

- *Subadditivity*: $\rho(X + Y) \leq \rho(X) + \rho(Y)$.

This property allows to decentralize the task of managing the risk arising from a collection of different positions: if separate risk limits are given to different “desks”, then the risk of the aggregate position is bounded by the sum of the individual risk limits.

Value at Risk at level $\alpha \in]0, 1]$, defined for random variables X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ by

$$V @ R_\alpha(X) = \inf\{m \in \mathbb{R} \mid \mathbb{P}[X + m < 0] \leq \alpha\} \quad (1)$$

is a monetary risk measure that is positively homogeneous but not subadditive and hence not convex (see **Value-at-Risk**). *Average Value at Risk* at level $\lambda \in]0, 1]$,

$$AV @ R_\lambda = \frac{1}{\lambda} \int_0^\lambda V @ R_\alpha(X) \, d\alpha \quad (2)$$

also called *Conditional Value at Risk*, *expected shortfall*, or *Tail Value at Risk* (see **Expected Shortfall**), is a coherent risk measure. Other examples are discussed in Section “Examples and Applications”.

It is sometimes convenient to reverse signs and to emphasize on the *utility* of a position rather than on its risk. Thus, if ρ is a convex risk measure, then $\phi(X) := -\rho(X)$ is called a *concave monetary utility functional*. If ρ is coherent, then ϕ is called a *coherent monetary utility functional*.

Acceptance Sets

A monetary measure of risk ρ induces the set

$$\mathcal{A}_\rho := \{X \in \mathcal{X} \mid \rho(X) \leq 0\} \quad (3)$$

of positions that are acceptable in the sense that they do not require additional capital. The set $\mathcal{A}_\rho$ is called

the *acceptance set* of ρ . One can show that ρ is a convex risk measure if and only if $\mathcal{A}_\rho$ is a convex set and that ρ is positively homogeneous if and only if $\mathcal{A}_\rho$ is a cone. In particular, ρ is coherent if and only if $\mathcal{A}_\rho$ is a convex cone. The acceptance set completely determines ρ , because

$$\rho(X) = \inf\{m \in \mathbb{R} \mid m + X \in \mathcal{A}_\rho\} \quad (4)$$

Moreover, $\mathcal{A} := \mathcal{A}_\rho$ satisfies the following properties.

$$\mathcal{A} \cap \mathbb{R} \neq \emptyset \quad (5)$$

$$\inf\{m \in \mathbb{R} \mid X + m \in \mathcal{A}\} > -\infty \text{ for all } X \in \mathcal{X} \quad (6)$$

$$X \in \mathcal{A}, Y \in \mathcal{X}, Y \geq X \implies Y \in \mathcal{A} \quad (7)$$

Conversely, one can take a given class $\mathcal{A} \subset \mathcal{X}$ of acceptable positions as the primary object. For a position $X \in \mathcal{X}$, we can then define

$$\rho_{\mathcal{A}}(X) := \inf\{m \in \mathbb{R} \mid m + X \in \mathcal{A}\} \quad (8)$$

If $\mathcal{A}$ satisfies the properties (5)–(7), then $\rho_{\mathcal{A}}$ is a monetary risk measure. If $\mathcal{A}$ is convex, then so is $\rho_{\mathcal{A}}$. If $\mathcal{A}$ is a cone, then $\rho_{\mathcal{A}}$ is positively homogeneous. Note that, with this notation, equation (4) takes the form $\rho_{\mathcal{A}_\rho} = \rho$. The validity of the analogous identity $\mathcal{A}_{\rho_{\mathcal{A}}} = \mathcal{A}$ requires that $\mathcal{A}$ satisfies a certain closure property.

Dual Representation

Suppose now that $\mathcal{X}$ consists of measurable functions on $(\Omega, \mathcal{F})$. A *dual representation* of a convex risk measure ρ has the form

$$\rho(X) = \sup_{Q \in \mathcal{M}} (E_Q[-X] - \alpha(Q)) \quad (9)$$

Here, $\mathcal{M}$ is a set of probability measures on $(\Omega, \mathcal{F})$ such that $E_Q[X]$ is well defined for all $Q \in \mathcal{M}$ and $X \in \mathcal{X}$. The functional $\alpha : \mathcal{M} \rightarrow \mathbb{R} \cup \{+\infty\}$ is called *penalty function*.

The elements of $\mathcal{M}$ can be interpreted as possible probabilistic models, which are taken more or less seriously according to the size of the penalty $\alpha(Q)$. Thus, the value $\rho(X)$ is computed as the worst-case expectation taken over all models $Q \in \mathcal{M}$ and penalized by $\alpha(Q)$ (see [8, 28, 42]).

In the dual representation theory of convex risk measures, one aims at deriving a representation (9)

in a systematic manner. The general idea is to apply convex duality. For every $Q \in \mathcal{M}$, we define the *minimal penalty function* of ρ by

$$\alpha_\rho(Q) := \sup_{X \in \mathcal{X}} (E_Q[-X] - \rho(X)) = \sup_{X \in \mathcal{A}_\rho} E_Q[-X] \quad (10)$$

With additional assumptions on the structure of $\mathcal{X}$ and on continuity properties of ρ , it is often possible to derive the representation

$$\rho(X) = \sup_{Q \in \mathcal{M}} (E_Q[-X] - \alpha_\rho(Q)) \quad (11)$$

via Fenchel–Legendre duality. In this case, ρ is coherent if and only if α_ρ takes only the values 0 and $+\infty$, and so

$$\rho(X) = \sup_{Q \in \mathcal{Q}_\rho} E_Q[-X] \quad (12)$$

where $\mathcal{Q}_\rho$ consists of all $Q \in \mathcal{M}$ with $\alpha_\rho(Q) = 0$. We now discuss some situations in which representation (11) can be obtained. In general, however, it may be necessary to consider extended sets $\mathcal{M}$ that also contain, for example, finitely additive set functions. Dual representation theory goes back to [1, 18, 26, 30, 32–34].

First, let $\mathcal{X}$ be the space of all bounded measurable functions on $(\Omega, \mathcal{F})$. Then, every convex risk measure ρ takes only finite values and is Lipschitz continuous with respect to the supremum norm. For $\mathcal{M}$, we can take the set of all probability measures on $(\Omega, \mathcal{F})$. The validity of the dual representation (9) implies that ρ is *continuous from above* in the sense that

$$X_n \searrow X \implies \rho(X_n) \nearrow \rho(X) \quad (13)$$

On the other hand, the condition of *continuity from below*,

$$X_n \nearrow X \implies \rho(X_n) \searrow \rho(X) \quad (14)$$

is equivalent to the validity of the *strong representation*

$$\rho(X) = \max_{Q \in \mathcal{M}} (E_Q[-X] - \alpha_\rho(Q)) \quad (15)$$

in which for every $X \in \mathcal{X}$ the maximum is attained by some $Q \in \mathcal{M}$. In particular, continuity from below is stronger than the continuity from above. Continuity

from above is equivalent to the so-called *Fatou property*,

$$\liminf_{n \uparrow \infty} \rho(X_n) \geq \rho(X) \text{ for any bounded sequence } (X_n) \text{ converging pointwise to } X \quad (16)$$

Continuity from below is equivalent to the stronger *Lebesgue property*,

$$\lim_{n \uparrow \infty} \rho(X_n) = \rho(X) \text{ for any bounded sequence } (X_n) \text{ converging pointwise to } X \quad (17)$$

Next, we fix a reference probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ and consider the case in which $\mathcal{X} = L^p := L^p(\Omega, \mathcal{F}, \mathbb{P})$ for some $p \in [1, \infty]$. This choice implicitly requires that $\rho(X) = \rho(\tilde{X})$ whenever $X = \tilde{X}$ $\mathbb{P}$ -almost surely. For $\mathcal{M}$, we take the set of all probability measures that are absolutely continuous with respect to $\mathbb{P}$ and whose density belongs to L^q , where $q = p/(p-1)$ is the dual exponent.

The space $\mathcal{X} = L^\infty$ can be regarded as a subset of the space of all bounded measurable functions, and so all corresponding results carry over. In addition, continuity from above (or, equivalently, the Fatou property) of ρ is now even equivalent to a dual representation (11) in terms of probability measures.

For a convex risk measure ρ on $\mathcal{X} = L^p$ with $1 \leq p < \infty$, the existence of a dual representation (11) is equivalent to the lower semicontinuity of ρ with respect to the standard L^p -norm. If ρ takes only finite values, then it is even Lipschitz continuous and admits a strong representation (15). Here, we assume for simplicity that $(\Omega, \mathcal{F}, \mathbb{P})$ is atomless and L^2 is separable.

For the discussion of the dual representation of convex risk measures on spaces of bounded measurable functions, we refer to [28, 37, 38]. For L^p spaces, see [23, 36] and the references therein. Representation theory on Orlicz spaces is considered in [9].

Examples and Applications

In this section, we take $\mathcal{X} = L^\infty(\Omega, \mathcal{F}, \mathbb{P})$. The class of convex risk measures comprises many of the common valuation methods in finance and economics. The risk-neutral expectation in a nice arbitrage-free

market model, for instance, clearly corresponds to a coherent risk measure. If the market model is incomplete, then the cost of superhedging a position $X \in \mathcal{X}$ is given by the coherent risk measure

$$\sup_{Q \in \mathcal{P}} E_Q[-X] \quad (18)$$

where $\mathcal{P}$ is the set of equivalent local martingale measures (see **Superhedging**). If one imposes additional convex trading constraints, the cost of superhedging is a convex risk measure whose representation (9) is explicitly known (see [24, 26, 28]).

Let us now consider the case in which valuation of positions $X \in \mathcal{X}$ is based on the expected utility $\mathbb{E}[u(X)]$ for a concave and strictly increasing function $u : \mathbb{R} \rightarrow \mathbb{R}$. Then, a position can be called *acceptable* if $E_Q[u(X)]$ is bounded from below by $u(c)$ for a given threshold c . The set

$$\mathcal{A} := \{X \in \mathcal{X} \mid E_Q[u(X)] \geq u(c)\} \quad (19)$$

is a valid and convex acceptance set. Hence, $\rho_{\mathcal{A}}$, defined by equation (8), is a convex risk measure called *utility-based shortfall risk measure*. It is continuous from below and admits the strong representation (15) with minimal penalty function

$$\alpha_{\rho}(Q) = \inf_{\lambda > 0} \frac{1}{\lambda} \left(\mathbb{E} \left[\tilde{u} \left(\lambda \frac{dQ}{d\mathbb{P}} \right) \right] - u(c) \right) \quad (20)$$

where $\tilde{u}(y) = \sup_x (u(x) - xy)$ denotes the convex conjugate function of u (see [26] or [28]). In the exponential utility case with $u(x) = -e^{-\theta x}$ for some $\theta > 0$, we obtain for $c = 0$ the *entropic risk measure*,

$$\rho^{\text{ent}}(X) = \frac{1}{\theta} \log \mathbb{E} [e^{-\theta X}] \quad (21)$$

Its minimal penalty function is given by $\alpha_{\rho}(Q) = \theta^{-1} H(Q|\mathbb{P})$, where

$$H(Q|\mathbb{P}) = \mathbb{E} \left[\frac{dQ}{d\mathbb{P}} \log \frac{dQ}{d\mathbb{P}} \right] \quad (22)$$

is the relative entropy of $Q \ll \mathbb{P}$ (see [28]). For the role of entropic risk measures in problems of risk transfer, (see [3]).

To introduce another closely related class of concave monetary utility functionals, let $g : [0, \infty[\rightarrow \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous convex function satisfying $g(1) < \infty$ and the superlinear growth

condition $g(x)/x \rightarrow +\infty$ as $x \uparrow \infty$. Associated to it is the *g-divergence*

$$I_g(Q|\mathbb{P}) := \mathbb{E} \left[g \left(\frac{dQ}{d\mathbb{P}} \right) \right], \quad Q \ll \mathbb{P} \quad (23)$$

as introduced in [14, 15]. The *g-divergence* $I_g(Q|\mathbb{P})$ can be interpreted as a statistical distance between the hypothetical model Q and the reference measure $\mathbb{P}$, so that $\gamma_g(Q) := I_g(Q|\mathbb{P})$ is a natural choice for a penalty function. The risk measure

$$\rho_g(X) := \sup_{Q \ll \mathbb{P}} (E_Q[-X] - I_g(Q|\mathbb{P})) \quad (24)$$

corresponding to such a divergence penalty, is continuous from below, and $I_g(\cdot|\mathbb{P})$ is its minimal penalty function. The corresponding concave utility functional, $\phi_g = -\rho_g$, was called *optimized certainty equivalent* in [5]. This name stems from the variational identity

$$\phi_g(X) = \sup_{z \in \mathbb{R}} (\mathbb{E}[u(X - z)] + z), \quad X \in L^{\infty} \quad (25)$$

where $u(x) = \inf_{z > 0} (xz + g(z))$ is the concave conjugate function of g (see [4, 47]). In [13], the name *divergence utility* is used. Note that the particular choice $g(x) = x \log x$ corresponds to the relative entropy, $I_g(Q|\mathbb{P}) = H(Q|\mathbb{P})$, and so ρ_g coincides with the entropic risk measure. Another important example is provided by taking $g(x) = 0$ for $x \leq \lambda^{-1}$ and $g(x) = \infty$ otherwise, so that the corresponding coherent risk measure is given by Average Value at Risk at level λ :

$$\begin{aligned} AV @ R_{\lambda}(X) &= \inf_{Q \in \mathcal{Q}_{\lambda}} E_Q[X] \quad \text{for} \\ \mathcal{Q}_{\lambda} &:= \left\{ Q \ll \mathbb{P} \mid \frac{dQ}{d\mathbb{P}} \leq \frac{1}{\lambda} \right\} \end{aligned} \quad (26)$$

see Definition (1) and **Expected Shortfall**. In this case, we have $u(x) = 0 \wedge x/\lambda$ and hence get the classical duality formula

$$AV @ R_{\lambda}(X) = \frac{1}{\lambda} \inf_{z \in \mathbb{R}} (\mathbb{E}[(z - X)^+] - \lambda z) \quad (27)$$

as a special case of equation (25). The mixtures of Average Value at Risk at various levels λ are called *spectral risk measures*. They are again coherent risk measures and discussed in more detail in **Spectral Measures of Risk**.

Many of these risk measures can be extended in a straightforward manner to spaces of unbounded random variables (see [23] for a systematic study of such extensions). For Gaussian random variables X , Value at Risk, Average Value at Risk, and the spectral risk measures all take the form

$$\rho(X) = \mathbb{E}[-X] + c \cdot \sigma(X) \quad (28)$$

with different constants c ; for the entropic risk measure, the standard deviation $\sigma(X)$ is replaced by the variance $\sigma^2(X)$.

Model uncertainty is another situation in which it is natural to consider risk measures, due to the interpretation of the measures Q in the dual representation (9) as suitably penalized probabilistic models. This idea already appears in robust statistics [34]. More recently, coherent and convex risk measures were applied in obtaining numerical representations of investors who are averse against both risk and model uncertainty [27, 29, 32, 43] or to define *measures of model uncertainty* [16].

In a financial market model, it makes sense to combine risk measurement with dynamic or static hedges. For instance, measuring the residual risk of a position after hedging by a convex risk measure ρ is equivalent to using the convex risk measure that arises as the *inf-convolution* of ρ and the superhedging risk measure, defined at the beginning of this section (cf. [3, 26, 28, 46] and the references therein).

Law-invariant Risk Measures

Here, we discuss those convex risk measures ρ on $\mathcal{X} = L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ that satisfy $\rho(X) = \rho(\tilde{X})$ for random variables $X, \tilde{X} \in \mathcal{X}$ that have the same law under $\mathbb{P}$. These risk measures are usually called *law invariant*. Examples from the preceding section are Average Value at Risk, the spectral risk measures, the utility-based shortfall risk measures, and the optimized certainty equivalents. Under mild conditions on the underlying probability space, every

law-invariant convex risk measure ρ can be represented in the form

$$\rho(X) = \sup_{\mu} \left(\int_{(0,1]} AV @ R_{\lambda}(X) \mu(d\lambda) - \beta(\mu) \right) \quad (29)$$

where the supremum is taken over all Borel probability measures μ on $]0,1]$ and $\beta(\mu)$ is a penalty for μ . Under the additional assumption of continuity from above, this representation was obtained in the coherent case by [41] and later extended by [17, 28, 31, 39]. More recently, it was shown in [35] that the condition of continuity from above can actually be dropped.

Conditional Convex Risk Measures and Time Consistency

A risk measure should take into account the available information, and it should do so in a consistent manner as new information comes in. Here, we limit the discussion to discrete time and fix a filtration $(\mathcal{F}_t)_{t=0,1,\dots}$ on $(\Omega, \mathcal{F}, \mathbb{P})$; for continuous time and the connection to backward stochastic differential equations (BSDEs), (see [10, 20]) and the references therein. A *conditional convex risk measure* at time t is now defined as a map

$$\rho_t : L^\infty \rightarrow L_t^\infty := L^\infty(\Omega, \mathcal{F}_t, \mathbb{P}) \quad (30)$$

which satisfies the obvious conditional versions of monotonicity, cash invariance, and convexity where the constants m and λ are replaced by functions in L_t^∞ . The associated *acceptance set*

$$\mathcal{A}_t := \{X \in L^\infty \mid \rho_t(X) \leq 0\} \quad (31)$$

is conditionally convex (i.e., $\alpha X + (1 - \alpha)Y \in \mathcal{A}_t$ for $X, Y \in \mathcal{A}_t$ and $\mathcal{F}_{t-1}$ -measurable α with $0 \leq \alpha \leq 1$) and it determines ρ_t via

$$\rho_t(X) = \text{ess inf} \{Y \in L_t^\infty \mid X + Y \in \mathcal{A}_t\} \quad (32)$$

The *Fatou property* is now equivalent to a *dual representation* of the form

$$\rho_t(X) = \text{ess sup}_{Q \in \mathcal{M}} (E_Q[-X \mid \mathcal{F}_t] - \alpha_t(Q)) \quad (33)$$

where the conditional penalty function α_t is given by

$$\alpha_t(Q) = \operatorname{ess\,sup}_{X \in \mathcal{A}_t} E_Q[-X | \mathcal{F}_t] \quad (34)$$

The inequality $\geq$ immediately follows from the definition of α_t , and the converse inequality is obtained by using the dual representation of the unconditional convex risk measure $\rho(X) := \mathbb{E}[\rho_t(X)]$ (cf. [6, 11, 21, 25, 45]).

For the conditional entropic risk measure,

$$\rho_t^{\text{ent}}(X) = \frac{1}{\theta} \log \mathbb{E}[e^{-\theta X} | \mathcal{F}_t] \quad (35)$$

the dual representation holds with

$$\alpha_t(Q) = \frac{1}{\theta} \widehat{H}_t(Q|P) \quad (36)$$

where

$$\widehat{H}_t(Q|P) := \mathbb{E} \left[\frac{Z}{Z_t} \log \frac{Z}{Z_t} \middle| \mathcal{F}_t \right] I_{\{Z_t > 0\}} \quad (37)$$

denotes the *conditional entropy* of $Q \in \mathcal{M}$ with respect to P , defined in terms of the densities $Z = dQ/dP$ and $Z_t = dQ/dP|_{\mathcal{F}_t}$.

In our dynamic setting, the key question is how the conditional risk assessments of a financial position at different times are connected to each other.

Definition A dynamic risk measure given by a sequence of conditional convex risk measures $(\rho_t)_{t=0,1,\dots}$ is called *time-consistent* if

$$\rho_{t+1}(X) \leq \rho_{t+1}(Y) \implies \rho_t(X) \leq \rho_t(Y) \quad (38)$$

and this is equivalent to recursiveness:

$$\rho_t = \rho_t(-\rho_{t+1}) \quad \text{for } t = 0, 1, \dots \quad (39)$$

To characterize time consistency in terms of acceptance sets and penalty functions, we define the “myopic” acceptance sets

$$\mathcal{A}_{t,t+1} := \{X \in L_{t+1}^\infty \mid \rho_t(X) \leq 0\} \quad (40)$$

and the corresponding “myopic” penalty functions

$$\alpha_{t,t+1}(Q) := \operatorname{ess\,sup}_{X \in \mathcal{A}_{t,t+1}} E_Q[-X | \mathcal{F}_t] \quad (41)$$

We also assume that the class $\mathcal{Q}^*$ of all equivalent probability measures Q with $\alpha_0(Q) < \infty$ is not

empty. Then time consistency is equivalent to each of the following conditions:

- $\mathcal{A}_t = \mathcal{A}_{t,t+1} + \mathcal{A}_{t+1}$ for $t = 0, 1, \dots$
- For any $Q \in \mathcal{M}$,

$$\alpha_t(Q) = \alpha_{t,t+1}(Q) + E_Q[\alpha_{t+1} | \mathcal{F}_t] \quad \text{for } t = 0, 1, \dots \quad (42)$$

- For any $Q \in \mathcal{Q}^*$ and any $X \in L^\infty$, the process $\rho_t(X) + \alpha_t(Q)$, $t = 0, 1, \dots$ (43)

is a Q -supermartingale

(cf. [2, 7, 11, 12, 19, 25]). Moreover, each condition implies that the dynamic risk measure admits a robust representation in terms of the set $\mathcal{Q}^*$, that is,

$$\rho_t(X) = \operatorname{ess\,sup}_{Q \in \mathcal{Q}^*} (E_Q[-X | \mathcal{F}_t] - \alpha_t(Q)) \quad (44)$$

for all $X \in L^\infty$ and all $t \geq 0$ (cf. [25]).

The entropic dynamic risk measure is time consistent as long as the parameter θ remains constant. On the other hand, time consistency fails for the dynamic risk measure defined by conditional Average Value at Risk. Under the assumption of *law invariance*, the entropic case is in fact the only time-consistent example, if we include the limiting cases $\theta = 0$ and $\theta = \infty$ corresponding to the conditional expected loss under $\mathbb{P}$ and the *conditional worst-case risk measure*,

$$\rho_t(X) = \operatorname{ess\,inf} \{Y \in L_t^\infty \mid Y \geq -X\} \quad (45)$$

respectively (cf. [40]). This suggests to consider weaker versions of time consistency. For example, the supermartingale property above implies that for each $Q \in \mathcal{Q}^*$ the process $\alpha_t(Q)$, $t = 0, 1, \dots$, is a Q -supermartingale, and this is equivalent to the weaker requirement

$$\rho_{t+1}(X) \leq 0 \implies \rho_t(X) \leq 0 \quad (46)$$

that is, $\mathcal{A}_t \subseteq \mathcal{A}_{t+1}$ for all $t = 0, 1, \dots$. In the law invariant case, such weaker notions of consistency may be used for a characterization of utility-based shortfall risk (cf. [48]). The notion of *prudence* introduced in [44] requires

$$X \in \mathcal{A}_t \implies -\rho_{t+s}(X) \in \mathcal{A}_t \quad \text{for all } s \geq 0 \quad (47)$$

and this is characterized by the fact that

$$\rho_t(X) - \sum_{k=0}^{t-1} \alpha_k(Q), \quad t = 0, 1, \dots \quad (48)$$

is a Q -supermartingale for any $Q \in \mathcal{Q}^*$ and any $X \in L^\infty$.

For an infinite time horizon, the supermartingale criteria for time consistency and for prudence both yield almost sure convergence of the capital requirements $\rho_t(X)$ to an asymptotic capital requirement $\rho_\infty(X)$. We may now ask whether the sequence is *asymptotically safe* in the sense that $\rho_\infty(X) \geq -X$, or even *asymptotically precise* in the sense of $\rho_\infty(X) = -X$; note that asymptotic precision can be viewed as a nonlinear analog of martingale convergence. Criteria in terms of acceptance sets and penalty functions are derived in [25] for the time-consistent case and in [44] for the case of prudence.

References

- [1] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**(3), 203–228.
- [2] Artzner, P., Delbaen, F., Eber, J.-M., Heath, D. & Ku, H. (2007). Coherent multiperiod risk adjusted values and Bellman's principle, *Annals of Operational Research* **152**, 5–22.
- [3] Barrieu, P. & El Karoui, N. (2005). Inf-convolution of risk measures and optimal risk transfer, *Finance and Stochastics* **9**(2), 269–298.
- [4] Ben-Tal, A. & Teboulle, M. (1987). Penalty functions and duality in stochastic programming via ϕ -divergence functionals, *Mathematical Operational Research* **12**, 224–240.
- [5] Ben-Tal, A. & Teboulle, M. (2007). An old-new concept of convex risk measures: the optimized certainty equivalent, *Mathematical Finance* **17**(3), 449–476.
- [6] Bion-Nadal, J. (2004). *Conditional Risk Measure and Robust Representation of Convex Conditional Risk Measures*, CMAP preprint 557, Ecole Polytechnique Palaiseau.
- [7] Bion-Nadal, J. (2006). *Dynamic Risk Measuring: Discrete Time in a Context of Uncertainty, and Continuous Time on a Probability Space*, CMAP preprint 596, Ecole Polytechnique Palaiseau.
- [8] Carr, P., Geman, H. & Madan, D. (2001). Pricing and hedging in incomplete markets, *Journal of Financial Economics* **62**, 131–167.
- [9] Cheridito, P. & Li, T. Risk measures on Orlicz hearts, *Mathematical Finance*, to appear.
- [10] Cheridito, P., Delbaen, F. & Kupper, M. (2005). Coherent and convex monetary risk measures for unbounded càdlàg processes, *Finance and Stochastics* **9**, 1713–1732.
- [11] Cheridito, P. Delbaen, F. & Kupper, M. (2006). Dynamic monetary risk measures for bounded discrete-time processes, *Electronic Journal of Probability* **11**, 57–106.
- [12] Cheridito, P. & Kupper, M. (2006). *Composition of Time-Consistent Dynamic Monetary Risk Measures in Discrete Time*, Preprint.
- [13] Cherny, A. & Kupper, M. (2007). *Divergence Utilities*, Preprint, Moscow State University.
- [14] Csaszar, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten, *Magyar Tudományos Akademia Mathematica Kutató International Közl* **8**, 85–108.
- [15] Csaszar, I. (1967). On topological properties of f -divergences, *Studia Scientiarum Mathematicarum Hungarica* **2**, 329–339.
- [16] Cont, R. (2006). Model uncertainty and its impact on the pricing of derivative instruments, *Mathematical Finance* **16**, 519–542.
- [17] Dana, R.-A. (2005). A representation result for concave Schur concave functions, *Mathematical Finance* **15**, 613–634.
- [18] Delbaen, F. (2002). Coherent measures of risk on general probability spaces, *Advances in Finance and Stochastics. Essays in Honour of Dieter Sondermann*, Springer-Verlag, pp. 1–37.
- [19] Delbaen, F. (2006). The structure of m-stable sets and in particular of the set of risk neutral measures, in *Mémoires Paul-André Meyer – Séminaire de Probabilités XXXIX*, M. Yor & M. Émery, eds, Springer, Berlin, pp. 215–258.
- [20] Delbaen, F., Peng, S. & Rosazza Gianin, E. (2008). *Representation of the Penalty Term of Dynamic Concave Utilities*, arXiv:0802.1121.
- [21] Detlefsen, K. & Scandolo, G. (2005). Conditional and dynamic convex risk measures, *Finance and Stochastics* **9**(4), 539–561.
- [22] El Karoui, N. & Ravanelli, C. (2008). Cash sub-additive risk measures and interest rate ambiguity, *Mathematical Finance*, to appear.
- [23] Filipović, D. & Svindland, G. (2008). *Convex Risk Measures Beyond Bounded Risks, or The Canonical Model Space for Law-Invariant Convex Risk Measures is $\$L^1$* , Preprint, University of Vienna.
- [24] Föllmer, H. & Kramkov, D. (1997). Optional decompositions under constraints, *Probability Theory and Related Fields* **109**, 1–25.
- [25] Föllmer, H. & Penner, I. (2006). Convex risk measures and the dynamics of their penalty functions, *Statistics and Decisions* **24**(1), 61–96.
- [26] Föllmer, H. & Schied, A. (2002). Convex measures of risk and trading constraints, *Finance and Stochastics* **6**, 429–447.
- [27] Föllmer, H. & Schied, A. (2002). Robust preferences and convex measures of risk, in *Advances in Finance and Stochastics*, Springer, Berlin, pp. 39–56.

-
- [28] Föllmer, H. & Schied, A. (2004). *Stochastic Finance: An Introduction in Discrete Time*, 2nd revised and extended edition, *de Gruyter Studies in Mathematics* 27 Walter de Gruyter, Berlin.
 - [29] Föllmer, H., Schied, A., Weber, S. (2007). Robust preferences and robust portfolio choice, in *Handbook on Mathematical Finance*, A. Bensoussan, ed, Elsevier, Amsterdam, forthcoming.
 - [30] Frittelli, M. & Rosazza Gianin, E. (2002). Putting order in risk measures, *Journal of Banking and Finance* **26**, 1473–1486.
 - [31] Frittelli, M. & Rosazza Gianin, E. (2005). Law-invariant convex risk measures, *Advances in Mathematical Economics* **7**, 33–46.
 - [32] Gilboa, I. & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior, *Journal of Mathematical Economics* **18**, 141–153.
 - [33] Heath, D. (2000). *Back to the Future*, Plenary Lecture, First World Congress of the Bachelier Finance Society, Paris.
 - [34] Huber, P. (1981). Robust statistics, *Wiley Series in Probability and Mathematical Statistics*, Wiley, New York.
 - [35] Jouini, E., Schachermayer, W. & Touzi, N. (2006). Law invariant risk measures have the Fatou property, *Advances in Mathematical Economics* **9**, 49–71.
 - [36] Kaina, M. & Rüschendorf, L. On convex risk measures on L_p -spaces, *Mathematical Methods in Operations Research*, to appear.
 - [37] Krätschmer, V. (2005). Robust representation of convex risk measures by probability measures, *Finance and Stochastics* **9**, 597–608.
 - [38] Krätschmer, V. (2007). *On Sigma-Additive Robust Representation of Convex Risk Measures for Unbounded Financial Positions in the Presence of Uncertainty About the Market Model*, Preprint, TU, Berlin.
 - [39] Kunze, M. (2003). *Verteilungsinvariante konvexe Risiko- maße*, Diplomarbeit, Humboldt-Universität zu Berlin.
 - [40] Kupper, M. & Schachermayer, W. (2008). *Representation Results for Law Invariant Time Consistent Functions*, Preprint TU, Vienna.
 - [41] Kusuoka, S. (2001). On law invariant coherent risk measures, *Advances in Mathematical Economics* **3**, 83–95.
 - [42] Larsen, K., Pirvu, T., Shreve, S. & Tütüncü, R. (2005). Satisfying Convex Risk Limits by Trading, *Finance and Stochastics* **9**(2), 177–195.
 - [43] Maccheroni, F., Marinaci, M. & Rustichini, A. (2006). Ambiguity aversion, robustness, and the variational representation of preferences, *Econometrica* **74**, 1447–1498.
 - [44] Penner, I. (2007). *Dynamic Convex Risk Measures: Time Consistency, Prudence, and Sustainability*, Ph.D. Thesis, Humboldt-Universität zu Berlin, Berlin.
 - [45] Riedel, F. (2004). Dynamic coherent risk measures, *Stochastic Processes and their Applications* **112**(2), 185–200.
 - [46] Schied, A. (2006). Risk measures and robust optimization problems, *Stochastic Models* **22**, 753–831.
 - [47] Schied, A. (2007). Optimal investments for risk- and ambiguity-averse preferences: a duality approach, *Finance and Stochastics* **11**(1), 107–129.
 - [48] Weber, S. (2006). Distribution-invariant risk measures, information, and dynamic consistency, *Mathematical Finance* **16**, 419–442.

Related Articles

Convex Duality; Expected Shortfall; Expected Utility Maximization; Duality Methods; Risk Measures; Statistical Estimation; Spectral Measures of Risk; Superhedging; Utility Function; Utility Indifference Valuation; Value-at-Risk.

HANS FÖLLMER & ALEXANDER SCHIED

Value-at-Risk

The Value at Risk (VaR) is the maximum likely loss that a portfolio experiences over a specified horizon period, where the key term is “likely” to be interpreted in terms of a specified probability, known as a *confidence level*. For example, if the daily VaR at the 95% confidence level is \$1m, then we would expect a maximum loss of \$1m on the best prospective 95 days out of 100. More formally, the VaR at the confidence level p can be defined as the $100 \times p\%$ quantile of the loss distribution [8, 9, 16]. Note, therefore, that the VaR is measured in units of loss: a positive VaR means that the likely worst outcome is a monetary loss, whereas a negative VaR implies that the likely worst outcome is a profit. Also note that although we have defined the VaR here in terms of a quantile of the loss distribution, the VaR can equivalently be defined in terms of a quantile of the profit/loss (P/L) distribution (i.e., the negative of our loss distribution) or in terms of a quantile of the distribution of financial returns (*see Market Risk*).

The values of the parameters on which the VaR is defined, the horizon period and the confidence level, would be chosen by the user and depend on the context and the use to which the VaR is put. The horizon might be anything from a fraction of a day to decades ahead, but is typically somewhere between 1 day and a month. Confidence levels are typically in the 95%+ range and can be well above 99%.

Virtually unknown in 1990, the VaR risk measure rose rapidly to prominence in the early 1990s, as leading financial institutions competed with one another to build risk models that would forecast daily VaRs on traded market portfolios. The best known was Morgan’s RiskMetrics model, which was published on the web in October 1994. The resulting debate on the merits and demerits of the RiskMetrics model did much to stimulate interest in and awareness of the issues involved in risk modeling [13, 14]. Since then, risk models have become much more sophisticated, and VaR methodologies have been used to forecast other types of risks such as credit, operational, liquidity, and cash flow risks.

Uses of VaR

VaR models have been used for many different risk management purposes. They can be used by senior

management to determine an institution’s overall risk target and can be used to determine risk targets and position limits down the line. They can be used to determine internal capital requirements, both at the level of the firm and at the level of individual business units. They can also be used to identify risk-expected return trade-offs and to assess investment, trading, and hedging decisions. They can be used to determine the remuneration of managers and traders in ways that take into account not just the profits made but also the risks taken to achieve those profits. And, finally, they can be used for risk reporting and risk disclosure purposes.

Attractions of VaR as a Risk Measure

The VaR has a number of attractive properties as a risk measure: (i) It is expressed in the simplest and most easily understood unit of measurement, namely, “lost money”. (ii) The VaR provides a common consistent risk measure across different positions and risk factors and can be applied to any type of portfolio. It enables us to compare the risk of an equity position, say, against that of a fixed-income position, and do so in a consistent way. The VaR can be applied to a broader range of risks than the traditional measures such as beta (which is restricted to equity and commodity risks), duration and convexity (which are restricted to fixed-income risks), and the Greeks (which are restricted to derivatives risks). (iii) The VaR enables us to aggregate the risks of individual positions in a way that takes into account how the risk factors in those positions interact, that is, the VaR can take account of correlations between different risk factors. For their part, most traditional risk measures do not easily allow for the “sensible” aggregation of component risks. (iv) The VaR is holistic in the sense that it takes into account all risk factors, whereas traditional risk measures look at one risk factor at a time (e.g., the Greeks) or collapse multiple risk factors into one (e.g., duration, beta). The VaR is also holistic in the sense that it can readily be applied to forecast a firm’s corporate-wide risk exposure, as well as the exposures of individual desks or business units. (v) The VaR is probabilistic and gives a risk manager a sense of the probabilities associated with possible loss outcomes. Traditional risk measures do not give any such information and only give the answers to questions about “what if?” scenarios.

Limitations of the VaR as a Risk Measure

The VaR also has major limitations as a risk measure. The underlying problem with the VaR is that it takes no account of the sizes of possible losses bigger than itself. It tells us the worst we can expect on the, say, 95 “good” days out of a possible 100, but tells us nothing about we can expect on the “bad” 5 remaining days. It is therefore “blind” to the sizes of tail losses and does not tell us how bad “bad” might be. This is a serious drawback for risk managers concerned with the managing of their firm’s exposure to “lower tail” loss outcomes.

A consequence of this “tail blindness” is that the VaR risk measure fails to satisfy the important property of subadditivity [3, 4]. A risk measure $\rho(\cdot)$ is said to be subadditive if, for *any* two position A and B , the risk measure always satisfies the condition $\rho(A + B) \leq \rho(A) + \rho(B)$, that is, the risk of the sum is always less than or equal to the sum of the risks. This subadditivity condition reflects the intuitive expectation that a “reasonable” risk measure should exhibit some or, at worst, no diversification when two positions are aggregated. We would certainly *not* expect the overall risk exposure as measured by a “respectable” risk measure to increase when individual risks are put together (*see Convex Risk Measures*).

Yet the VaR does not satisfy this basic property. For example, consider the case where we have 100 bonds and 100 possible equally likely states of nature, in each one of which one of the bonds defaults and inflicts a loss of \$10 on the holder. The probability of loss on any given bond is therefore 1%, but the probability of loss on a fully diversified portfolio is 100%. Thus, if an investor holds any one of these bonds, the VaR at the 95% confidence level will be 0, but if the investor holds a fully diversified bond portfolio, the VaR at the 95% confidence level will be equal to $\$10/100 = \0.1 . This example suffices to prove that the VaR is not subadditive.

VaRs “tail blindness” can lead to other problems. For example, if a prospective investment involves the possibility of low-probability high loss, a VaR-based decision calculus might suggest that the investor should go ahead with the investment regardless of the size of the possible loss, provided only that the probability of loss is small enough. In effect, the risk involved can fall under the VaRs radar screen. Such a categorical acceptance of a risk—regardless

of the size of the possible loss, provided that only the probability involved is sufficiently low—undermines sensible risk-return analysis and can leave an investor very exposed without realizing it.

A further problem that can occur in the context of decentralized decision making—that is, when traders or asset managers are managing other people’s funds—is the danger of traders or asset managers “gaming” VaR systems that are used to control or remunerate their risk-taking. A standard example is where traders facing a VaR-based risk target have an incentive to sell out-of-the-money options that lead to greater income in most states of the world and the occasional large hit when the firm is unlucky. If the options are suitably chosen, the probabilities of them expiring in the money will be sufficiently low that they will have no impact on the VaR. The trader then benefits from the higher income and hence higher bonuses earned in “normal” times when the options expire worthless. The fact that the VaR does not take into account the possible large losses involved can distort incentives and encourage traders to “game” the VaR system and promote their own interests at the expense of the institutions that employ them.

Alternative Risk Measures

These problems have encouraged research into alternative risk measures—most particularly, other quantile-based risk measures that share the attractions of the VaR but avoid its main drawbacks. The seminal work in this area is the theory of coherent risk measures proposed by Artzner *et al.* [3, 4]. The best-known coherent risk measure is the expected shortfall (ES) also referred to as *conditional VaR* (*CVaR*), which is the loss we can expect if a tail loss should occur (*see Convex Risk Measures; Expected Shortfall; Market Risk*). Hence, although the VaR gives us the maximum loss we can expect if a tail loss does not occur, the ES tells us what to expect in the tail itself. The ES is therefore not “blind” to the tail in the way that the VaR is. At the same time, the ES shares the attractive properties of the VaR—it is measured in units of lost money, provides a common consistent measure of risk across different positions, is probabilistic, and so forth. Another type of coherent risk measure is a spectral risk measure that takes into account a user’s risk aversion ([1, 2]; *see Spectral Measures of Risk*). It is also interesting to note

that the outcomes of scenario analyses are coherent risk measures as well (*see Stress Testing*). Thus, the ES, spectral risk measures and scenario analyses each give us financial risk measures that are “respectable” in a sense that the VaR is not.

When to Choose the VaR?

This raises a natural question: when would a rational user choose the VaR as a preferred risk measure? The answer is that although the VaR is not generally a suitable risk measure for the reasons already mentioned, it *can* be suitable in special case circumstances where we restrict either the return distribution or the user’s utility function (or, if you prefer, the user’s risk preferences). An example of the former would be where a user believes that he/she faces normally (or, more generally, elliptically) distributed risks. In such circumstances, the VaR obeys the sub-additivity property and might be used both to measure risks in a portfolio analysis framework and to determine capital requirements. Examples of the latter would be where a user has a quadratic utility function (which implies that the user does not care about the skewness or kurtosis of the return distribution), a lexicographic utility function (in which the risk-expected return trade-off becomes degenerate) or, in the context of the theory of lower partial moments, where the user has a lower partial moment of zero (which reflects a rather extreme form of negative risk aversion [12]). However, each of these special cases is implausible: most empirical return distributions are not elliptical, and experts would regard quadratic and lexicographic utility functions and lower partial moments of zero as inconsistent with any notions of “well-behaved” risk preferences. The implausibility of these special cases reinforces the point that the VaR is not, in general, an advisable risk measure to choose.

This said, the special case circumstances where VaR is appropriate include probability-of-ruin problems, where the VaR is the quantile associated with the probability of ruin [11]. An important example of this type of problem is where the VaR might be used to determine the capital requirements necessary to achieve some target probability of insolvency. Typically, the target insolvency probability would be set to some very small value p and the capital requirement would be equal to the VaR at the confidence level $1 - p$. Since p is very small, the VaR itself

should then be estimated by using extreme value (EV) methods rather than some arbitrary approach picked off the shelf, as it were (*see Operational Risk*).

Methods to estimate VaR are discussed in **Market Risk**.

VaR in Practice

There is considerable evidence that real-world VaR estimates are often highly inaccurate. This evidence is both direct, in studies such as [6, 7, 15], and indirect. Indirect evidence comes from the many instances since the early 1990s where firms suffered large losses, such as alleged “25 sigma” losses, that are much greater than their VaR models anticipated [10]. The causes of this inaccuracy include model risk (i.e., that we do not in practice know what the “true” distributions actually are; *see Models*), parameter risk (i.e., that, for given any distribution, we do not know the “true” values of its parameters but have to work with estimates of those parameters instead), and implementation risk (i.e., that different institutions will implement any given VaR system in different ways, so leading to different VaR estimates).

In practice, VaR systems often perform fairly well under “normal” stable market conditions, but perform badly during crises. A common problem is where VaR models predicated on “everyday” confidence levels (such as 95 or even 99%) perform very poorly when “tail events” when very large losses occur. Other common problems are the radicalization of correlations and the disappearance of market liquidity during crises. VaR models can also perform badly in circumstances where they are used to control or remunerate traders, who respond by “gaming” the models and exploiting their weaknesses. There is also the problem of systemic risk, where a bank’s risk models perform poorly because they failed to allow for other institutions following similar risk management strategies—a problem analogous to the mistake a cinema-goer makes when he fails to anticipate that everyone else will also run for the exit in the event of a fire [5].

All these problems can contribute to institutions experiencing much larger losses than their risk models anticipated. They also highlight the importance of backtesting risk models and complementing risk forecasts with stress tests and scenario analyses that seek to determine an institution’s vulnerability to “what

if?” events that might not be present in the data used to calibrate those models ([9]; see **Stress Testing; Backtesting**).

References

- [1] Acerbi, C. (2002). Spectral measures of risk: a coherent representation of subjective risk aversion, *Journal of Banking and Finance* **26**, 1505–1518.
- [2] Acerbi, C. (2004). Coherent representations of subjective risk aversion, in *Risk Measures for the 21st Century*, G. Szegö, ed, John Wiley & Sons, New York, p. 147–207.
- [3] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1997). Thinking coherently, *Risk* **10**, 68–71.
- [4] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**, 203–228.
- [5] Basak, S. & Shapiro, A. (2001). Value-at-risk-based risk management: optimal policies and asset prices, *Review of Financial Studies* **14**, 371–405.
- [6] Beder, T. (1995). VaR: seductive but dangerous, *Financial Analysts Journal* **51**, 12–24.
- [7] Berkowitz, J. & O’Brien, J. (2002). How accurate are value-at-risk models at commercial banks? *Journal of Finance* **57**, 1093–1112.
- [8] Crouhy, M., Galai, D. & Mark, R. (2001). *Risk Management*. McGraw Hill, New York.
- [9] Dowd, K. (2005). *Measuring Market Risk*, 2nd Edition, John Wiley & Sons, Chichester.
- [10] Dowd, K., Cotter, J., Humphrey, C.G. & Woods, M. (2008). How unlucky is 25-sigma? *Journal of Portfolio Management*, **34**(4), 76–80.
- [11] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extreme Events for Insurance and Finance*, Springer Verlag, Berlin.
- [12] Grootveld, H. & Hallerbach, W.G. (2004). Upgrading value-at-risk from diagnostic metric to decision variable: a wise thing to do?, in *Risk Measures for the 21st Century*, G. Szegö, ed, Wiley, New York, p. 33–50.
- [13] Lawrence, C. & Robinson, G. (1995). How safe is RiskMetrics? *Risk* **8**, 26–29.
- [14] Longestae, J. & Zangari, P. (1995). A transparent tool. *Risk* **8**, 30–32.
- [15] Marshall, C. & Siegel, M. (1997). Value at risk: implementing a risk measurement standard. *Journal of Derivatives* **4**, 91–110.
- [16] McNeil, A.J., Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, Tools*, Princeton University Press, Princeton.

Further Reading

Beirlant, J., Goegebeur, Y., Teugels, J., Seger, J., De Waal, D. & Ferro, C. (2004). *Statistics of Extremes: Theory and Applications*, John Wiley & Sons, Chichester.

Related Articles

Backtesting; Convex Risk Measures; Expected Shortfall; Economic Capital Allocation; Extreme Value Theory; Market Risk; Models; Operational Risk; Ruin Theory; Simulation-based Estimation; Spectral Measures of Risk; Stress Testing; Risk Measures; Statistical Estimation.

KEVIN DOWD

Expected Shortfall

The synonymous terms conditional Value at Risk (CV@R) and expected shortfall (ES) refer to a particular statistics of a random variable, which is widely adopted as a measure of portfolio risk. It is also known as tail Value at Risk (TV@R) or average Value at Risk (AV@R) among other names. Here we adopt the name ES.

The “ES of a random variable X ”, denoted by $ES_\alpha(X)$, is most easily defined in plain English as *minus the average of the α lowest outcomes of X* , where $\alpha \in [0, 1]$ is a chosen confidence level (e.g., $\alpha = 5\%$). There are several equivalent mathematical expressions of this definition. We discuss the most useful ones in the section “Mathematical Expressions for ES”.

In financial risk management, the “ES of a portfolio π ” is usually meant as the expected shortfall $ES_\alpha(r_T^{(\pi)})$ of the return $r_T^{(\pi)} = \pi_T - (1 + R)\pi_0$ between the current date $t = 0$ and a chosen future date $t = T$, in perfect analogy with the practice introduced by V@R (see **Value-at-Risk**). Here, R is the risk-free interest on $[0, T]$. It is then clear that $ES_\alpha(r_T^{(\pi)})$ represents the *average portfolio loss at time T in the α worst cases*.

Measures of risk similar to ES (in particular TCE, equation (2), and “average shortfall” [10]) appeared in the practice of risk management already in the early 1990s, along with the diffusion of V@R methods. The main motivation for their introduction was to remedy the fact that V@R is totally blind to the severity of the losses contained in the tail beyond the specified confidence level. The idea of averaging the outcomes contained in the tail instead of detecting only its threshold quantile is so natural that it came simultaneously from several academicians and practitioners, and this is probably the reason why no consensus was ever achieved on the name of this risk measure.

The interest around ES increased considerably after the appearance of [5] with the advent of coherent measures of risk (CMRs) (see **Convex Risk Measures**) and the consequent criticism that was addressed to V@R when it was recognized to be not subadditive, that is, not compatible with the risk diversification principle. ES, in fact, in the modern version introduced in [1, 12], turned out to be a subadditive measure of risk and, in fact, it provides

the best-known example of a coherent measure of risk.

A lively debate has arisen in the past years on whether ES should replace V@R in the regulatory framework of the Basel Committee on Banking Supervision. A number of arguments [3, 5, 6] have been given to display the superiority of a coherent measure of risk for capital adequacy purposes, like absence of regulatory arbitrage opportunities, huge computational advantages related to subadditivity and convexity, and absence of risk paradoxes. No sound arguments have ever been given to display why V@R should be more appropriate than ES, instead. Yet, to date, ES still plays no role in the Basel II Accord (see **Regulatory Capital**).

Mathematical Expressions for ES

We denote with $F_X(x) = \mathbb{P}\{X \leq x\}$ the probability distribution function of X in a chosen probability space and we define for any $p \in (0, 1)$ the lower p -quantile $q_p(X) = \inf\{x \in \mathbb{R} | F_X(x) \geq p\}$ and the upper p -quantile $q^p(X) = \inf\{x \in \mathbb{R} | F_X(x) > p\}$. A (generic) p -quantile X_p is defined as any value in the interval $[q_p(X), q^p(X)]$.

For any $\alpha \in (0, 1]$, ES_α can then be expressed as

$$ES_\alpha(X) = -\frac{1}{\alpha} \int_0^\alpha X_p \, dp \quad (1)$$

where X_p is a p -quantile whose choice is irrelevant, showing that there exists a unique definition for ES. For V@R, on the contrary, which is defined as $V@R_\alpha(X) = -X_\alpha$, this choice does indeed affect the result, leading to possible alternative different notions (upper and lower V@R, for instance) when the quantiles $p \mapsto X_p$ turn out not to be single valued.

For the extreme case $\alpha = 0$, the definition is naturally extended by placing $ES_0(X) = -\text{ess inf } X = -\inf\{x \in [-\infty, +\infty] | F_X(x) > 0\}$, while for $\alpha = 1$ one easily recognizes that $ES_1(X) = -\mathbb{E}[X]$.

Note how equation (1) clearly expresses the fact that ES is indeed *minus the average of the α lowest outcomes of X* .

It is important to distinguish ES from the so-called tail conditional expectation (TCE) defined as [5]

$$TCE_\alpha(X) = -\mathbb{E}[X | X \leq q_\alpha(X)] \quad (2)$$

This expression can be shown to coincide with $ES_\alpha(X)$ only under special conditions on the distribution function, for instance, when $F_X(x)$ is a continuous function. However, for general distribution functions, ES and TCE differ and it is only the former, which can be shown to be subadditive [1, 12]. Equation (2) is, in any case, a useful expression for ES when one works with parametric distributions whose continuity is assumed to hold *a priori*.

The correction to expression (2) needed to express ES on general (noncontinuous) probability distributions is contained in the following formula [1]:

$$ES_\alpha(X) = -\frac{1}{\alpha} \{ \mathbb{E} [X \mathbf{1}_{\{X \leq X_\alpha\}}] - X_\alpha(F_X(X_\alpha) - \alpha) \} \quad (3)$$

where the choice of the α -quantile X_α is again irrelevant. This formula shows that $ES_\alpha(X) = TCE_\alpha(X)$ if and only if $F_X(X_\alpha) = \alpha$.

ES can also be defined as the solution of a convex minimization problem [9, 11, 12]:

$$ES_\alpha(X) = \min_s \left\{ -s + \frac{1}{\alpha} \mathbb{E} [s - X]^+ \right\} \quad (4)$$

attaining the minimum in any $s \in [q_\alpha(X), q^\alpha(X)]$. This formulation of ES is very important because it allows to obtain fast algorithms for portfolio ES-optimization [11, 12].

Properties of ES

ES satisfies the following important properties for any $\alpha \in [0, 1]$ and random variables X, Y [1, 7, 9]:

(i) *monotonicity*

$$ES_\alpha(X) \geq ES_\alpha(Y) \quad \text{if } X \leq Y \text{ a.s.} \quad (5)$$

(ii) *translational invariance*

$$ES_\alpha(X + c) = ES_\alpha(X) - c \quad \text{if } c \in \mathbb{R} \quad (6)$$

(iii) *positive homogeneity*

$$ES_\alpha(\lambda X) = \lambda ES_\alpha(X) \quad \text{if } \lambda \geq 0 \quad (7)$$

(iv) *subadditivity*

$$ES_\alpha(X + Y) \leq ES_\alpha(X) + ES_\alpha(Y) \quad (8)$$

Formally, these properties coincide with the axioms of coherence [5]. However, it has to be kept in mind that the concept of coherence is not just a property of a statistics of a generic random variable but a property of a portfolio statistics with a precise economic interpretation. To speak of coherence of a portfolio measure of risk, it is also necessary that the random variables X and Y in these conditions represent future “portfolio values” (as opposed to “portfolio returns”). This fact is crucial for the correct economic interpretation of condition (6) in the spirit of capital adequacy. A CMR is, in fact, supposed to decrease by the amount c when a risk-free capital of future deterministic value c (and not a portfolio yielding a deterministic return c) is added to any portfolio X .

Hence, strictly speaking, the *expected shortfall of a portfolio* $ES_\alpha(r_T^{(\pi)})$, widely adopted by practitioners as an alternative to $V@R$, although monotonic, translational invariant, positive homogeneous, and subadditive with respect to its return argument, is, nevertheless, *not* a CMR in the sense of [5]. It is the *expected shortfall of a portfolio value* $ES_\alpha(\pi_T)$, which represents a truly CMR. The connection between the two notions is however just a constant offset, thanks to property (2).

$$ES_\alpha(\pi_T) = ES_\alpha(r_T^{(\pi)}) - (1 + R)\pi_0 \quad (9)$$

Properties (5)–(8) are, in any case, extremely important also for return-based measures, in particular for the convexity properties they induce on portfolio optimization problems. That is why some authors (e.g., [3, 6]) (somehow improperly) call them *axioms of coherence* also when referred to statistics acting on portfolio returns.

Particularly important is the subadditivity property, which $V@R$ fails to satisfy. Inequality (8) formalizes the fact that the risk (expressed by ES) of the sum of two portfolios X and Y never exceeds the sum of their individual risks. The difference between right- and left-hand side of the inequality expresses the hedging benefit of merging the two portfolios. In a CMR this benefit is positive, and just in exceptional cases vanishing, but never negative, a situation that would give rise to paradoxes and

regulatory arbitrages (see, for instance, [3–5] for a discussion).

An important immediate consequence of properties (7) and (8) is convexity.

(v) *convexity*

$$\begin{aligned} ES_\alpha(\mu X + (1 - \mu)Y) \\ \leq \mu ES_\alpha(X) + (1 - \mu)ES_\alpha(Y) \quad \text{if } \mu \in [0, 1] \end{aligned} \quad (10)$$

which states that the risk of a blend of two portfolios is less than the blend of the individual risks. This expresses the clearest formalization of the risk diversification principle and is of utmost importance for portfolio optimization problems [3, 11, 12], because they generally turn out to be convex.

Other important properties of ES are related to the concepts of portfolio preference under first or second stochastic dominance. We recall that a random variable X is said to dominate Y according to first-order stochastic dominance (notation: $X \succ^{1st} Y$) if $F_X(z) \leq F_Y(z)$ for all $z \in \mathbb{R}$, while X is said to dominate Y according to second-order stochastic dominance (notation: $X \succ^{2nd} Y$) if $\int_{-\infty}^z [F_Y(t) - F_X(t)] dt \geq 0$ for all $z \in \mathbb{R}$. First order implies second-order stochastic dominance. It can be shown [7–9] that ES is compatible with both these preferences.

(vi) *monotonicity with respect to first- and second-order stochastic dominance*

$$ES_\alpha(X) \leq ES_\alpha(Y) \quad \text{if } X \succ^{1st, 2nd} Y \quad (11)$$

which tells us that if a portfolio X is preferable to Y under either first- or second-order stochastic dominance, its ES will be smaller.

Furthermore, one can show comonotonic additivity [7, 8].

(vii) *comonotonic additivity*

$$\begin{aligned} ES_\alpha(X + Y) &= ES_\alpha(X) + ES_\alpha(Y) \quad \text{if} \\ &X, Y \text{ comonotonic} \end{aligned} \quad (12)$$

Comonotonic additivity represents the limiting case of subadditivity, achieved when two random variables

happen to be comonotonic and therefore ineffective to hedge each other. It is in this case that the hedging benefit of subadditivity is expected to vanish. Comonotonic additivity is not implied by the axioms of coherence and this is sometimes argued to be a weakness of the axioms, because noncomonotonic additive CMRs can give rise to capital regulatory arbitrages [4].

Furthermore, ES belongs to the class of law-invariant, comonotonic additive CMRs [8], also known as spectral measures of risk (SMRs) [2] (see **Spectral Measures of Risk**). Actually, the set $\{ES_\alpha | \alpha \in [0, 1]\}$ of all possible ES is the smallest set whose convex hull corresponds to the class of SMRs. In other words, any SMR can be uniquely expressed as a convex combination of ESs [8].

Law-invariance plays a special role as the distinguishing property of the CMRs, which may be estimated from data sets.

Estimation of ES

A number of cases are known where exact expressions of ES can be computed directly from equation (1) by direct calculation (analytical or numerical) knowing the distribution of X . However, in general cases, and in particular when parametric distributions are not assumed *a priori* like in many risk management platforms, ES is obtained by estimation.

Given a sample of N independent and identically distributed outcomes x_i , of a random variable X , a consistent estimator of $ES_\alpha(X)$ can be obtained by [1, 3]

$$\begin{aligned} \widehat{ES}_\alpha(X) &= -\frac{1}{N\alpha} \left(\sum_{i=1}^{\lfloor N\alpha \rfloor} x_{i:N} \right. \\ &\quad \left. + (N\alpha - \lfloor N\alpha \rfloor)x_{\lfloor N\alpha \rfloor+1:N} \right) \\ &= \frac{1}{\lfloor N\alpha \rfloor} \sum_{i=1}^{\lfloor N\alpha \rfloor} x_{i:N} + o(1/N) \end{aligned} \quad (13)$$

where $\lfloor n \rfloor$ denotes the integer part of n and $x_{i:N}$ denotes the ordered statistics, that is, the vector $\{x_{i:N}\}_{i=1, \dots, N}$ of values $x_{1:N} \leq \dots \leq x_{N:N}$ obtained sorting the vector $\{x_i\}_{i=1, \dots, N}$. This expression converges to $ES_\alpha(X)$ for $N \rightarrow \infty$. It can be shown to

be nothing but the ES of the empirical distribution of the sample, and as such, it is clearly coherent for any finite N .

In equation (13) the $o(1/N)$ term can be neglected for simplicity in those applications where N is large like some Monte Carlo simulations, but it is not at all negligible, for instance, in a historical simulation of $N = 252$ data, with a confidence level of say, 5%. In general, it is always safer, and not computationally penalizing, to adopt the complete estimator.

An equivalent expression [11, 12] can be obtained by exploiting equation (4):

$$\widehat{ES}_\alpha(X) = \min_s \left\{ -s + \frac{1}{N\alpha} \sum_{i=1}^N [s - x_i]^+ \right\} \quad (14)$$

which has the solution $s \in [x_{\lfloor N\alpha \rfloor}, x_{\lfloor N\alpha \rfloor + 1}]$. From a computational point of view, there may be cases where the cost of minimization of equation (14) is smaller than the cost of the sorting procedure needed for equation (13). These formulas are essential tools for any risk management application based on exact Monte Carlo or historical simulation.

References

- [1] Acerbi, C., Tasche, D. (2002). On the coherence of expected shortfall, *Journal of Banking and Finance* **26**, 1487–1503.
- [2] Acerbi, C. (2002). Spectral measures of risk: a coherent representation of subjective risk aversion, *Journal of Banking and Finance* **26**, 1505–1518.
- [3] Acerbi, C. (2004). Coherent representations of subjective risk aversion, in *Risk Measures for the 21st Century*, G. Szego, ed., Wiley & Sons, pp. 147–207.
- [4] Acerbi, C. (2007). Coherent measures of risk in everyday market practice, *Quantitative Finance* **7**(4), 359–364.
- [5] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**(3), 203–228.
- [6] Embrechts, P., Frey, R. & McNeil, A.J. (2005). *Quantitative Risk Management*, Princeton University Press.
- [7] Föllmer, H. & Schied, A. (2004). *Stochastic Finance*, 2nd Edition, de Gruyter.
- [8] Kusuoka, S. (2001). On law invariant coherent risk measures, *Advances in Mathematical Economics* **3**, 83–95.
- [9] Pflug, G. (2000). Some remarks on the value-at-risk and the conditional value-at-risk, in *Probabilistic Constrained Optimization: Methodology and Applications*, S. Uryasev, ed., Kluwer Academic Publishers, pp. 278–287.
- [10] Rappoport, P. (1993). *A New Approach: Average Shortfall* J.P. Morgan, Fixed Income Research. (44–71) pp. 325–399.
- [11] Rockafellar, R.T. & Uryasev, S. (2000). Optimization of conditional value-at-risk, *Journal of Risk* **2**(3), 21–42.
- [12] Rockafellar, R.T. & Uryasev, S. (2001). Conditional value-at-risk for general loss distributions, *Journal of Banking and Finance* **26**, 1443–1471.

Related Articles

Convex Risk Measures; Spectral Measures of Risk; Value-at-Risk.

CARLO ACERBI

Economic Capital Allocation

In a bank, risk is usually quantified in terms of risk capital (or economic capital). The reason for the close connection between risk and capital is the fact that the main purpose of the bank's capital is to provide protection against extreme losses, that is, capital invested in safe and liquid assets should ensure solvency of the bank even in adverse economic scenarios. Hence, the capital requirements of a bank are determined by its risk profile. The actual calculation is based on risk measures like Value-at-Risk (VaR) or expected shortfall; see [12] or **Convex Risk Measures** for an exposition of the theory of risk measures.

Techniques for measuring risk are not only required for capital management but are also prerequisites for profitability analysis and portfolio optimization. The development of the theoretical relationship between risk and expected return is built on two economic theories: portfolio theory and capital market theory [18, 19, 25]. Portfolio theory deals with the selection of portfolios that maximize the expected returns consistent with individually acceptable levels of risk, whereas capital market theory focuses on the relationship between security returns and risk. These theories also provide a natural framework for measuring profitability. The profitability analysis is commonly carried out by expressing the risk–return relationship as simple rational functions of risk and return components. The two basic variants of these so-called risk adjusted ratios are known as *RORAC* or *RAROC*, respectively; see [20] for details.

Portfolio optimization and the profitability analysis of individual business units, portfolios, and transactions require the quantification of risk at the respective levels of the bank's hierarchy. In line with portfolio theory, the main objective of this analysis is to measure the contribution of each unit to the total risk of the bank and to allocate risk capital accordingly. For a mathematical formalization of the allocation problem, we assume that a risk measure ρ has been fixed and that X is a portfolio that consists of subportfolios $X_1, \dots, X_m$. The objective is to distribute the risk capital $k := \rho(X)$ of the portfolio $X = X_1 + \dots + X_m$ to its subportfolios, that is, to

compute risk contributions $k_1, \dots, k_m$ of $X_1, \dots, X_m$ with $k = k_1 + \dots + k_m$.

In recent years, theoretical and practical aspects of different allocation schemes have been analyzed in several articles; see for instance [5–8, 11, 14, 22, 23, 26–28]. An allocation scheme proposed by several authors is the allocation by the gradient or Euler principle: the capital allocated to the subportfolio X_i of X is the derivative of the associated risk measure ρ at X in the direction of X_i (see formula (9) for a precise formalization). In the following, we review a simple axiomatization of capital allocation proposed in [15]. The main axioms are the property that the entire risk capital of a portfolio is allocated to its subportfolios and a diversification property that is closely linked to the subadditivity of the underlying risk measure. It turns out that in this framework the Euler principle is an immediate consequence of the proposed axioms.

Axioms for Capital Allocation

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, L^0 the space of all equivalence classes of real-valued random variables on Ω and V a subspace of the vector space L^0 , which represents portfolios. More precisely, we will identify each portfolio X with its loss function, that is, X is an element of V and $X(\omega)$ specifies the loss of X at a future date in state $\omega \in \Omega$. We assume that a function $\rho : V \rightarrow \mathbb{R}$ has been defined. For each $X \in V$, $\rho(X)$ specifies the risk capital associated with portfolio X . At this point, we do not require that the risk measure ρ has specific properties.

The axiomatization in [15] is based on the assumption that the capital allocated to subportfolio X_i only depends on X_i and X but not on the decomposition of the remainder $X - X_i = \sum_{j \neq i} X_j$ of the portfolio. Hence, a capital allocation can be considered as a function Λ from $V \times V$ to $\mathbb{R}$. Its interpretation is that $\Lambda(X, Y)$ represents the capital allocated to the portfolio X considered as a subportfolio of portfolio Y .

Definition 1 (Axiomatization of Capital Allocation). *A function $\Lambda : V \times V \rightarrow \mathbb{R}$ is called a capital allocation with associated risk measure ρ if it satisfies the condition $\Lambda(X, X) = \rho(X)$ for all $X \in V$, that is, if the capital allocated to X (considered as stand-alone portfolio) is the risk capital $\rho(X)$ of X . The following requirements for Λ are proposed. The capital allocation Λ is called*

$$\begin{array}{ll} \text{linear :} & \forall a, b \in \mathbb{R}, X, Y, Z \in V \end{array} \quad (1)$$

$$\begin{array}{ll} \Lambda(aX + bY, Z) = a\Lambda(X, Z) + b\Lambda(Y, Z) & \\ \text{diversifying :} & \forall X, Y \in V \end{array} \quad (2)$$

$$\begin{array}{ll} \Lambda(X, Y) \leq \Lambda(X, X) & \\ \text{continuous at } Y \in V : & \forall X \in V \end{array} \quad (3)$$

$$\lim_{\epsilon \rightarrow 0} \Lambda(X, Y + \epsilon X) = \Lambda(X, Y)$$

The first two axioms ensure that the risk capital of the portfolio equals the sum of the (contributory) risk capital of its subportfolios and that the capital allocated to a subportfolio X never exceeds the risk capital of X considered as a stand-alone portfolio. If Λ is continuous at $Y \in V$, then small changes to the portfolio Y only have a limited effect on the risk capital of its subportfolios. We refer to [7, 8] for alternative axiomatizations based on game-theoretic concepts.

The above axiom system provides a close link between risk measures and capital allocation rules. First, given a capital allocation Λ the corresponding risk measure ρ is obviously given by the values of Λ on the diagonal, that is, $\rho(X) = \Lambda(X, X)$. Conversely, assume that a risk measure ρ is

a capital allocation Λ_ρ by

$$\Lambda_\rho(X, Y) := h_Y^\rho(X) \quad (8)$$

The set H_ρ can be interpreted as a collection of (generalized) scenarios: the capital allocated to a subportfolio X of portfolio Y is simply the loss of X under scenario h_Y^ρ .

The following theorem (Theorem 4.2 in [15]) states the equivalence between positively homogeneous, subadditive (but not necessarily coherent) risk measures and linear, diversifying capital allocations.

Theorem 1 (Existence of Capital Allocations). *Let $\rho: V \rightarrow \mathbb{R}$.*

$$\begin{array}{ll} \text{positively homogeneous :} & \forall a \geq 0, X \in V \end{array} \quad (4)$$

$$\begin{array}{ll} \rho(aX) = a\rho(X) & \\ \text{subadditive :} & \forall X, Y \in V \end{array} \quad (5)$$

$$\rho(X + Y) \leq \rho(X) + \rho(Y)$$

Then, the corresponding capital allocation Λ_ρ can be constructed as follows: let V^* be the set of real linear functionals on V and for a given risk measure ρ consider the subset

$$H_\rho := \{h \in V^* \mid h(X) \leq \rho(X) \text{ for all } X \in V\} \quad (6)$$

It is an easy consequence of the Hahn–Banach theorem (see, for instance, Theorem II.3.10 in [10]) that for a positively homogeneous and subadditive risk measure ρ

$$\rho(X) = \max\{h(X) \mid h \in H_\rho\} \quad (7)$$

for all $X \in V$. Hence, for every $Y \in V$ there exists an $h_Y^\rho \in H_\rho$ with $h_Y^\rho(Y) = \rho(Y)$. This allows to define

1. *If there exists a linear, diversifying capital allocation Λ with associated risk measure ρ , then ρ is positively homogeneous and subadditive.*
2. *If ρ is positively homogeneous and subadditive, then Λ_ρ is a linear, diversifying capital allocation with associated risk measure ρ .*

If a linear, diversifying capital allocation Λ is moreover continuous at a portfolio $Y \in V$, then it is uniquely determined by the directional derivative of its associated risk measure, as the next theorem (Theorem 4.3 in [15]) shows.

Theorem 2 (Uniqueness of Capital Allocations). *Let ρ be a positively homogeneous and subadditive risk measure and $Y \in V$. Then the following three conditions are equivalent:*

1. Λ_ρ is continuous at Y , that is, for all $X \in V$ $\lim_{\epsilon \rightarrow 0} \Lambda_\rho(X, Y + \epsilon X) = \Lambda_\rho(X, Y)$.
2. The directional derivative

$$\lim_{\epsilon \rightarrow 0} \frac{\rho(Y + \epsilon X) - \rho(Y)}{\epsilon} \quad (9)$$

exists for every $X \in V$.

3. There exists a unique $h \in H_\rho$ with $h(Y) = \rho(Y)$.

If these conditions are satisfied then $\Lambda_\rho(X, Y)$ equals the directional derivative (9) for all $X \in V$, that is, Λ_ρ is given by the Euler principle.

If the equivalent conditions in Theorem 2 are not satisfied for $Y \in V$, then the two-sided directional derivative (9) does not exist for all X and the capital allocation $\Lambda_\rho(\cdot, Y)$ is not uniquely defined. In a recent article, Cherny and Orlov [6] propose two different solutions to this problem: either the replacement of the two-sided directional derivative by a one-sided directional derivative or a modification of the axiomatization in Definition 1 obtained by adjusting the continuity axiom and adding law invariance. They show that for a spectral risk measure [1, 27], (see also **Spectral Measures of Risk**), there exists a unique capital allocation satisfying the proposed axiom system. Capital allocations for spectral risk measures are also investigated in [23]. An axiomatic characterization of capital allocations of coherent risk measures [3] is given in [16].

The Euler principle is an allocation scheme that has been proposed by several authors; see [28] for an overview. Tasche [26] argues that allocation based on the Euler principle provides the right signals for performance measurement. Another justification for the Euler principle is given by Denault [8] using the framework of cooperative game theory. He shows that the Euler principle is the only fair allocation principle for a differentiable coherent risk measure. The existence of the directional derivative (9) is analyzed in several articles, for example, in [4, 11, 13, 27]. Alternative allocation techniques have been explored in the actuarial sciences; see, for example, [9, 21]. A comparison of different combinations of risk measures and allocation methods can be found in [29].

Capital Allocation for Specific Risk Measures

In classical portfolio theory, risk is measured by standard deviations [19]. More precisely, let c be a nonnegative real number and define the risk measure ρ_c^{Std} and the capital allocation Λ_c^{Std} by

$$\rho_c^{\text{Std}}(X) := c \cdot \text{Std}(X) + E(X) \quad (10)$$

$$\Lambda_c^{\text{Std}}(X, Y) := \begin{cases} c \cdot \text{Cov}(X, Y) / \text{Std}(Y) + E(X) & \text{if } \text{Std}(Y) > 0, \\ E(X) & \text{if } \text{Std}(Y) = 0 \end{cases} \quad (11)$$

The risk measure ρ_c^{Std} is positively homogeneous and subadditive but not coherent for $c > 0$. The covariance allocation Λ_c^{Std} is linear and diversifying. It is derived from the risk measure ρ_c^{Std} using the construction principle (8). If $\text{Std}(Y) > 0$ then Λ_c^{Std} is continuous at Y and equals the directional derivative (9) by Theorem 2.

The current standard in the finance industry is to define the risk capital in terms of a quantile of the loss distribution (see **Value-at-Risk**): the Value-at-Risk $\text{VaR}_\alpha(X)$ of the portfolio X at a specified confidence level $\alpha \in (0, 1)$ is defined by

$$\text{VaR}_\alpha(X) := \inf\{l \in \mathbb{R} \mid \mathbb{P}(X \leq l) \geq \alpha\} \quad (12)$$

VaR has an intuitive economic interpretation, that is, it specifies the capital needed to absorb losses with probability α , and has even achieved the high status of being written in industry regulations. However, VaR also has an obvious limitation as a risk measure: in general, it is not subadditive if applied to portfolios with nonelliptic distributions. Theorem 1 implies that in the general case there do not exist linear, diversifying capital allocations for VaR. However, under regularity conditions (see, for example, [26]), the directional derivative (9) exists for VaR_α and equals

$$E(X|Y = \text{VaR}_\alpha(Y)) \quad (13)$$

The most prominent coherent risk measure is expected shortfall (see, for instance, [2, 24], **Expected Shortfall**) given by

$$\text{ES}_\alpha(X) = \frac{1}{1 - \alpha} \int_\alpha^1 \text{VaR}_u(X) \, du \quad (14)$$

Instead of fixing a quantile at a particular confidence level α , expected shortfall averages VaR across the entire tail specified by α . Coherence implies that expected shortfall is positively homogeneous and subadditive. Hence, application of equation (8) yields

$$\begin{aligned} \Lambda_{\alpha}^{\text{ES}}(X, Y) &:= \left(\int X \cdot \mathbf{1}_{\{Y > \text{VaR}_{\alpha}(Y)\}} d\mathbb{P} \right. \\ &\quad \left. + \beta_Y \int X \cdot \mathbf{1}_{\{Y = \text{VaR}_{\alpha}(Y)\}} d\mathbb{P} \right) / (1 - \alpha), \quad \text{with} \\ \beta_Y &:= \frac{\mathbb{P}(Y \leq \text{VaR}_{\alpha}(Y)) - \alpha}{\mathbb{P}(Y = \text{VaR}_{\alpha}(Y))} \\ &\quad \text{if } \mathbb{P}(Y = \text{VaR}_{\alpha}(Y)) > 0 \end{aligned} \quad (15)$$

which is a linear, diversifying capital allocation with respect to ES_{α} . If

$$\begin{aligned} \mathbb{P}(Y > \text{VaR}_{\alpha}(Y)) &= 1 - \alpha \\ \text{or } \mathbb{P}(Y \geq \text{VaR}_{\alpha}(Y)) &= 1 - \alpha \end{aligned} \quad (16)$$

then $\Lambda_{\alpha}^{\text{ES}}$ is continuous at Y and equals the directional derivative (9). In particular, equation (16) holds if $\mathbb{P}(Y = \text{VaR}_{\alpha}(Y)) = 0$; in that case, $\Lambda_{\alpha}^{\text{ES}}(X, Y)$ takes the particularly intuitive form

$$\Lambda_{\alpha}^{\text{ES}}(X, Y) = E(X \mid Y > \text{VaR}_{\alpha}(Y)) \quad (17)$$

The impact of capital allocation techniques on portfolio management has been analyzed in a case study in [17], which compares expected shortfall allocation and covariance allocation in large credit portfolios.

References

- [1] Acerbi, C. (2004). Coherent representation of subjective risk-aversion, in *Risk Measures for the 21st Century*, G. Szego, ed, Wiley, Chichester.
- [2] Acerbi, C. & Tasche, D. (2002). On the coherence of expected shortfall, *Journal of Banking and Finance* **26**, 1487–1503.
- [3] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**(3), 203–228.
- [4] Carlier, G. (2008). Differentiability properties of rank-linear utilities, *Journal of Mathematical Economics* **44**(1), 15–23.
- [5] Cherny, A.S. (2007). Pricing with coherent risk, *Probability Theory and Its Applications* **52**(3), 506–540.
- [6] Cherny, A.S. & Orlov, D.V. (2007). *On Two Approaches to Coherent Risk Contribution*, Preprint, Moscow State University.
- [7] Delbaen, F. (2000). *Coherent Risk Measures*. Lecture notes, Scuola Normale Superiore di Pisa.
- [8] Denault, M. (2001). Coherent allocation of risk capital, *Journal of Risk* **4**(1), 1–34.
- [9] Dhaene, J., Goovaerts, M. & Kaas, R. (2003). Economic capital allocation derived from risk measures, *American Actuarial Journal* **7**(2), 44–59.
- [10] Dunford, N. & Schwartz, J. (1958). *Linear Operators—Part I: General Theory*, Interscience Publishers, New York.
- [11] Fischer, T. (2003). Risk capital allocation by coherent risk measures based on one-sided moments, *Insurance: Mathematics and Economics* **32**(1), 135–146.
- [12] Föllmer, H. & Schied, A. (2004). *Stochastic Finance—An Introduction in Discrete Time*, 2nd Edition, Walter de Gruyter, Berlin.
- [13] Gouriéroux, C., Laurent, J.-P. & Scaillet, O. (2000). Sensitivity analysis of values at risk, *Journal of Empirical Finance* **7**, 225–245.
- [14] Hallerbach, W. (2003). Decomposing portfolio value-at-risk: a general analysis, *Journal of Risk* **5**(2), 1–18.
- [15] Kalkbrener, M. (2005). An axiomatic approach to capital allocation, *Mathematical Finance* **15**(3), 425–437.
- [16] Kalkbrener, M. (2007). *An Axiomatic Characterization of Capital Allocations of Coherent Risk Measures*, Preprint, Deutsche Bank, Frankfurt.
- [17] Kalkbrener, M., Lotter, H. & Overbeck, L. (2004). Sensible and efficient capital allocation for credit portfolios, *Risk* **17**(1), S19–S24.
- [18] Lintner, J. (1965). Security prices, risk and maximal gains from diversification, *Journal of Finance* **20**(4), 587–615.
- [19] Markowitz, H. (1952). Portfolio selection, *Journal of Finance* **7**, 77–91.
- [20] Matten, C. (2000). *Managing Bank Capital: Capital Allocation and Performance Measurement*, 2nd Edition, Wiley, New York.
- [21] Myers, S.C. & Read, J.A. (2001). Capital allocation for insurance companies, *Journal of Risk and Insurance* **68**(4), 545–580.
- [22] Overbeck, L. (2000). Allocation of economic capital in loan portfolios, in *Proceedings Measuring Risk in Complex Stochastic Systems, Lecture Notes in Statistics*, Vol. 147, W. Hardle & G. Stahl, eds, Springer, Berlin, 1999.
- [23] Overbeck, L. (2004). Spectral capital allocation, in *Economic Capital, A Practitioner Guide*, A. Dev, ed, Risk Books, London.
- [24] Rockafellar, R.T. & Uryasev, S. (2002). Conditional value-at-risk for general loss distributions, *Journal of Banking and Finance* **26**, 1443–1471.
- [25] Sharpe, W. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk, *Journal of Finance* **19**(3), 425–442.

- [26] Tasche, D. (1999). *Risk Contributions and Performance Measurement*, Preprint, Munich University of Technology.
- [27] Tasche, D. (2002). Expected shortfall and beyond, *Journal of Banking and Finance* **26**(7), 1519–1533.
- [28] Tasche, D. (2007). *Euler Allocation: Theory and Practice*, Preprint, Fitch Ratings, London.
- [29] Urban, M., Dittrich, J., Klüppelberg, C. & Stölting, R. (2004). Allocation of risk capital to insurance portfolios, *Blätter der DGVFM* **26**, 389–406.

Related Articles

Convex Risk Measures; Economic Capital; Risk-adjusted Return on Capital (RAROC); Value-at-Risk.

MICHAEL KALKBRENER

Spectral Measures of Risk

The term *spectral measures of risk* (SMRs) [1] denotes a subclass of coherent measures of risk (CMRs) [4] (see **Convex Risk Measures**) characterized by the additional properties of law invariance and comonotonic additivity, which make them particularly appropriate for financial risk management applications. In financial mathematics, this class was first defined in [1] and [7], which studied it from different perspectives. It was however soon realized that this class is in close correspondence with so-called distortion measures already introduced in 1996 in actuarial mathematics [9] in a different axiomatic framework from that of CMRs.

The most popular example of SMR is given by conditional Value at Risk/expected shortfall (*ES*) (see **Expected Shortfall**), which is, in fact, the building block of all the class of SMRs as will be seen.

An SMR of a random variable (r.v.) X denoted by $\rho_\phi(X)$ can be defined as *minus the ϕ -weighted average of all possible outcomes of X sorted from the lowest to the highest*, where ϕ is a weight function, called the *risk spectrum* of the SMR, which needs to be positive, normalized, and decreasing.

In financial risk management, the SMR of a portfolio π usually means the SMR $\rho_\phi(r_T^{(\pi)})$ of the return $r_T^{(\pi)} = \pi_T - (1 + R)\pi_0$ between the current date $t = 0$ and a chosen future date $t = T$, R being the risk-free interest. Clearly, $\rho_\phi(r_T^{(\pi)})$ expresses the *ϕ -weighted average of all possible losses of π sorted in decreasing order*.

Strictly speaking, it is only the SMR $\rho_\phi(\pi_T)$ of the “value” (and not of the “return”) of a portfolio that is a true CMR for the same reasons as explained for *ES* in **Expected Shortfall**. In this article, we discuss the mathematical properties of an SMR as a statistic of a generic r.v. X and loosely use the term *coherent* to mean monotonic, translational invariant, positive homogeneous, and subadditive.

Formal Definition of SMR

For the sake of simplicity, we overlook some technical details such as integrability of r.v.’s and continuity of risk measures. For a more detailed exposition, we refer to [6].

SMRs can be defined as the most general convex combinations of all possible *ES*s. Let us consider a probability space with probability measure P and let X be an r.v. A generic SMR $\rho^\mu(X)$ can therefore be uniquely written as

$$\rho^\mu(X) = \int_0^1 ES_\alpha(X) d\mu(\alpha) \quad (1)$$

where $d\mu$ is a measure on the interval $[0, 1]$ and, *vice versa*, any such $d\mu$ defines an SMR.

For the sake of financial intuition, this expression can be recast into a mixture of quantiles of the variable X . Let $F_X^{-1} : (0, 1) \rightarrow \mathbb{R}$ denote an inverse function of the probability distribution function $F_X(x)$ of X . Choose, for instance, the lower p -quantile function $F_X^{-1}(p) := q_p(X) = \inf\{x \in \mathbb{R} | F_X(x) \geq p\}$. One can show that $\rho^\mu = \rho_\phi$ with

$$\rho_\phi(X) = - \int_0^1 \phi(p) F_X^{-1}(p) dp \quad (2)$$

with ϕ defined by

$$\begin{aligned} \phi(p) &= \mu(\{0\}) \delta_0(p) & p = 0 \\ \phi(p) &= \int_{[p,1]} d\mu(\alpha) \alpha^{-1} & p \in (0, 1] \end{aligned} \quad (3)$$

Alternatively, one can adopt equation (2) as a definition of SMR. The *risk spectrum* ϕ must be of the form $\phi = c\delta_0(p) + \tilde{\phi}$ with $c \in [0, 1]$ and $\tilde{\phi} : [0, 1] \rightarrow \mathbb{R}$ and satisfy the conditions

$$\begin{aligned} \phi(p) &\geq 0 & 0 < p \leq 1 \\ \int_{[0,1]} \phi(p) dp &= 1 \\ \phi(p_1) &\geq \phi(p_2) & 0 < p_1 < p_2 \leq 1 \end{aligned} \quad (4)$$

which can be shown to be necessary and sufficient for ρ_ϕ in equation (2) to be a coherent measure of risk [1]. We see therefore that equation (2) can indeed be read as *minus the ϕ -weighted average of all the outcomes of X sorted in decreasing order*, where ϕ is a positive, normalized, and decreasing function. In general, ϕ also admits a Dirac delta term $c\delta_0$ in the origin, but the singular case $c \neq 0$ is of no practical interest.

Expression (2) together with conditions (4) tells us that a mixture of quantiles is coherent only when they are weighted in such a way that worse events

are given higher weights. This is an important lesson we learn when dealing with quantile-based measures of risk. We immediately also learn that for such a measure to be coherent it needs to weight all the left tail without neglecting any portion of its extreme values.

If we consider α -Value at Risk ($V@R_\alpha$) (see **Value-at-Risk**), which is not a CMR (and hence not an SMR), we see that it can also be expressed through formula (2), but its risk spectrum turns out to be a Dirac delta $\phi^{V@R_\alpha}(p) = \delta_\alpha(p)$, which is clearly not decreasing on $[0, 1]$. Here, we see that the reason why $V@R$ is not a CMR lies exactly in the fact that it completely neglects the worst cases at quantiles smaller than α .

Taking ES_α (see **Expected Shortfall**) as an example of SMR, it is easy to show that its risk spectrum is given by $\phi^{ES_\alpha}(p) = \alpha^{-1} \mathbf{1}_{p \leq \alpha}$, namely, it is flat on $[0, \alpha]$ and vanishing elsewhere. This is a perfectly coherent choice corresponding to an equally weighted average of all the cases beyond the α quantile only.

The function ϕ is also termed a *risk-aversion function* of the SMR [1] because it expresses different aversions encoded in ρ_ϕ to different quantiles of the profit–loss distribution.

The relationship with the class of “distortion measures” defined in [9] is established by writing

$$\rho_\phi(X) = D_g(-X) \quad (5)$$

where the distortion measure D_g is defined as the *Choquet integral* (see, for instance, [6])

$$\begin{aligned} D_g(Y) &:= \int Y \, d(g \circ P) = \int_0^\infty g(1 - F_Y(y)) \, dy \\ &\quad + \int_{-\infty}^0 [g(1 - F_Y(y)) - 1] \, dy \end{aligned} \quad (6)$$

and the function $g : [0, 1] \rightarrow \mathbb{R}$ is a “concave distortion” of the probability measure P . The function g is related to ϕ by

$$\begin{aligned} g(p) &= 0 & p &= 0 \\ g(p) &= c + \int_{(0,p]} \phi(q) \, dq & p &\in (0, 1] \end{aligned} \quad (7)$$

The expression $D_g(-X)$ is another possible definition of SMR, provided that $g : [0, 1] \rightarrow \mathbb{R}$ is increasing and concave and satisfies $g(0) = 0$ and $g(1) = 1$.

In [9], the definition of “distortion measures” is slightly less general, as it is given on positive-valued loss variables Y only, which reflects the actuarial context. Notice also the different convention between the actuarial literature (where losses are modeled as positive variables) and the financial literature (where profits are modeled as positive variables). Apart from these technical details, Wang’s work [9] can be considered a precursor of the development of SMRs in all respects.

As an example of SMR, consider the one-parameter family class of exponential risk spectra given by

$$\phi_\gamma(p) = \frac{e^{-p/\gamma}}{\gamma(1 - e^{-1/\gamma})} \quad (8)$$

It is easy to show that this expression satisfies equation (4) and therefore defines an SMR for any $\gamma > 0$. This risk spectrum is nonvanishing on all the quantiles range $p \in [0, 1]$. Smaller values of γ provide SMRs that are more focused on the extreme left tail of the distribution.

The corresponding concave distortion is easily obtained from equation (7):

$$g(p) = \frac{1 - e^{-p/\gamma}}{1 - e^{-1/\gamma}} \quad (9)$$

Properties of SMRs and Their Characterization

Any SMR ρ_ϕ can be shown to be coherent [1, 7] (see **Convex Risk Measures**). In other words, it satisfies the following properties [4]:

(i) *monotonicity*

$$\rho_\phi(X) \geq \rho_\phi(Y) \quad \text{if } X \leq Y \text{ almost surely.} \quad (10)$$

(ii) *translational invariance*

$$\rho_\phi(X + c) = \rho_\phi(X) - c \quad \text{if } c \in \mathbb{R} \quad (11)$$

(iii) *positive homogeneity*

$$\rho_\phi(\lambda X) = \lambda \rho_\phi(X) \quad \text{if } \lambda \geq 0 \quad (12)$$

(iv) *subadditivity*

$$\rho_\phi(X + Y) \leq \rho_\phi(X) + \rho_\phi(Y) \quad (13)$$

These conditions represent the cornerstone of axiomatic financial risk theory and strong arguments have been given to show that any measure of risk violating them turns out to be inappropriate for capital allocation purposes. In particular, subadditivity embodies the risk diversification principle and prevents the occurrence of paradoxical regulatory arbitrages. Adopting a nonsubadditive measure of risk to compute the regulatory capital of a portfolio, in fact, one can, in principle, obtain a lower value by just splitting the portfolio into two parts and computing the regulatory capital on each part [3].

Conditions (iii) and (iv) imply, in particular, convexity

(v) *convexity*

$$\begin{aligned} \rho_\phi(\mu X + (1 - \mu) Y) \\ \leq \mu \rho_\phi(X) + (1 - \mu) \rho_\phi(Y) \quad \text{if } \mu \in [0, 1] \end{aligned} \quad (14)$$

which plays a fundamental role in asset allocation as it entails the convexity of portfolio optimization problems (see, for instance, [2, 8]).

It is particularly important to identify the additional properties that characterize SMRs as a subclass of CMRs. It can be shown that SMRs are, in fact, all CMRs that are also *law invariant* and *comonotonic additive* [6, 7].

A measure of risk is said to be *law invariant* if it only depends on the probability distribution function of its argument, or, in other words, if it takes the same value on any two r.v.'s that are identically distributed.

(vi) *law invariance*

$$\rho_\phi(X) = \rho_\phi(Y) \quad \text{if } X \sim Y \quad (15)$$

This property may appear puzzling at first sight because every statistic obviously satisfies it. It has, however, to be kept in mind that not all CMRs are “statistics” because they are defined without necessarily specifying a probability space, but by specifying only a set of events [4] (see **Convex Risk Measures**). Hence, law invariance is a stringent condition on CMRs, and, in fact, it is the condition characterizing all CMRs that are also statistics and can therefore *be estimated from data* [2]. This condition is therefore often considered as inevitable for practical risk

management applications, where portfolios are confronted with a model of the “real-world” probability distribution.

The further characterizing property of SMRs within CMRs is comonotonic additivity

(vii) *comonotonic additivity*

$$\begin{aligned} \rho_\phi(X + Y) &= \rho_\phi(X) + \rho_\phi(Y) \quad \text{if} \\ X, Y &\text{ comonotonic} \end{aligned} \quad (16)$$

This means that for SMRs the subadditivity diversification benefit of adding two portfolios X and Y vanishes in the limiting case when they happen to be comonotonic, which means they are perfectly dependent on each other (for a definition of comonotonicity, see, for instance, [5]). This property is also crucial, particularly in the context of capital adequacy in association with subadditivity, because, if it does not hold, another kind of regulatory arbitrage opportunity appears. By exploiting subadditivity, one could lower the capital requirement of two distinct portfolios by just putting them together, even in the comonotonic case when they provide no mutual hedge at all. It seems, in other words, that the diversification principle is correctly represented by conditions (iv) and (vii) together and not by subadditivity alone.

Law invariance and comonotonic additivity, therefore, enrich the set of coherency axioms (i)–(iv). Owing to these additional properties, SMRs are considered to be a particularly interesting subset of CMRs for concrete risk management applications (see, for instance, [3] for a discussion).

Further properties of SMRs are monotonicity with respect to both first- and second-order stochastic dominance [6], which are important preference criteria among portfolios in utility theory. We recall that an r.v. X is said to dominate Y according to first-order stochastic dominance (notation: $X \succ^{1st} Y$) if $F_X(z) \leq F_Y(z)$ for all $z \in \mathbb{R}$, while X is said to dominate Y according to second-order stochastic dominance (notation: $X \succ^{2nd} Y$) if $\int_{-\infty}^z [F_Y(t) - F_X(t)] dt \geq 0$ for all $z \in \mathbb{R}$. First-order implies second-order stochastic dominance. SMRs turn out to be compatible with the notions of preference among portfolios induced by stochastic dominance.

(viii) *monotonicity with respect to first- and second-order stochastic dominance*

$$\rho_\phi(X) \leq \rho_\phi(Y) \quad \text{if} \quad X \stackrel{1^{st}, 2^{nd}}{\succ} Y \quad (17)$$

This means that an SMR always detects higher risk on dominated portfolios.

It is important to stress that conditions (vi)–(viii) do not hold, in general, for CMRs.

Estimation of SMRs

Thanks to law invariance, SMRs can be estimated from data. This is fundamental both when an empirical distribution for the variable X is given and when a closed-form solution for $\rho_\phi(X)$ is not known, but one can sample from the distribution of X to perform a Monte Carlo simulation.

Consider an SMR ρ_ϕ with a given risk spectrum ϕ . Given a sample of N i.i.d. outcomes x_i , of an r.v. X , a consistent estimator of $\rho_\phi(X)$ can be obtained by [1, 2]

$$\hat{\rho}_\phi(X) = - \sum_{i=0}^N \phi_i x_{i:N} \quad (18)$$

where $x_{i:N}$ denotes the ordered statistics, that is, the vector $\{x_{i:N}\}_{i=1,\dots,N}$ of values $x_{1:N} \leq \dots \leq x_{N:N}$ obtained sorting the vector $\{x_i\}_{i=1,\dots,N}$ and the weights ϕ_i are given by

$$\phi_i = \int_{(i-1)/N}^{i/N} \tilde{\phi}(p) dp + \delta_{i,1} c \quad (19)$$

It can be shown that $\hat{\rho}_\phi(X)$ converges to $\rho_\phi(X)$ for $N \rightarrow \infty$. $\hat{\rho}_\phi(X)$ is, in fact, nothing but the SMR ρ_ϕ of the empirical distribution of the sample and as such it is coherent by construction for any finite N .

As an example, consider the family of SMRs defined by equation (8). It can be seen immediately that the spectrum is nonsingular ($c = 0$) and that the weights of the estimator (18) are given by

$$\phi_i = \frac{e^{1/N\gamma} - 1}{1 - e^{-1/\gamma}} e^{-i/N\gamma} \quad (20)$$

References

- [1] Acerbi, C. (2002). Spectral measures of risk: a coherent representation of subjective risk aversion, *Journal of Banking and Finance* **26**, 1505–11518.
- [2] Acerbi, C. (2004). Coherent representations of subjective risk aversion, in *Risk Measures for the 21st Century*, G. Szego, ed, John Wiley & Sons, pp. 147–207.
- [3] Acerbi, C. (2007). Coherent measures of risk in everyday market practice, *Quantitative Finance* **7**(4), 359–3364.
- [4] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**(3), 203–228.
- [5] Embrechts, P., Frey, R. & McNeil, A.J. (2005). *Quantitative Risk Management*, Princeton University Press.
- [6] Föllmer, H. & Schied, A. (2004). *Stochastic Finance*, 2nd Edition, de Gruyter.
- [7] Kusuoka, S. (2001). On law invariant coherent risk measures, *Advances in Mathematical Economics* **3**, 83–95.
- [8] Rockafellar, R.T. & Uryasev, S. (2001). Conditional value-at-risk for general loss distributions, *Journal of Banking and Finance* **26**, 1443–11471.
- [9] Wang, S. (1996). Premium calculation by transforming the layer premium density, *Astin Bulletin* **26**, 71–92.

Related Articles

Convex Risk Measures; Expected Shortfall; Value-at-Risk.

CARLO ACERBI

Market Risk

Market risk models were developed in the 1990s following the growth in proprietary trading in banking institutions. As an example, the chairman of Morgan asked his staff: “At the close of each business day, tell me what the market risks are across all businesses and locations.” This led to quantitative risk measurements systems, such as Value at Risk (VAR), which aggregate risk at the top level of the institution, based on the most current positions.

The impetus for using VAR models came not only from the industry but also from financial regulators. Notably, the Basel Committee on Banking Supervision [2] allowed commercial banks to use their internal VAR models as the basis for their market risk charge, starting in 1998.

VAR summarizes the worst loss over a target horizon that will not be exceeded with a given level of confidence. For instance, in 1994, Morgan’s daily VAR was approximately \$15 million at the 95% level of confidence over a horizon of one day. In other words, the bank should not expect to lose more than \$15 million in 95 days out of 100. Conversely, it should expect to lose \$15 million or more in the remaining 5 days out of 100. This single number describes the risk of the entire trading portfolio of the bank. Until then, risk was measured and managed at the level of a desk or business unit.

Portfolio Distributions

VAR took hold as a universally accepted measure of risk because of its simplicity. It summarizes the downside risk in one measure, easy to understand, which is expressed in terms of a loss in the relevant currency. Because it is associated with a confidence level, it is sometimes called a *statistical measure of risk*.

The most useful aspect of VAR system, however, is the process that led to this number. To compute VAR, the institution has to set up systems to collect position and market data for the entire firm. From this information, its risk manager can build the entire probability density function of profits and losses over the risk horizon. As shown in Figure 1, VAR is just a quantile of the distribution. Specifying a confidence

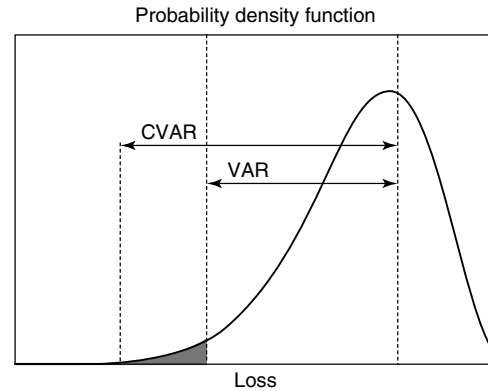


Figure 1 Summarizing the distribution of losses

level c , this is the cutoff point such that the area to its left is precisely $1 - c$.

If the distribution is normal, or more generally elliptical, the quantile can be derived from the standard deviation σ and a multiplier that is distribution-specific. For a normal distribution, for instance, the standard normal deviate that corresponds to the 99% confidence level is $\alpha = 2.33$. Therefore, the 99% VAR can be obtained by multiplying σ by 2.33.

The entire distribution, however, is of interest. We should examine whether the distribution is symmetric or skewed. Skewness to the left implies a greater probability of large losses than gains. Another question is, what is the expected loss when VAR is exceeded? This is known as the *conditional VAR* (CVAR). This must be greater than VAR by construction. In most cases, CVAR is only slightly greater than VAR. In other cases, however, it can be much larger. In such situations, VAR limits may not be effective, as rare losses could bankrupt the institution.

As a risk measure, CVAR has other useful properties. Ideally, risk measures should obey several properties to be “coherent”, as suggested by Artzner *et al.* [1]. If a portfolio has lower returns for all states of the world, its risk must be greater (monotonicity); adding cash to the portfolio should reduce its risk by this amount (translation invariance); scaling a portfolio should simply scale its risk (homogeneity); and merging portfolios cannot increase risk (subadditivity). VAR does not obey the last property, unlike CVAR. Hence, merging two portfolios could sometime produce a higher VAR than the sum of the two separate VAR measures. When distributions are

elliptical, however, VAR measures are coherent (*see Convex Risk Measures*).

The VAR framework can also be used to uncover progressively larger but rarer losses, by increasing the confidence level. However, the empirical distribution becomes more irregular in the extreme tails due to the scarcity of data. To some extent, this problem can be alleviated by extreme value theory (EVT), which provides a theoretically sound analytical density function for the smoothing of the tails. Even so, historical data have severe limitations in some cases.

For instance, a trader may have a position on a foreign currency that is fixed against the dollar and for which there is no history of devaluations. Because the recent history shows no volatility, a VAR measure based on this would mistakenly indicate that the position is riskless, which may not be the case. In the case of corporate credits, there is usually no history of defaults by the creditors in the portfolio. A typical solution consists of assessing default probability from a history of similar creditors, perhaps within the same credit rating group. Even so, credit ratings may be flawed, as we have seen in the case of structured products that received very high credit ratings based on a recent history of home price appreciation, but led to massive defaults as soon as home prices turned lower in 2007.

This is why VAR systems must be supplemented by stress-tests scenarios. The risk manager needs to devise scenarios derived from history, perhaps from

a very long time ago, and from prospective events. Stress tests can be used to estimate the effects of these scenarios on the portfolio. Such measures of risk are nonstatistical, yet are essential to evaluate the risk profile of the portfolio.

Structure of Risk Measurement Systems

The potential for losses can be attributed to two sources. On the one hand are the exposures, derived from active positions taken by the trader or portfolio manager. On the other hand are the movements in the risk factors, which are outside their control. This dichotomy is reflected in the structure of risk management systems, which is described in Figure 2.

The left-hand side describes the portfolio positions, which takes as input trades from the front office. The right-hand side describes the risk factors, which has as input datafeeds with current market prices. Positions and risk factor distributions are brought together in the risk engine, which generates a distribution of portfolio values that can be summarized, for instance, by its VAR.

Different VAR methods make different assumptions for the modeling of positions and risk factors. Because modern risk measurement involves large-scale aggregation, positions are routinely simplified or mapped to the risk factors. Positions can be replaced by their linear exposures to the risk factors,

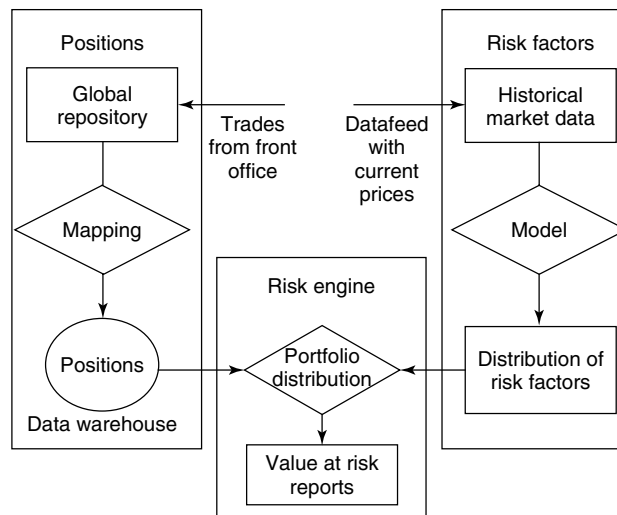


Figure 2 Structure of risk measurement systems

by the quadratic exposures, or using full repricing. The distributions of risk factors can be modeled using a normal distribution, or the historical data, or using Monte Carlo simulations (MCSs). For more details on implementation, see [6, 7].

Local Versus Full Valuation

Conceptually, risk measurement methods can be separated into local valuation and full valuation methods, as shown in Figure 3. The left branch describes local valuation methods, also known as *analytical methods*, which provide closed-form solutions to risk measures. These include linear models and nonlinear models. Linear models are based on the covariance matrix approach. This matrix can be simplified using factor models, or even a diagonal model, which is a one-factor model. Nonlinear models take into account the first and second partial derivatives. The latter are called *gamma* or *convexity* (see **Bond**). Next, the right branch describes full valuation methods, which include historical or MCSs.

To illustrate different approaches, consider a simple instrument, a bond whose value depends on its yield to maturity, which can be taken as the single driving risk factor. The potential for loss depends on the exposure of the bond value to the yield and on potential movements in this yield dy . The linear exposure involves modified duration D^* and dollar duration D^*P . We can write

$$(dP) = (-D^*P) \times (dy) \quad (1)$$

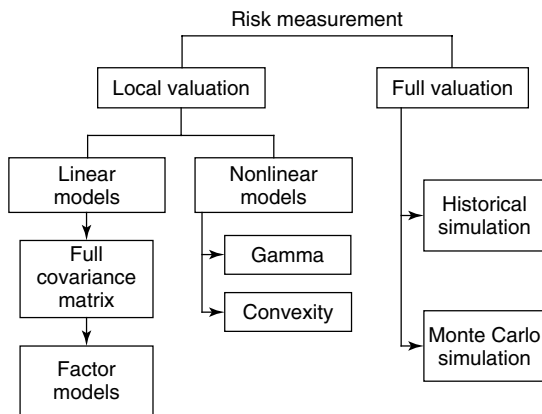


Figure 3 Comparison of risk measurement methods

Hence, the worst loss in the value of the bond at a specified confidence level can be derived from the worst movements in the yield at the same confidence level:

$$(\text{Worst } dP) = (-D^*P) \times (\text{Worst } dy) \quad (2)$$

More generally, the linear model implies a direct link between the bond's VAR and that of the risk factor. Because VAR is reported as a positive number, we take absolute values and have

$$\text{VAR}(dP) = |D^*P| \times \text{VAR}(dy) \quad (3)$$

As an example, take a position in a bond worth \$100 million. We wish to assess VAR over a horizon of one month at the 99% confidence level. Modified duration D^* is four years. From the distribution of yield changes, we observe that the worst increase in yields at the same confidence level and over the same horizon is 0.9%. Putting all of this together, the bond's VAR is $(4 \times \$100) \times 0.9\% = \3.6 million. Note that we priced the bond only once, at the initial yield, which is why this is called a *local valuation method*.

This VAR measure is an approximation only because the bond price is not a simple linear function of the yield. For more precision, we could revalue the bond at the worst yield. Defining y_0 as the initial yield, this gives

$$(\text{Worst } dP) = P[y_0 + (\text{Worst } dy)] - P[y_0] \quad (4)$$

We call this approach full valuation because it requires repricing the asset at the new yield. This is more precise but, unfortunately, is more complex than a simple, linear valuation method, especially if the instrument requires valuation through lengthy simulations. More generally, this method requires full revaluation over the entire risk space, which involves a very large number of data points.

Ideally, we would like to keep the simplicity of the local valuation method while accounting for nonlinearities in the payoffs patterns. We can expand the price-yield function using a Taylor approximation:

$$\begin{aligned} dP &\approx \frac{\partial P}{\partial y} dy + \left(\frac{1}{2}\right) \frac{\partial^2 P}{\partial y^2} (dy)^2 \\ &= (-D^*P) dy + \left(\frac{1}{2}\right) CP(dy)^2 \end{aligned} \quad (5)$$

where the second-order term involves a coefficient called *convexity* C . Here, the valuation is still local because we only value the bond once, at the original point. The first and second derivatives are also evaluated at the local point.

Because the price is a monotonic function of the underlying yield, we can use the Taylor expansion to find the worst down move in the bond price from the worst move in the yield. This leads to a simple adjustment for VAR:

$$\text{VAR}(dP) = |D^*P| \times \text{VAR}(dy) - \left(\frac{1}{2}\right) (CP) \text{VAR}(dy)^2 \quad (6)$$

More generally, this method can be applied to derivatives, whose value depends on the underlying asset as a function $f(S)$. The Taylor approximation is

$$df \approx \frac{\partial f}{\partial S} dS + \left(\frac{1}{2}\right) \frac{\partial^2 f}{\partial S^2} dS^2 = \Delta dS + \left(\frac{1}{2}\right) \Gamma dS^2 \quad (7)$$

where delta, Δ , is the first derivative and gamma, Γ , is now the second derivative, like convexity.

For a long call option position, for example, the worst value is achieved when the underlying price moves down by $\text{VAR}(dS)$. With $\Delta > 0$ and $\Gamma > 0$, the VAR for the derivative is now

$$\text{VAR}(df) = |\Delta| \times \text{VAR}(dS) - \left(\frac{1}{2}\right) \Gamma \times \text{VAR}(dS)^2 \quad (8)$$

This method is called *delta-gamma* because it provides an analytical, second-order correction to the delta-normal VAR. This explains why long positions in options, with positive gamma, have lower risk than with a linear model. This type of analysis can be used to understand how the option characteristics affect its VAR. Unfortunately, this analysis is most useful when there is one only risk factor. With multiple risk factors, the second-order corrections become unwieldy. As a result, this analysis is not applicable to large portfolios.

Mapping

Whichever VAR method is used, the first step consists of mapping. Mapping is the process by which the current values of the portfolio positions are replaced by

global exposures on the risk factors. Mapping arises because of the fundamental nature of VAR, which is portfolio measurement at the highest level. As a result, this is usually a very large-scale aggregation problem. It would be too complex and time consuming to model all positions individually as risk factors. Furthermore, this is unnecessary because many positions are driven by the same set of risk factors. Once a portfolio has been mapped on the risk factors, any of the three VAR methods can be used to build the distribution of profits and losses.

Typical portfolios consist of a large number of instruments, say M . The risk manager must then decide on a small number of risk factors that effectively spans the risk space for the portfolio strategy. Define the number of risk factors as N . In the first step, each position is decomposed into a list of N exposures. In the second step, the exposures are aggregated across instruments. This yields global exposures x_i on each of the risk factors. The final step then consists of combining the exposures with the distribution of the risk factors using one of the three VAR methods.

Consider, for example, a portfolio consisting of many government bonds. The value of each bond represents the sum of the present values of cash flows at different point in times. So, in theory, we could have a number of risk factors that represent the number of days until the longest maturity, typically 30 years. This would involve too many factors. Instead, we could reduce the number of risk factors to three maturities only, for example. Figure 4 illustrates the mapping process with six bonds, which are replaced by three sets of exposures. These are summed across the portfolio, which is then prepared for the next step.

The choice of the number of risk factors must be done taking into account the trading strategies. For example, most fixed-income mutual funds follow a simple, long-only strategy, in which case one factor may be sufficient. The exposure to this factor is then the dollar duration. The portfolio can then be described by its total duration. This choice, however, still leaves out exposures to changes in the shape of the yield curve and would be totally inappropriate for a leveraged trading portfolio that can take long and short positions. If this portfolio is duration-neutral, the risk model would erroneously indicate that there is no risk. More factors are needed for this second type of portfolio. A third type of portfolio could also

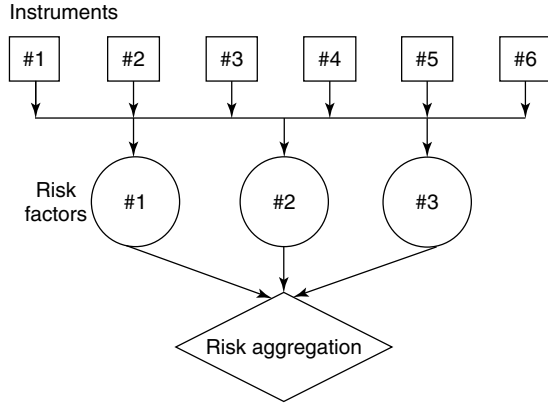


Figure 4 Mapping process

contain options, in that case additional factors such as implied volatilities are needed. In general, more complex strategies require more risk factors. Next, we turn to the different VAR methods.

Delta-normal Method

The delta-normal method is the simplest VAR approach. This method, which is sometimes called *variance-covariance method*, assumes that the portfolio exposures are linear in the risk factors and that the risk factors are jointly normally distributed. As such, it is a local valuation method. The joint normal assumption is very useful, because a linear combination of normal variables is itself normally distributed. As a result, VAR can be expressed in analytical form, directly from the portfolio variance:

$$\sigma^2(R_{p,t+1}) = x_t' \Sigma_{t+1} x_t \quad (9)$$

where Σ_{t+1} is the forecast of the covariance matrix over the horizon, x_t represents the dollar exposures on the risk factors, obtained via mapping and current as of t , and σ^2 is the forecast variance of the portfolio dollar return $R_{p,t+1}$. If needed, the covariance matrix can be simplified to factor models. VAR is directly obtained from the dollar volatility and the standard normal deviate α that corresponds to the confidence level c :

$$\text{VAR}^{\text{DN}} = \alpha \sigma(R_{p,t+1}) \quad (10)$$

More generally, this method can be adapted to portfolio distributions that differ from the normal, by

changing α . VAR can change over time as a result of changes in the positions x_t or in the forecast of the covariance matrix Σ_{t+1} . In the RiskMetrics approach developed by Morgan [8], variances and covariances are forecast using an exponentially weighted moving average (EWMA) model. This is a particular case of the class of generalized autoregressive conditional heteroskedasticity (GARCH) model developed by Engle [3]. Such models are widely employed for predicting time variation in risk because they are parsimonious and fit the data rather well.

Figure 5 shows an application to daily volatility forecasts of the S&P 500 stock index. The graph shows wide swings in the volatility over the short term. In addition, we observe longer term cycles, with elevated volatility from 1999 to 2002, followed by a much quieter period until 2007.

The EWMA model generates forecasts of the covariance between two risk factors i and j as follows. As of today, we observe the previous forecast $\sigma_{ij,t}$ and innovations in the two risk factors R_i and R_j , taken as deviation from their respective means. The forecast for tomorrow is then constructed as

$$\sigma_{ij,t+1} = \lambda \sigma_{ij,t} + (1 - \lambda)(R_{i,t} \times R_{j,t}) \quad (11)$$

This also applies to forecasts of the variance, in that case the two risk factors are the same. This model builds the covariance forecast in recursive form, which not only ensures some persistence in the forecast but also accounts for recent shocks. The relative weight is given by the decay factor λ , which is between 0 and 1. RiskMetrics selected $\lambda = 0.94$ for daily data and $\lambda = 0.97$ for monthly data. Although in theory, this choice could differ across the series, in practice using a common decay factor creates better properties for the covariance matrix.

The model can be also expressed in terms of all previous innovations. Taking the simple case of the variance, this gives

$$\sigma_{t+1}^2 = (1 - \lambda)(R_t^2 + \lambda R_{t-1}^2 + \lambda^2 R_{t-2}^2 + \dots) \quad (12)$$

which shows that the variance forecast is a weighted average of previous squared innovations, with exponentially declining weights on older observations. With $\lambda = 0.94$, the weight on the latest observation is 6%. The weight on the previous observation is $0.94 \times 6\%$, and so on. More than half of the weights is placed on the last 12 observations.

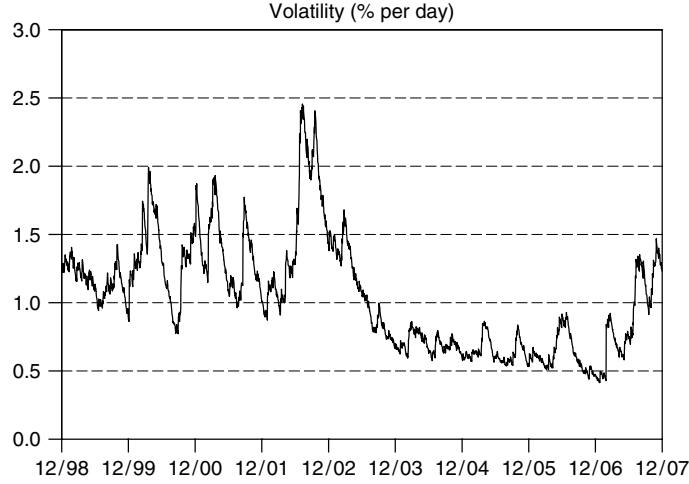


Figure 5 Forecast of daily volatility for S&P 500 Index using EWMA model

In summary, the main benefit of the delta-normal approach is its appealing simplicity. It provides a closed-form solution that is quick to compute. This analytical solution also lends itself well to a decomposition of VAR into its main risk drivers. This provides useful economic intuition.

This method, however, has drawbacks. It cannot account for nonlinear effects. Therefore, some of the risk will be missed for options or for instruments with imbedded options, such as mortgages. In addition, this method may also underestimate the occurrence of large losses because of its reliance on a normal distribution. In practice, most financial series have heavier tails than the normal distribution.

Historical Simulation Method

The historical simulation (HS) method applies current exposures to a time series of recent historical changes in the risk factors, for example, over the last 250 days. Define the current time as t ; we observe data from 1 to t . The current portfolio value is P_t , which is a function of the current N risk factors $f_{i,t}$:

$$P_t = P[f_{1,t}, f_{2,t}, \dots, f_{N,t}] \quad (13)$$

We sample the factor movements from the historical distribution, without replacement. The factor movements are drawn jointly for each day, which preserves the historical pattern of correlations. Call

Δf_i^k the hypothetical movement in risk factor i for the simulation numbered k . These are drawn from the historical distribution:

$$\Delta f_i^k \sim \{\Delta f_{i,1}, \Delta f_{i,2}, \dots, \Delta f_{i,t}\} \quad (14)$$

From this, we can construct hypothetical factor values, starting from the current one, $f_i^k = f_{i,t} + \Delta f_i^k$. This new set of risk factor values can be used to construct a hypothetical value of the current portfolio under the new scenario, $P^k = P[f_1^k, f_2^k, \dots, f_N^k]$ using full valuation.

We can now compute changes in portfolio values from the current position $R_p^k = (P^k - P_t)$, sort these returns, and pick the one that corresponds to the c th quantile, $R_p(c)$. The HS VAR is obtained from the difference between the average and the quantile,

$$\text{VAR}^{\text{HS}} = \text{AVE}[R_p] - R_p(c) \quad (15)$$

Often, the mean is ignored. The advantage of this method is that it is easy to explain. The method is intuitive because each return is associated to an actual realization of risk factors, which gives insight into what could go wrong. The method also appears “distribution-free”, because it makes no specific distributional assumption, other than assuming that the recent past is relevant. This is an improvement over the normal distribution because historical data typically contain fat tails. In addition, because the portfolio can be priced under full revaluation, this approach

can handle options and other nonlinear instruments. This explains why this is perhaps the most popular VAR method. It is also interesting to note that this apparatus can easily handle stress tests, where scenarios are prespecified by the risk manager instead of drawn from the recent history.

Historical simulation has drawbacks, however, as discussed by Finger [4]. It usually relies on a short historical moving window, such as one year, to infer movements in market prices. This may be too short to provide adequate representation of the risk space.

In addition, this method does not adequately track time variation in risk. Figure 5 indeed reveals long-term cycles in volatility. With a long window, the method is very slow to pick up changes in the risk environment. An alternative method, called *filtered simulation*, is proposed by Hull and White [5]. This consists of first fitting a time-series volatility model, such as the EWMA to each risk factor. This produces a series of returns scaled by their volatility forecast σ_t

$$\varepsilon_t = R_t / \sigma_t \quad (16)$$

Historical simulation is then applied to the sequence of scaled returns ε_t , which are then multiplied by the latest volatility forecast for the next day σ_{t+1} . The advantage of this approach is that it does not assume that the scaled returns follow a normal distribution. This method combines any fat tails in the data with the more recent volatility forecast, which should lead to more precise estimates of the risk forecasts. The disadvantage of these more reactive methods, however, is that they create a more volatile capital charge unless there is some smoothing of the daily VAR measures.

Monte Carlo Simulation Method

The MCS method is similar to historical simulation, except that the movements in risk factors are generated by drawings from some prespecified distribution:

$$\Delta f^k \sim g(\theta) \quad (17)$$

where g is the joint distribution (e.g., normal, Student's t , or other) with parameters θ . The risk manager samples pseudorandom values for the risk factors from this distribution. As in the case of historical simulation, these are used to price the portfolio for different scenarios. The returns are then sorted to produce the desired VAR.

This method is the most flexible, because it allows full revaluation as well as complex distributions for the risk factors. On the other hand, it is computationally more onerous. Monte Carlo methods also create inherent sampling variability because of the randomization process. Different sequences of random numbers will lead to different results. It may take a large number of iterations to converge to a stable VAR measure. Finally, the Monte Carlo method requires users to make detailed assumptions about the form and parameters of the stochastic process. The process makes the output less amenable to economic interpretation than other methods. Therefore, it is subject to model risk. Risk managers need to understand how sensitive the results are to the model assumptions.

Backtesting

Whichever method is used, it is essential to compare the risk forecasts with the actual distribution of revenues. This provides a reality check on the accuracy of models. Backtesting involves systematically comparing historical VAR measures with the subsequent daily returns. Each exceedance is called an *exception*. Assume, for example, that VAR is measured at the 99% level of confidence. If the model is well specified, we should observe on average 2.5 exceptions over the course of a year. This is because 100–99%, or 0.01×250 trading days gives 2.5. In practice, some deviation around this number is to be expected. If there are too many exceptions, however, the risk manager will have to conclude that the model is flawed. Such backtesting provides valuable feedback to users about the accuracy of the VAR system. The procedure can also be used to search for possible improvements in the models.

One final drawback of risk measurement models should be mentioned. All models assume that the current positions are frozen over the horizon. In practice, trading will occur, which can change the trading risk profile. More exceptions can occur if intraday risk is greater than at the close of the trading day. This is why exception tests should also be performed on a hypothetical return series. This represents a frozen portfolio, obtained from fixed positions applied to the actual returns on all securities, measured from close to close. This also abstracts from fee income or interest payments that are included

in actual revenues. A bank experiencing too many exceptions using both actual and hypothetical returns should conclude that its risk model is fundamentally flawed.

Conclusions

Quantitative approaches to measuring market risk have become widespread. Risk is now increasingly measured at the top level of the institution, which involves the large-scale aggregation of data across many instruments and types of financial risks.

In spite of its quantitative aspects, risk management is still largely an art form, however. The design of risk architecture involves many tradeoffs. The goal of risk systems should be to produce reasonably accurate estimates of risk at a reasonable cost and within a reasonable time frame. Most importantly, the risk manager needs to be keenly aware of limitations in quantitative risk measures.

References

- [1] Artzner, P., Delbaen, F., Eber, J.M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**, 203–228.
- [2] Basel Committee on Banking Supervision. (1996). *Amendment to the Basel Capital Accord to Incorporate Market Risk*, BIS, Basel.
- [3] Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation, *Econometrica* **50**, 987–1007.
- [4] Finger, C. (2006). *How Historical Simulation Made Me Lazy*, Working Paper, RiskMetrics Group, New York.
- [5] Hull, J. & White, A. (1998). Incorporating volatility updating into the historical simulation method for value-at-risk, *Journal of Risk* **1**, 5–19.
- [6] Jorion, P. (2006). *Value at Risk: The New Benchmark for Managing Financial Risk*, McGraw-Hill, New York.
- [7] Jorion, P. (2007). *Financial Risk Manager Handbook*, Wiley, New York.
- [8] Morgan, J.P. (1995). *RiskMetrics Technical Manual*, J.P. Morgan, New York.

Related Articles

Backtesting; Bond; Convex Risk Measures; Expected Shortfall; Rare-event Simulation; Risk Exposures; Risk Management: Historical Perspectives; Risk Measures: Statistical Estimation; Stress Testing; Value-at-Risk.

PHILIPPE JORION

Models

When I was in grade school we used to build model airplanes out of kits. The frame was made out of precut pieces of balsa wood, each having been carefully pinned according to the plans along a preprinted arc to obtain the appropriate curvature, and then cemented, piece to piece, with airplane glue. The fuselage was made out of tissue paper, first glued to the balsa frame, trimmed, then dampened with water to shrink it taut, and finally, when dry, lacquered and painted to make it stiff and realistic. The engine was merely a long rubber band that ran the internal length of the fuselage, from propeller block at the nose to tail, wound up by rotating the propeller many times and then let loose to unwind for a flight of perhaps 10 seconds at best. If you were a really ambitious model builder, you followed the instructions very carefully: you were supposed to sand off any excess glue on the frame so as leave no imperfections at all.

What was “model” about the model airplane? The Zippy in the kit I remember building was smaller than a real Zippy (I presume there was somewhere in the world of real airplanes an actual Zippy); it was lighter; it was made out of totally different materials. But it did capture two essential features of the putatively real Zippy: appearance and flight. The model looked a lot like an airplane, and it could (briefly) fly.

Nevertheless, the model was not the Zippy. It was not the thing in itself. It was a model Zippy. It lacked seats, ailerons, proper windows, and doors among many other real-life details, and focused on only a few important features. What features are important depends on the model user. In my case, aged about 10, appearance and flight sufficed. At age three or four, crudely shaped wings, a body, and a throaty airplane engine noise might have been enough. A few years older and I would have wanted a combustion engine and radio control. But, however complex, none of these model Zippys would have been a Zippy.

What constrains the construction of a model Zippy?

First, the user and his needs. What aspects of the actual complex airplane and its surroundings is the user most interested in emulating, testing, playing, or tinkering with? An engineer needs a different

model than a child does. Second, engineering and construction: how to put together the model reliably and effectively, with as much accuracy in the key features as possible. Third, science: deep down below, even though the Wright Brothers probably did not know the partial differential equations of fluid flow, the possibility of heavier-than-air flight is built upon the science of mechanics and aerodynamics, on Newton’s laws and the Navier–Stokes equations.

Models in Physics

Scientific models are different. Resemblance is not enough. Scientific models aim at *divination*—foretelling the future—and controlling it, and there are two different approaches physicists use in creating such models.

Fundamental Models

The first approach is to aim at what physicists call a *fundamental model*, which, by nature, describes the *dynamics* behind events in the world. A fundamental model consists of a system of principles, usually formulated mathematically, that are then used to draw *causal* inferences about future behavior. *Dynamics* and *causality* are the essential characteristics. A fundamental model, particularly a successful one, is more of a theory than a model. To be a little pedantic, fundamental models proclaim that “These are the laws of the universe.” They describe the dynamics in God’s terms; they seek to state eternal truths, like Moses descending from the mountain.

An example all physicists are familiar with is Newton’s laws, which state that

$$F = ma \quad (1)$$

$$F = \frac{GMm}{r^2} \quad (2)$$

These are laws of cause and effect. Mass causes gravitational force. Force produces acceleration. Newton’s theory isolates the appropriate variables to use, and specifies a causal relationship between them.

The gap between a successful theory and the part of the universe it describes is virtually nonexistent:

the theory is the universe, not a model of the universe; the universe is the theory.

Phenomenological Models

The second type of model is what physicists call a *phenomenological* model. Phenomenological models are also used to make predictions, but they do not state absolute principles; instead they make pragmatic analogies between things you would like to understand and things you already do understand from more fundamental models. The analogies can be descriptive and useful, but analogies are self-limiting and there is often a toylike quality to them. In physics, one does not delude oneself into thinking of analogies as truth.

Phenomenological models do not say “This is a law.” Instead, they say “Approximately, you can think of this part of the world as being a lot like this other kind of thing you already understand more deeply.” Phenomenological models describe the world in man’s language rather than God’s.

A good example is the liquid drop model of the nucleus, where you think of an atomic nucleus as behaving much like an oscillating drop of fluid, even though you know that at a more reductionist level it is composed of protons and neutrons, themselves particles. Calibrating the liquid drop’s parameters to match the known properties of the nucleus, you can then use the model to compute and predict values of other yet unmeasured properties.

The gap between a successful phenomenological model and the part of the universe it describes is finite, actually, quite large. The model is an approximation, an effigy, a realistic-looking wax apple, Parrhasius’s painted curtain that fooled his fellow artist, a wonderful resemblance but not the thing itself.

Models in Finance

What is the point of a model in finance?

It takes only a little experience to see that it is not the same as the point of a model in physics or applied mathematics. Here is a simple but prototypical financial model. How do you estimate the price of a seven-room apartment on Park Avenue if someone tells you the market price of a typical two-room apartment in Battery Park City? Most likely,

you figure out the price per square foot of the two-room apartment. Then you multiply by the square footage of the Park Avenue apartment. Finally, you make some rule-of-thumb corrections for location, park views, light, facilities, and so on.

The model’s critical parameter is the implied price per square foot. You calibrate the model to Battery Park City. Then you use it to interpolate or extrapolate to Park Avenue. The price per square foot is *implied* from the market price of the Battery Park City apartments; it is not the construction price per square foot because there are other variables—exposure, quality of construction, neighborhood—that are subsumed into that one number.

The Aim of Financial Models

The way property markets use implied price per square foot illustrates the functions of financial models more generally.

Models are Used to Rank Securities by Value.

Implied price per square foot can be used to rank and compare many similar but not identical apartments. Apartments have manifold properties, as outlined above. Implied price per square foot provides a simple one-dimensional scale on which to *begin* ranking apartments by value. The single number given by implied price per square foot does not truly reflect the value of the apartment; it provides a starting point after which more qualitative factors must be taken into account.

Similarly, yield to maturity for bonds allows you to compare the values of many similar but not identical bonds, each with a different coupon and/or maturity, by mapping their yields onto a linear scale. The same applies to price/earnings (P/E) for stocks and option-adjusted spread (OAS) for mortgages or callable bonds. All of these metrics reduce a many-dimensional problem to a single-dimensional one. The Black–Scholes implied volatility of options provides a similar way of collapsing instruments with multiple qualities (strike, expiration, underlier, etc.) onto a single value scale and then making pragmatic modifications to it.

Models are Used to Interpolate from Liquid Prices to Illiquid Ones. In finance, models are used less for divination than to interpolate or extrapolate from the known dollar prices of liquid securities to

unknown dollar values of illiquid securities—in this case from the Battery Park City price to the Park Avenue value. Models in finance do not predict the future; instead they allow you to compare different prices in the present. Similarly, OAS is used to interpolate from relatively liquid bonds to less liquid ones. Correspondingly, the Black–Scholes model proceeds from a known stock price and a riskless bond price to the unknown price of a hybrid security, an option, similar to the way one estimates the value of fruit salad from its constituent fruits, or, inversely, the way one estimates the price of one fruit from the price of the other fruits and the salad.

None of these metrics are strictly accurate or correct, but they all provide immensely helpful ways of beginning to estimate value.

Models Transform Intuitive Linear Quantities into Nonlinear Dollar Values. In physics, a theory predicts the future. In finance, a model is also a means of translating intuition into dollar values. The apartment-value model transforms price per square foot into the dollar value of the apartment. It is intuitively easier to start from price per square foot (or per room) because it captures much of the variability of apartment prices. Similarly, P/E describes much of the variability of share prices. It is less easy to develop intuition about yield to maturity, option-adjusted spread, default probability, or return volatility than it is to think about price per square foot. Nevertheless, all of these parameters are clearly related to value and easier to think about than dollar value itself. They are intuitively graspable, and the more sophisticated you get, the richer your intuition. Models develop by leapfrogging from simple, intuitive, mental concepts (volatility, for example) to the mathematics that describes it (geometric Brownian motion and the Black–Scholes model) to richer mental concepts (the volatility smile) to experience-based intuition about them to newer models (stochastic volatility models, for example) that incorporate the newer concept.

In contrast to a fundamental or phenomenological model, the gap between a successful financial model and the correct value is nearly indefinable, because the fair value is finance’s *fata morgana*, undefined by prices, which themselves are not stationary, and so model success is temporary at best. If fair value were strictly calculable, there would be no markets.

Table 1 The aims of modeling

Field	Model aims
Physics	Reproduction, divination
Hobbyists	Resemblance
Finance	Ranking, interpolation, intuition

The qualities of models in different fields is summarized in Table 1.

The Foundations of Financial Engineering

Science—mechanics, electrodynamics, or molecular biology, for example—seeks to discover the fundamental principles that describe the world, and is usually reductive. Engineering is about using those principles, constructively, for a purpose.

Mechanical engineering is concerned with building devices based on the principles of mechanics (Newton’s laws) suitably combined with empirical rules about more complex forces (friction, for example) that are too difficult to derive from first principles. Electrical engineering is the study of how to create useful electrical devices based on Maxwell’s equations and solid-state physics. Bioengineering is the art of building prosthetics and other biologically active devices based on the principles of biochemistry, physiology, and molecular biology.

What about financial modeling, financial engineering, or quantitative finance? In a logically consistent world, financial engineering, layered above a base of solid financial science, would be the study of how to create functional financial devices—convertible bonds, warrants, credit default swaptions, and so on—that perform in desired ways, not just at expiration, but throughout their lifetime. Which brings us to financial science, the study of the fundamental laws of financial objects, whether they are stocks, interest rates, or whatever else a theory uses as atomic constituents. Here, unfortunately, lie dragons.

Brownian motion, the underpinning of much of quantitative finance, is indeed science, but it is accurate science only for small particles bumped around by invisible atoms. For stocks, the standard theory of geometric Brownian motion is an idealization that, while it captures some of the essential features of price uncertainty, is not a very good description of the detailed characteristics of their price

distributions. Markets are plagued and blessed with anomalies that disagree with standard and nonstandard theories. So, while we in financial engineering are rich in techniques (stochastic calculus, optimization, the Hamilton–Jacobi–Bellman equation, etc.), we do not (yet? ever?) have the right laws of science to exploit.

What solid laws and concepts do we have for building our ranking and translation models? In truth, only one.

The One Law of Financial Modeling

According to a legend, Hillel, a famous Jewish sage, was asked to recite the essence of God’s laws while standing on one leg. “Do not do unto others as you would not have them do unto you,” he is supposed to have said. “All the rest is commentary. Go and learn.”

You can summarize the essence of quantitative finance on one leg too: “If you want to know the value of a security, use the known price of another security that is as similar to it as possible. All the rest is modeling. Go and build.”

“Security” here refers not only to a single security but also to a portfolio of securities. The wonderful thing about this law—it is valuation by analogy—when compared with almost everything else in economics, is that it dispenses with utility functions, the unobservable hidden variables whose ghostly presence permeates most of *faux* quantitative economic theory. Financial economists refer to their essential principle as the *law of one price*, or the principle of *no riskless arbitrage*, which states that:

Any two securities with identical future payoffs, no matter how the future turns out, should have identical current prices.

The law of one price, this valuation by analogy, is the only genuine law we have in quantitative finance, and it is not a law of nature.^a It is a general reflection on the practices of human beings, who, when they have enough time and enough information, will grab a bargain when they see one. The law usually holds in the long run, in well-oiled markets with enough savvy participants, but there are always short- or even longer-lived and persistent exceptions.

How do you use the law of one price to determine value? If you want to estimate the unknown value of a target security, you must find some other replicating

portfolio, a collection of more liquid securities that, collectively, has the same future payoffs as the target, no matter how the future turns out. The target’s value is then simply the value of the replicating portfolio.

Where do models enter? It takes a model to show that the target and the replicating portfolio have identical future payoffs under all circumstances. To demonstrate payoff identity, (i) you must specify what you mean by “all circumstances”, for each security, and (ii) you must find a strategy for creating a replicating portfolio that, in each future scenario or circumstance, will have payoffs identical to those of the target. That is what the Black–Scholes options model does; it tells you exactly how to replicate or manufacture fruit salad—an option—out of fruit—stock and bonds. The appropriate price should be the cost of manufacture.

The tricky part of building these models is specifying what you mean by “under all circumstances.” In Black–Scholes all circumstances means a future in which stock returns are normally distributed and stock prices move continuously. Of course, that is not exactly true. Trying to specify “all circumstances” reminds me of the 1960s movie *Bedazzled*, starring Peter Cook and Dudley Moore. It retells the story of Faust, in which Dudley Moore plays a short-order cook in a Wimpy’s in London who sells his soul to the devil in exchange for seven chances to specify the circumstances under which he can achieve his romantic aims with the Wimpy’s waitress he covets. Each time the devil asks him to specify the romantic scenarios under which he believes he will succeed, he cannot get them quite specific enough. He says he wants to be alone with the waitress in a beautiful place where they are both in love with each other. He gets what he wants—a snap of the devil’s fingers and he and his mutually beloved are instantaneously transported a country estate, where he is a guest of the owner, her husband, whom her principles will not allow her to betray. In the final episode, he wishes for them to be alone together and totally in love in a quiet place where no one will bother them. He gets his wish: the devil makes them both nuns in nunnery where everyone has taken a vow of silence. This is the same difficulty you have specifying future scenarios in financial models—like the devil, markets always outwit you. Even if they are not strictly random, their

vagaries are too rich to capture in a few sentences or equations.

Model Risk

Risk is future uncertainty. A coin flip is risky. You know the current state of the coin, but not the future one. You can, however, perform an infinite number of mental flips and reliably calculate the probability distribution of heads and tails, which will match a physical coin's probability distribution to the extent that the coin is separable from its surroundings and uninfluenced by them.

In this sense, a liquid stock price is risky. You (more or less) know the current price and have no idea about the direction of its future change. But you cannot perform an infinite number of mental stock price moves with any reliability; the stock, the market and the world are not clearly separable and influence each other, so that the probability distribution of stock prices cannot be accurately known (and may not be time invariant). The history of the world does not affect a coin flip. The history of the world does have a bearing on the next change in a stock's price. The risk of a stock price is qualitatively different from the risk of a coin flip.

Financial models interpolate from liquid to illiquid prices by analogy, and must necessarily change over time as the economic environment changes or as market participants become more sophisticated. The Black–Scholes model, for example, used to be regarded as adequate for valuing exotic options prior to the market crash of 1987, but is now often replaced by a range of extended models that incorporate local volatility, stochastic volatility, or jumps. One does not know the truly correct current model, let alone the future one, and so the correct model is uncertain not only in the future but in the present too [3]. The word “risk” therefore inaccurately describes the indeterminate nature of financial models. If you want to describe this state of ignorance as risk, then do not forget that it is a shorthand term for uncertainty, for something much vaguer than probabilistic risk. There is no ensemble of models each with an associated known probability of being right.^b

Idolatry

The greatest danger in financial modeling is therefore the age-old sin of idolatry. Financial markets are alive

but a model is a limited human work of art. A model may be entrancing but no matter how hard you try, you will not be able to breathe true life into it. To confuse the model with the world is to embrace a future disaster driven by the belief that humans obey mathematical rules.

Financial modelers must therefore compromise, must firmly decide what small part of the financial world is of greatest current interest, decide on its key features, and make a mock-up of only those. A model cannot include everything. If you are interested in everything, you are interested in too much. A successful financial model must have limited scope; you must work with simple analogies; in the end, you are trying to rank complex objects on a low-dimensional scale. In physics there may one day be a Theory of Everything; in finance and the social sciences, you are lucky if there is a usable theory of anything.

Models are best regarded as a collection of parallel inanimate thought universes you can explore. Each universe should be consistent, but the actual financial and human world, unlike the world of matter, is going to be vastly more complex and vital than any model you make of it. You are always trying to shoehorn the real world into one of the models you have to see how usefully it approximates the key features you are concerned with.

The right way to engage with a model is, as a fiction reader does, to temporarily suspend disbelief, and then push it as far as you can. The success of the theory of options valuation, the best model economics can offer, is the story of a platonically simple theory, taken more seriously than it deserves and then used extravagantly, with hubris, as a crutch to human thinking. “If a fool would persist in his folly, he would become wise,” wrote Blake in *The Marriage of Heaven and Hell*. That is what options markets have done with options theory.

A little hubris is good. But catastrophes strike when hubris evolves into idolatry. Somewhere between these two extremes, a little north of common sense but still south of idolatry, lies the wise use of conceptual models.

Appendix: Avoiding Modeling Errors

- Expect neither miracles nor mechanical formulas; expect only aids to rational thinking.

- Many financial phenomena are not mathematically modelable in a useful way. You are worse off thinking you have a model and relying on it than simply realizing you do not, and then allowing for a margin of error.
 - Since financial models translate intuitive qualities into dollar values, and since financial data are often sparse, unreliable, and unstationary, models should be relatively simple. Complexity in the face of lack of highly detailed knowledge is pointless.
 - Deal with concepts first, then mathematical formulation, then calibration.
 - Markets are by definition vulgar, and correspondingly the most useful models are wisely vulgar too, using variables that the crowd uses to describe the phenomena they observe. In physics it pays to drop down deep, several levels below what you can observe (think of Kepler, Newton, Einstein, Schrodinger), formulate an elegant principle, and then rise back to the surface again to work out the observable consequences. In finance, which lacks deep scientific principles, it is better to use models that have as direct as possible a path between observation and consequences. (Of course, over time, crowds and markets get smarter and the definition of vulgarity changes to encompass increasingly sophisticated concepts.)
 - Markets learn from experience and proceed to adopt new paradigms and make new mistakes. What is a good model today may be inappropriate tomorrow. Be sure to understand what your model is assuming and what it is ignoring.
 - Regard models as interdisciplinary endeavors. Financial models are generally not back-of-the-envelope formulas handed over to “coders” to turn into executable instructions. Modeling is multidisciplinary: it touches on the practicality of doing business, on financial theory, on mathematical modeling and computer science, on computer implementation, and on the construction of user interfaces. Models end up as computational computer programs embedded in human and machine interfaces that are themselves computer programs. The risks lie in the knowledge of the business, the applicability of the financial model, the mathematics and numerical analysis used to solve it, the computer science used to implement and present it, and in the transmission of information and knowledge accurately from one part of the model, in the larger sense of the word, to the next. It helps to be knowledgeable in all of these areas in order to notice an error and then diagnose it.
 - To avoid errors, it is important to have modelers, programmers, and users who all work closely together, understand each other’s domains well enough to know what constitutes a warning symptom, and have a good strategy for testing a model and its limits. Too much specialization is harmful.
 - Test complex models in simple cases first. Often there are simple special-case known solutions; a deep-in-the-money call option is a forward, for example. Test the model at these boundaries first.
 - Do not ignore small numerical discrepancies. Track down their origin. Small disagreements often serve as warnings of potentially large disagreements and disastrous errors under other scenarios.
 - Usable financial models are software, and are subject to all the problems of writing, debugging, and maintaining large-scale software systems. Be reluctant to give up content for the sake of analytical elegance.
 - Work closely with end users (traders, salespeople, etc.) to embed the software in their environment in the most convenient way.
 - Informed and patient users who clearly comprehend both the model and the method of solution are invaluable in discovering model errors.
 - Efficient testing requires a good user interface.
 - Diffuse a new model slowly outward through the organization, from its developers to developers of other models to the end users.
 - Pride of ownership: one of the best defenses against modeling errors is to ensure that both models and systems are built by people who love what they do, take pride in their work, and are rewarded for doing it well.
- For a more detailed discussion of modeling errors and how to avoid them, see [2]. See also [4].

End Notes

^aThe time value of money, the benefits of diversification and the value of the right to choose are other useful principles. I thank Marcos Carreira for these comments.

^bThis point has also been made by R. Cont [1].

References

- [1] Cont, R. (2006). Model uncertainty and its impact on the pricing of derivative instruments, *Mathematical Finance* **16**(3), 519–547.
- [2] Derman, E. (1996). Model risk, *Risk* **9**(5), 139–145.
- [3] Derman, E. (2001). Markets and models, *Risk* **14**(7), 48–50.
- [4] Crouhy, M., Mark, R. & Galai, D. (2000). *Risk Management*, Chapter 15, McGraw Hill.

Related Articles

Ambiguity; Arbitrage Pricing Theory; Arbitrage Strategy; Efficient Market Hypothesis; Model Calibration; Predictability of Asset Prices.

EMANUEL DERMAN

Operational Risk

A basic ingredient of the Basel II guidelines of the Basel Committee on Banking Supervision is the introduction of the new risk class of operational risk (OR). OR is the risk of loss resulting from inadequate or failed internal processes, people, and systems or from external events. This definition includes legal risk, but excludes strategic and reputational risk. OR complements the more traditional risk classes of market risk (MR) and credit risk (CR), but it is more concerned with the overall quality assurance of banking operations, and as such borrows methodology more from such fields as reliability engineering and insurance mathematics. For the latter, see, for instance, McNeil *et al.* [17, Chapter 10] and Panjer [20]. The capital charge for OR may range from 10 to 20%, say, of overall risk capital and for some banks top the charge for MR. Under the so-called Pillar I of Basel II, banks can choose among three increasingly quantitative approaches for the measurement of OR. The first two, namely, the basic indicator (BIA) and the standardized (SA) approach, yield volume-based measures for OR: risk capital is charged in proportion to gross income, either companywide in the case of the BIA or averaged across a number (typically 7) of business lines (BL) for the SA.

Larger international banks have to use the so-called advanced measurement approach (AMA), which, upon introduction in January 2008, is mainly referred to as the *loss distribution approach* (LDA). Under the Basel II setup, banks using the LDA have to prove to the local regulator that they have the necessary quantitative risk management environment. Further, insurance of OR losses up to 20% is allowed. Capital estimates are to be based on internal, external, as well as expert opinion data. Essentially, banks are free to choose their own analytic approach for the calculation of an LDA-based capital charge for OR. In the second section, we review some of the approaches used. The third section contains some comments on further work.

The LDA Approach

Basel II prescribes the use of Value-at-Risk as a risk measure for OR, with a confidence level of 99.9% and a measurement horizon of 1 year. Hence,

denoting the total OR loss for the coming year T by L_T , the risk capital charge becomes

$$\text{RC}_{OR}^T = \text{VaR}_{99.9\%}^T(L_T) = F_{L_T}^{\leftarrow}(0.999) \quad (1)$$

where $F_{L_T}^{\leftarrow}$ denotes the generalized inverse of the distribution function (df) of the total loss random variable L_T . This function is also referred to as the *quantile function*. The variety of LDA models used by the industry now stems from the granularity of the stochastic frequency-severity models of L_T implemented, as well as from the extreme tail-fitting of F_{L_T} in order to estimate the 99.9% quantile.

For internal data of banks, the basic structure of the OR data warehouse is as follows:

$$\mathcal{X} = \{X_k^{T-i,b,r} : i = 1, \dots, n, \quad b = 1, \dots, B, \\ r = 1, \dots, R, \quad k = 1, \dots, N^{T-i,b,r}\} \quad (2)$$

where

i	stands for the number of past observation years;
B	stands for the different BL, typically $B = 8$;
R	denotes the different risk types (RT); often $R = 7$;
N	(frequency) denotes the number of losses in year $T - i$, BL b , and RT r ; and
X	(severities) denotes the successive loss amounts in these classes.

Finally, banks are allowed to cap the loss amounts recorded from below, for example, losses are only recorded above USD 20 000. As a consequence, the loss random variable (rv) to be modeled becomes

$$L_T = \sum_{b=1}^B \sum_{r=1}^R \sum_{k=1}^{N^{T,b,r}} X_{k,+}^{T,b,r} \quad (3)$$

where the “+” denotes the above capping: a daunting task indeed! It should be remarked at this point that the calculation of a capital charge for OR is methodologically very different from that for MR and CR. For MR, the confidence level is 99% and the modeling horizon is 10 days; this becomes 99.9% and 1 year for CR. Besides this, the underlying processes are totally different; for OR only losses are to be modeled and obviously, a bank does not “take positions” in OR. A consequence of this is that, in our opinion, regulators have moved a bit too fast up the hill of the ultimate risk management goal: a unified quantitative capital charge formula across all risk

classes. We firmly believe that, before reaching this ultimate goal, we have to walk up and down the quantitative risk management hill a couple of times. The final satisfaction with the results obtained may crucially depend on this.

What have we learned so far from the various quantitative impact studies (QIS) presented by the regulators (see [18] and [8])? First of all, severity data are very heavy-tailed. Moscadelli [18] even reports infinite-mean models for several of the BL-aggregated loss data. Data are furthermore very skewed. Finally, though frequency is no doubt important, it is mainly severity (“the one claim causes ruin”-type of paradigm) that calls the tune. In [18], extreme value theory (EVT) is used for the high-quantile estimation of $F_{L_T}^{\leftarrow}(0.999)$. On the other hand, Dutta and Perry [8] question the use of EVT, as the final capital charge may be very unstable with respect to the starting values of the resulting estimation procedure. See [5] and [19] for more information on this. Dutta and Perry [8] champion the use of the so-called *g-and-h* df as a flexible family of loss severity models for OR data. An rv X has a *g-and-h* df if for $Z \sim N(0, 1)$, and $a, b, g, h \in \mathbb{R}$,

$$X = a + b \frac{e^{gZ} - 1}{g} e^{1/2hZ^2} \quad (4)$$

For a comparison of EVT versus *g-and-h* based estimates of the OR capital charge, see [7], [6], and [15]. The first two references explain the methodological difficulties that would be present when using an EVT-based analysis for real *g-and-h* data, whereas the latter reference, based on the industry OR data, claims that EVT-based risk capital calculations work well till a confidence level of about 95%, whereas for higher confidence levels the *g-and-h* approach seems to deliver more reliable (acceptable) capital values. In our view, the jury is still out. For readers interested in a complete discussion of OR modeling in a large international bank, see [1]; in this article, various modeling issues that we had to omit here because of space constraint have been touched upon. For early contributions to the LDA, see for instance [2, 13, 14].

So far, we have only briefly mentioned the stochastic modeling of the internal OR loss data. The regulators require a combination of internal with external and expert opinion data. Several companies are providing industrywide OR loss data on which banks can calibrate their internal loss distribution

curves. At the moment, it is still early to comment on the validity and the ultimate effect of this. Furthermore, in the realm of OR, expert opinion is called for; it may be borne in mind that we are mainly estimating rare-frequency, high-severity events. As a consequence, a Bayesian approach presents itself. Within the (actuarial) realm of credibility theory, such an approach has for instance been worked out by Bühlmann *et al.* [4]. For a textbook treatment on credibility theory and its link to empirical Bayes, see [3]. Shevchenko and Wüthrich [21] and Lambrigger *et al.* [16] use the concept of conjugate priors within a Bayesian framework.

Further Work

The brief discussion above hopefully highlights some of the key issues/problems for the Pillar I quantitative modeling of OR. Our main concern is that for such extremely heavy-tailed data, VaR-based risk management is, at best breaking new ground and at worst not really applicable. This brings us to some of the recent research on using VaR as a one-fit-all-size risk management tool. For instance, many banks aggregate data BL-wise, estimate the resulting VaRs, $\widehat{\text{VaR}}_1, \dots, \widehat{\text{VaR}}_B$, and add these to obtain

$$\text{VaR}_+ = \sum_{b=1}^B \widehat{\text{VaR}}_b \quad (5)$$

Then they have to come up with a so-called diversification-based capital reduction argument to yield

$$\text{RC}_{\text{OR}}^T < \text{VaR}_+ \quad (6)$$

A more careful look at this issue shows that one is in danger of running head on into the wall of potential nonsubadditivity of the VaR, that is,

$$\text{VaR}(L_1 + L_2) > \text{VaR}(L_1) + \text{VaR}(L_2) \quad (7)$$

in cases important for OR practice; see [19] and the references therein for a discussion on this.

A related mathematical problem we would like to point out at this point is the following: suppose $X_1, \dots, X_d$ are d one-period risks, $\Psi : \mathbb{R}^d \rightarrow \mathbb{R}$ a function corresponding to a financial position $\Psi(X_1, \dots, X_d)$, and ρ a risk measure so that we want to calculate (estimate)

$$\rho(\Psi(X_1, \dots, X_d)) \quad (8)$$

Full determination of equation (8) requires the knowledge of the joint df $F_{X_1, \dots, X_d}$ which in many cases is not known. As a consequence, only upper and lower bounds on equation (8) can be calculated. This problem is dealt with in [9–11]. In [12], the authors discuss the difference between BL-wide aggregation of the VaR-estimates versus RT-wide aggregation. The final difference very much depends on the specific stochastic properties of the frequency-severity models for each cell of the OR loss data matrix.

Conclusion

The Pillar I based quantitative modeling of OR data still has numerous problems: lack of historical data, extreme heavy-tailedness leading to possible super-additivity of the VaR and hence, the questioning of diversification effects based on the VaR, and finally, the overall statistical uncertainty in the modeling of extremes for data of OR type. Before a sufficiently reliable method for OR capital estimation can be found, much more data-oriented work is needed.

References

- [1] Aue, F. & Kalkbrener, M. (2006). LDA at work: Deutsche Bank's approach to quantifying operational risk, *Journal of Operational Risk* **1**(4), 49–93.
- [2] Baud, N., Frachot, A. & Roncalli, T. (2003). *How to Avoid Over-Estimating Capital Charge for Operational Risk*, OpRisk & Compliance.
- [3] Bühlmann, H. & Gisler, A. (2005). *A Course in Credibility Theory and its Applications*, Springer-Verlag, Berlin.
- [4] Bühlmann, H., Shevchenko, P.V. & Wüthrich, M.V. (2007). A “toy” model for operational risk quantification using credibility theory, *Journal of Operational Risk* **2**(1), 3–19.
- [5] Chavez-Demoulin, V., Embrechts, P. & Nešlehová, J. (2006). Quantitative models for operational risk: extremes, dependence and aggregation, *Journal of Banking and Finance* **30**, 2635–2658.
- [6] Degen, M. & Embrechts, P. (2008). EVT-based estimation of risk capital and convergence of high quantiles, *Advances of Applied Probability* **40**(3), 696–715.
- [7] Degen, M., Embrechts, P. & Lambrigger, D.D. (2007). The quantitative modelling of operational risk: between g -and- h and EVT, *ASTIN Bulletin* **37**(2), 265–291.
- [8] Dutta, K. & Perry, J. (2006). *A Tale of Tails: An Empirical Analysis of Loss Distribution Models for Estimating Operational Risk Capital*, Federal Reserve Bank of Boston, Working Paper No. 6–13.
- [9] Embrechts, P. & Puccetti, G. (2006). Bounds for functions of dependent risks, *Finance and Stochastics* **10**, 341–352.
- [10] Embrechts, P. & Puccetti, G. (2006). Bounds for functions of multivariate risks, *Journal of Multivariate Analysis* **97**(2), 526–547.
- [11] Embrechts, P. & Puccetti, G. (2008). Aggregating risk capital, with an application to operational risk, *Geneva Risk and Insurance Review* **31**(2), 71–90.
- [12] Embrechts, P. & Puccetti, G. (2008). Aggregating operational risk across matrix structured loss data, *Journal of Operational Risk* **3**(2), 29–44.
- [13] Frachot, A., Moudoulaud, O. & Roncalli, T. (2003). Loss distribution approach in practice, in Chapter 15 of *The Basel Handbook: A Guide for Financial Practitioners*, M. Ong, ed., Risk Books.
- [14] Frachot, A., Roncalli, T. & Salomon, E. (2004). *Correlation and Diversification Effects in Operational Risk Modelling*, OpRisk & Compliance.
- [15] Jobst, A.A. (2007). Operational risk: the sting is still in the tail but the poison depends on the dose, *Journal of Operational Risk* **2**(2), 3–59.
- [16] Lambrigger, D.D., Shevchenko, P.V. & Wüthrich, M.V. (2007). The quantification of operational risk using internal data, relevant external data and expert opinion, *Journal of Operational Risk* **2**(3), 3–27.
- [17] McNeil, A.J., Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, Tools*, Princeton University Press, Princeton.
- [18] Moscadelli, M. (2004). *The Modelling of Operational Risk: Experience with the Analysis of the Data Collected by the Basel Committee*. Temi di discussione, Banca d'Italia.
- [19] Nešlehová, J., Embrechts, P. & Chavez-Demoulin, V. (2006). Infinite-mean models and the LDA for operational risk, *Journal of Operational Risk* **1**(1), 3–25.
- [20] Panjer, H.H. (2006). *Operational Risk*, John Wiley & Sons, Hoboken, NJ.
- [21] Shevchenko, P.V. & Wüthrich, M.V. (2006). The structural modeling of operational risk via Bayesian inference, *Journal of Operational Risk* **1**(3), 3–26.

VALÉRIE CHAVEZ-DEMOULIN & PAUL
EMBRECHTS

Regulatory Capital

The Road to Basel II

Regulatory authorities impose minimum capital requirements and monitor banks through on- and off-site supervision to solve the asymmetric information problem created by the opacity of banks' business and their high leverage. Besides depositor protection, securing the banking sector's key role in the economy, including the monetary transmission process, the credit supply and the provision of payment systems, is an equally important motivation for bank regulation.^a

The development of the "1988 Basel Capital Accord", in short "Basel I", is closely linked to developments in the banking industry after the collapse of the fixed exchange rate system of Bretton Woods. Deregulation of the banking sector, high volatility of financial market prices, particularly exchange rates and interest rates, and increasing competition that fueled rapid lending growth, ultimately contributed to an increasing number of breakdowns in the banking industry. The failure of the German Herstatt Bank in 1974 was an early warning signal. Basel I was the regulatory answer to these developments and became the first common regulatory framework for large and internationally active banks.

The risk sensitivity of minimum capital requirements for credit risk is implemented (and limited) in Basel I by conditioning the risk weight on the borrower type. Loans to sovereigns or banks, for example, receive lower risk weights than corporate exposures. The securitization of financial claims, which has developed into a huge market over the last decade, was not covered at all by Basel I, although rules were often implemented on a national level. At the turn of the millennium, innovations in the financial industry and risk management practice in large, internationally active banks had outpaced the Accord by creating ample opportunities for regulatory capital arbitrage (see [28], for examples).

The 1996 amendment extends the unilateral focus of the Accord on credit risk by adding rules for market risks. Market risks comprise the risk of losses in on- and off-balance sheet positions caused by changes in market prices. More specifically, the amendment requires banks to hold capital for interest rate risk and equities risk in the trading book,

foreign currency risk, and commodities risk. It also introduces, for the first time, an option for banks to choose between a simpler, less risk-sensitive *standardized method* and a more *advanced internal models approach* that allows the use of banks' own *Value at Risk* models for calculating the required capital cushion.

Different approaches distinguished by the degree of risk sensitivity and individual complexity introduce an evolutionary aspect into the Basel II framework. Banks can migrate to a more complex approach with overall lower capital requirements provided they meet higher standards in risk management and risk measurement procedures. Its key objectives as stated by the Basel Committee on Banking Supervision were the focus on large, internationally active banks, the macroprudential focus on the stability of the banking system and the goal to enhance competitive equality.

An important way to promote both financial stability and a level playing field is to strengthen risk sensitivity of the minimum capital requirements. This reduces opportunities of *regulatory capital arbitrage*, which may otherwise encourage banks to carry out business transactions entailing risks that are not adequately captured by regulatory capital requirements and to avoid transactions that are overly burdened.

The risk sensitivity of the capital requirements for credit risk, for example, is increased in Basel II by replacing the static risk weights of Basel I with risk weights that either depend on external ratings or are based on a stylized credit risk model and require banks' own estimates of default probabilities as input. Furthermore, operational risk is included in the minimum capital requirements as a new risk category. This transition to model-based minimum capital requirements has not always been without its detractors. Indeed, concerns were raised about the limitations of quantitative models, advocating simpler approaches for regulatory purposes (see [17], especially, for market risk models).

Banking supervision operates somewhere between the two conceptual extremes of rules-based and principles-based regulation. Basel II contains important elements of both concepts. The calculation of a regulatory capital charge based on banks' own estimates of risk components which are inputs to regulatory risk-weight functions is an example of a rules-based approach. The regulatory approval of

internal models both for market risk in the trading book and for operational risk is an example of a principles-based approach. In the early stages of its development, a principles-based internal models approach for credit risk was still considered a viable option for the new framework [26]. On the basis of the results of an internal study, the Basel Committee acknowledged that the state of model development at that time could not ensure sufficient comparability across institutions and that a meaningful validation of model parameters and assumptions would be prevented by data limitations [5]. As a consequence, internal models are not eligible for the calculation of regulatory capital for credit risk. The risk-weight functions are defined in the new framework and only the risk components, for example, default probabilities, have to be estimated by banks' internal methods.

Basel II rests on three pillars, the first of these being the *minimum capital requirements*. The *supervisory review process*, which constitutes the second pillar, provides guidance for supervisory practice and banks' internal capital adequacy assessment processes. Whereas the first pillar is a typical rules-based approach, the second pillar is principles-based on account of its very purpose. The minimum capital requirements and guidance on the supervisory review process are complemented by disclosure requirements to enforce *market discipline*, the third pillar of the new framework.

Since this article is concerned with the methodological underpinnings of the new framework, its main focus is on the first pillar and here on the most risk-sensitive, model-based methods. The second pillar is only touched upon while the third is omitted altogether.

Capital Ratios as Solvency Indicators

The regulatory capital ratio links a bank's loss-absorbing capital in the numerator to its aggregated risks, measured by risk-weighted assets, in the denominator. If a bank has recognized internal models for market risk and operational risk, the *total capital ratio* (TCR), which should not fall below 8% is defined in Basel II as follows:

$$TCR = \frac{TC - RD_{CrR} - Ded}{RWA_{CrR} + 12.5 VaR_{MkR} + 12.5 VaR_{OpR}} \quad (1)$$

TC	: <i>total capital</i> consisting of core capital and supplementary capital
RD_{CrR}	: <i>regulatory difference</i> between expected loss and provisions for credit risk (only)
Ded	: deductions (e.g., certain investments in subsidiaries and securitization-related positions)
RWA_{CrR}	: <i>risk-weighted assets</i> for credit risk in the banking book
VaR_{MkR}	: <i>Value at Risk</i> for market risk in the trading book (internal models approach)
VaR_{OpR}	: <i>Value at Risk</i> for operational risk (advanced measurement approach).

Total capital can be decomposed into core capital (tier 1), for example, paid-up share capital, and supplementary capital (tier 2 and tier 3). Tier 2 capital consists of, for example, general loan loss reserves and subordinated debt (with a maturity of at least five years). Tier 3 capital consists of short-term subordinated debt and can cover market risk alone and this only if the bank has a recognized internal model. Besides the 8% minimum for the TCR, banks also need to set aside a minimum of 4% for their tier 1 ratio, which contains only tier 1 capital in the numerator.

Minimum requirements for credit risk in the banking book typically make the highest contribution to a commercial bank's total regulatory capital charge. They can be measured by two different approaches: the *revised standardized approach* (RSA) and the *internal ratings-based approach* (IRBA). The latter is further differentiated in the *foundation IRBA* and the more complex and risk-sensitive *advanced IRBA*. The RSA risk weights depend on the borrowers' external ratings, whereas the IRBA risk weights are model based and require banks' own estimates of the probability of default, the (expected) loss given default, and the maturity of the exposure. The parameterization of the risk-weight functions differs between asset classes, for example, sovereigns, corporates, and retail clients.

Its widespread use in the industry favors the *Value at Risk* (see **Value-at-Risk**) as the risk measure for the calculation of regulatory capital charges. The capital under a Value at Risk rule K_{VaR} is set sufficiently high so that it covers portfolio losses L

with a probability α over a one-year horizon:

$$K_{VaR} = \inf \{k \mid \Pr(L \leq k) \geq \alpha\} \quad (2)$$

The “confidence level” or solvency probability α for credit risk and operational risk, for example, is set to 99.9%. This value is consistent with empirical findings suggesting that large internationally active banks even reach higher target levels so that the new minimum requirements should not represent a binding constraint [27].

The Value at Risk of a credit portfolio can be decomposed into the *expected loss* (EL) and the *unexpected loss* (UL). Whereas EL is the mean of the loss distribution, UL is defined as the difference between an adverse quantile of the portfolio’s loss distribution (e.g., given by a confidence level of 99.9%) and EL. Expected losses are traditionally covered by general provisions (or general loss reserves) and also considered in the calculation of the interest rate margin whereas UL is covered by a capital cushion. In the IRBA, the risk-weighted assets cover only UL. Therefore, if EL is covered by provisions or revenues and if the model is correct, the likelihood that the bank stays solvent at the risk horizon is equal to the confidence level, consistent with equation (2).

The risk-weighted assets for credit risk in the IRBA are computed by aggregating over the UL-based capital charges UL_n of all N individual credit claims in the portfolio:

$$RWA_{CrR} = 12.5 \cdot \sum_{n=1}^N UL_n \quad (3)$$

To avoid distortions of the level playing field by differences in banks’ EL coverage by provisions, any shortfall needs to be deducted in equal shares from tier 1 and tier 2 capital in the numerator of the capital ratio. This shortfall or *regulatory difference* can be negative if the total eligible provisions (TEP) exceed EL. In this case, the regulatory difference RD_{CrR} is recognized in the tier 2 capital, but only up to a limit of 0.6% of risk-weighted assets for credit risk:

$$RD_{CrR} = \max \{EL - TEP, 0\} - \min \{\max \{TEP - EL, 0\}, 0.006 \cdot RWA_{CrR}\} \quad (4)$$

In the internal models approach for market risk and in the *advanced measurement approach* (AMA)

for operational risk, regulatory capital is measured by the Value at Risk and, therefore, the model output needs to be scaled by a factor of 12.5 to be consistent with the risk-weighted assets for credit risk in the denominator of the capital ratio in equation (1).

Capital Requirements for Credit Risk

Asymptotic Single-factor Model

The capital charge for a single asset measures its marginal contribution to the capital requirement for the total portfolio. A *decentralized capital rule*, often referred to as *portfolio invariance* property, requires that the capital charge for an asset depends only on its own risk characteristics and not on the composition of the rest of the portfolio. It is convenient for regulatory capital requirements because it makes it possible to compute the capital charge of a portfolio by adding up the independently calculated capital charges of its assets. This property generally does not hold under a Value at Risk based capital rule because the marginal contribution of a single asset depends on the correlation with the other assets in the portfolio. Therefore, any change in the portfolio composition can affect the capital charge of an individual asset (see **Economic Capital Allocation**).

Gordy [21] was able to show that portfolio invariance can be achieved for a limiting case given a single-factor version of a common credit risk model. In the following, we refer to this model as the *asymptotic single risk factor* (ASRF) model for reasons that will soon become clear. Large-sample approximations had already been used earlier for computational shortcuts but only with additional homogeneity assumptions for the portfolio ([39] and also see **Large Pool Approximations**).

The ASRF model is a stylized version of an asset value model that belongs to the class of *conditionally independent factor models* [38]. The key variable is the stochastic default-trigger or ability-to-pay variable Y_n of the n th firm, which is often loosely interpreted as the standardized asset value return of the borrower. Y_n depends on a single systematic risk factor X and an idiosyncratic risk factor U_n :

$$Y_n = \sqrt{\rho} X + \sqrt{1 - \rho} U_n \quad (5)$$

The parameter ρ denotes the asset correlation between any pair of firms. The risk factors X and U_n

are pairwise independent and standard normally distributed, such that Y_n is also standard normal. Default is triggered if Y_n falls below a default threshold c_n at the end of a risk horizon of (typically) one year. The (unconditional) probability of default is given by $\Phi(c_n)$ where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

Gordy [21] identifies the specific conditions for replacing L in equation (2) asymptotically by $\mathbb{E}(L|q_X(\alpha))$, which leads to a simple capital formula where $q_X(\alpha)$ denotes the $(1 - \alpha)$ -quantile of the distribution of X . Apart from technical conditions, such as the monotonicity of losses in X in the aggregate, two economically relevant conditions need to be fulfilled for a portfolio invariant capital rule. First, the dependence structure of the model is driven (solely) by the dependence of the default-trigger Y_n on the common systematic risk factor X . Second, portfolio invariance requires that the credit portfolio is perfectly fine grained, which means that the idiosyncratic risk has been eliminated by diversification.

Under these two assumptions, the required capital UL_n to cover the UL of the n -th asset over a one-year horizon is given by the following formula:

$$UL_n = EAD_n LGD_n [p_n(\Phi^{-1}(1 - \alpha)) - PD_n] \quad (6)$$

with

$$p_n(\Phi^{-1}(1 - \alpha)) = \Phi \left(\frac{\Phi^{-1}(PD_n) - \sqrt{\rho} \Phi^{-1}(1 - \alpha)}{\sqrt{1 - \rho}} \right)$$

The aggregation of UL_n over all assets gives the capital of the portfolio. The term $p_n(\Phi^{-1}(1 - \alpha))$ in equation (6) can be interpreted as the conditional default probability of the borrower, conditional on the value $\Phi^{-1}(1 - \alpha)$ of the systematic risk factor. The value of α is 0.999, implying a solvency probability of 99.9%. The asset correlation parameter ρ is defined dependent on the asset class.

The ASRF model's property of portfolio invariance ensures that the contribution of each asset to the UL capital of the portfolio requires the bank to estimate only three risk components as inputs to the risk-weight formula: the *exposure at default* (EAD), the *loss given default* (LGD), and the *probability of*

default (PD). For credit exposures with a time to maturity different from one year, UL_n from equation (6) is multiplied by a maturity adjustment factor, which requires the *effective time to maturity* as an additional parameter. These four risk components and the asset correlation parameter are further discussed in the following subsection.

The Risk Components EAD, LGD, and PD

The EAD is defined as the expected gross exposure upon default of the obligor and for on-balance sheet items as no less than the currently drawn amount. Credit conversion factors are applied in the case of noncash products such as undrawn credit commitments. In the advanced IRBA, banks use own estimates, whereas in the foundation IRBA, the draw-down factors of undrawn credit lines are given ([7], para. 474). Empirical and theoretical work on EADs in traditional bank lending is scarce. There appears to be some evidence that EAD decreases with credit quality. This may, however, also be the consequence of stricter limits and covenants for borrowers below investment grade [6].

LGDs are estimates of the loss severity of credit exposures upon default. They are estimated by the bank in the advanced IRBA and for retail exposures in the foundation IRBA as well. Mounting empirical evidence suggests that, at least in certain asset categories, loss severity can systematically fluctuate over time (see, e.g., [3] or [1]), inducing a correlation between default rates and recovery rates. Instead of incorporating this correlation in the ASRF model, banks have to account for systematic risk in their LGD estimates. Therefore, for exposures for which recovery rates are volatile over the economic cycle, the bank must estimate *downturn LGDs*, that is, estimates that are appropriate for an economic downturn if those are more conservative than the long-run average. In order to spare banks the estimation of two different LGDs, the downturn LGD is used in (6) also for subtracting EL, although EL depends in theory on the *unconditional* expectation of loss severity. For defaulted loans, banks have to compute a best-estimate LGD on the basis of current economic conditions and facility status ([7], para. 468 and 471).

The PD parameter measures the probability of the borrower entering default over a one-year horizon. The default event occurs either if the bank considers that the obligor is unlikely to pay its credit obligation

or if the obligor is past due more than 90 days ([7], para. 452). All banks adopting the IRBA must have an internal rating methodology through which every borrower is assigned a rating grade, which, in turn, determines the PD.

PD estimates of banks typically lie in a continuum between two methodological concepts: *point-in-time* PDs, which focus on current conditions and are often estimated from market prices, and *through-the-cycle* PDs, which account for the fluctuation of default rates over a business cycle. The exact definition of through-the-cycle PDs is controversial as are their pros and cons for regulatory capital purposes ([6], pp. 10–27 for a rigorous definition and discussion of different methodological concepts of internal ratings and PDs). Arguments in favor of point-in-time PDs are their consistency with a mark-to-market valuation of the portfolio and the easier validation owing to the shorter one-year horizon. Through-the-cycle PDs are, by construction, less volatile over time; this reduces cyclical effects of regulatory capital requirements. The Basel II framework gives banks some flexibility concerning their PD estimation, which mirrors the diversity of methodologies used in banks.

Asset Correlation

The asset-correlation parameter ρ in equation (6) fully determines the dependence between the asset value returns of two borrowers in the ASRF model and, in this way, the dependence structure of the portfolio. It is, therefore, by construction, a key driver of UL capital. During the development of Basel II, the original proposal of a flat asset correlation of 0.2 was replaced by the following function depending on the PD of the borrower:

$$\rho = \rho_{\max} - (\rho_{\max} - \rho_{\min})(1 - e^{-k PD_n}) \quad (7)$$

For wholesale lending, $\rho_{\max}$ and $\rho_{\min}$ are set to 0.24 and 0.12, respectively (with $k = 50$). For retail lending, the correlations are lower. Apart from reducing the steepness of the risk-weight curve, and, in this way, mitigating cyclical fluctuations of UL capital, the PD dependence of the asset correlation also reflects an on-average lower systematic/higher idiosyncratic risk of firms with a lower credit quality ([33] for empirical evidence).

In principle, the asset correlation can be estimated from historical data, such as stock returns, rating

migrations, or default rates. Empirical work on the estimation of asset correlations faces various limitations, for example, the scarcity of default events, missing traded equity for certain borrower types, and instability of correlations over time. As a consequence, it has produced quite diverse results (see **Correlation Risk** and for an overview of empirical results see [25]). Comparing the asset-correlation parameters in the risk-weighted functions with empirical correlation estimates is, however, misleading because the former are the result of the calibration of the ASRF model to credit risk portfolio models with a richer correlation structure, based on real credit portfolios (see the section Basel II Calibration and Procyclicality).

For small and medium enterprises in the corporate asset class, the asset-correlation parameter also increases with the size of the firm, which is measured by yearly turnover. The dependence on firm size reflects indicative empirical findings that systematic risk is higher for large firms than for small and medium enterprises (see, e.g., [18] or [33]).

Maturity Adjustments

The UL capital charge given by equation (6) captures only losses from default events over a one-year horizon. The market value of a loan with a longer maturity, however, is also sensitive to a change of the borrower's credit quality over this horizon. The maturity adjustment factor $M(T)$, which is a function of the loan's time to maturity T measured in years, accounts for the impact of migration risk by calibrating the model in a mark-to-market setting.

$$M(T) = \frac{1 + b(PD_n) \cdot (T - 2.5)}{1 - 1.5 b(PD_n)} \quad (8)$$

with

$$b(PD_n) = (0.11852 - 0.05478 \ln(PD_n))^2 \quad (9)$$

The adjustment factor $M(T)$ enters equation (6) as a multiplicative factor. It measures the relative change in required capital (per euro of the capital required for a 2.5-year loan) for a time to maturity greater than one year. For maturities between one and five years, $M(T)$ has been calibrated to a mark-to-market model. As a consequence, the capital figure obtained should be understood as the output of a mark-to-market model although the underlying model

structure of UL_n is that of a one-period default-mode model.

Since the maturity adjustment factor depends on the PD, the relative difference between capital charges for one and five years increases from 26% for a PD of 10% up to 242% for a PD of 0.03%. The size of the maturity adjustment has been criticized on the grounds that regulatory capital rules should focus on tail risk events rather than on migration risk. In fact, maturity adjustments quickly decrease for higher confidence levels as migration risk becomes less important [29]. Another criticism has been voiced from a macroprudential perspective. Relatively high values of the adjustment factor for long maturities and correspondingly low values for short maturities can provide incentives for short-term lending. Therefore, from a macro perspective, they can increase the risk of a credit crunch at times of stress and reinforce procyclical effects of the framework.

Credit Risk Mitigation

Compared with the previous Accord, the recognition of credit risk mitigation techniques has been greatly refined and expanded in Basel II, with respect to the methodology as well as the scope of eligible collateral. Two different methods are already available in the RSA to take account of risk mitigation. In the *simple approach*, the risk weights of the credit exposure are substituted by the risk weights of the protection provider. In the more risk-sensitive *comprehensive approach*, fluctuations in the value of the exposure as well as in the collateral are taken account of by the following formula, which defines the EAD after credit risk mitigation (E^*):

$$E^* = E(1 + H_e) - C(1 - H_c - H_{fx}) \quad (10)$$

- E : EAD without recognition of credit risk mitigation
- H_e : haircut for changes in the value of the credit exposure
- H_c : haircut dependent on the collateral
- H_{fx} : haircut for currency risk
- C : market value of the collateral.

The various haircuts in equation (10) are instrument dependent and either tabulated or can be estimated in more advanced approaches.

Credit derivatives and guarantees are typical instruments to address *counterparty credit risk* in

over-the-counter derivatives and security-financing transactions. This risk is driven by the joint effect of the uncertainty of the future exposure and the credit quality of the counterparty. A loss occurs only if the contract has a positive market value when the counterparty defaults; it also depends on the costs from replacing the contract by a similar transaction with a different counterparty. In this context, Basel II offers three approaches to address credit risk mitigation.

The most risk-sensitive approach is the *internal-model method*. In this approach, which is subject to regulatory approval, the EAD is the product of a multiplier and the *effective expected positive exposure (EPE)*. The effective EPE is defined as weighted average over time of effective expected exposures over the first year. Effective expected exposure, in turn, is the maximum expected exposure that occurs at a specific date or at a previous date ([7], para. 25–68).

In Basel II, the multiplier is set to 1.4. Subject to regulatory approval, it can also be estimated by the bank, recognizing a floor of 1.2. Applying a multiplier is important because the EPE can be interpreted as a loan-equivalent value only under certain assumptions: the portfolio needs to be infinitely granular and the exposures need to be mutually independent on a counterparty level and also from the counterparty's credit quality, that is, there is no wrong way risk. In addition, the EPE calculation assumes a static portfolio and does not consider *roll-over risk* from replacing transactions that mature before the one-year horizon. Canabarro et al. [15] and Wilde [40] explore the sensitivity of the multiplier against these assumptions.

Basel II Calibration and Procyclicality

The capital requirements under Basel II are the result of combining a bottom-up with a top-down calibration approach. In the bottom-up approach, appropriate risk weights for different asset classes are determined *relative* to each other to achieve consistency with banks' internal risk models. Figure 1 shows the risk-weight functions of selected asset classes. For a meaningful comparison, the LGD values represent average values from the fifth quantitative impact study of the Basel Committee [8]. Loans in asset classes with higher systematic risk and longer maturities generally carry higher risk weights for the

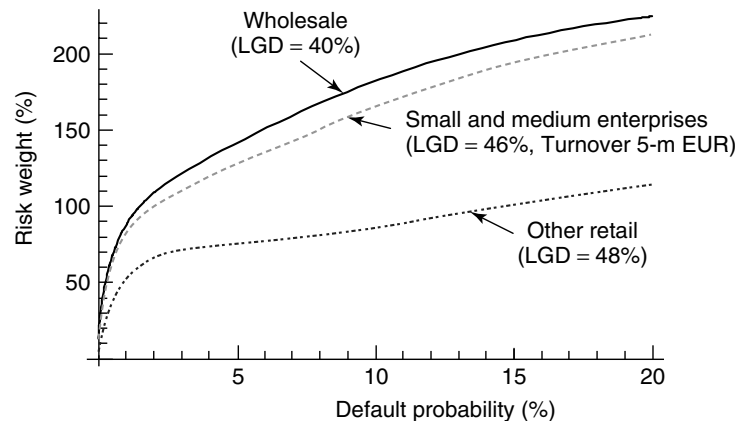


Figure 1 Risk-weight functions of different asset classes. This figure shows the risk-weight functions of three asset classes dependent on the probability of default. The loss given default (LGD) of every asset class is the average value for large, internationally active G10 banks, taken from the fifth quantitative impact study of the Basel Committee. The effective maturity for wholesale and small and medium enterprises is 2.5 years

same default probability (e.g., wholesale credit relative to retail credit).

Parallel to the bottom-up approach, the Basel Committee pursues a top-down approach with the goal to broadly maintain the level of capital requirements, while providing incentives for the more risk-sensitive approaches of the revised framework. The top-down approach was implemented by five quantitative impact studies in which banks were invited to compute the capital requirements for their current portfolios under the current accord as well as under the envisaged Basel II framework. One result of these exercises was the introduction of a scaling factor of 1.06 which is applied to risk weights for credit risk. Subsequent to the finalization of the Basel II framework, ongoing efforts have continued to monitor the level of capital requirements.

The dynamics of minimum capital requirements which are driven by the economic cycle, are not least as important as the level of capital. Since default rates and recovery rates are often systematically affected by the business cycle, banks adopting an IRBA might face higher capital requirements in an economic downturn. Stricter lending policies could then restrict the credit supply and generate negative feedback effects on the real economy, a phenomenon that is often described as *procyclicality*. In anticipation of the dynamic impact of the new IRBA, various mitigants have already been built into the new framework, for example, a flattening of the risk-weight functions

by a negative dependency of the asset correlation parameter on PD and the requirements to estimate downturn LGDs and to carry out procyclicality stress tests.

Empirical research has confirmed that fluctuations in regulatory capital requirements under the new framework are expected to be economically significant (see, e.g., [16] or [30]). Major empirical obstacles, for instance, uncertainties about the dynamic properties of PD and LGD estimates in real IRBA systems, have contributed to widely dispersed results. Furthermore, financial innovation, for example, the emergence of securitization and the originate-to-distribute model, has affected the traditional bank lending channel in ways that are difficult to predict. Procyclical effects, therefore, still pose a challenge for empirical research, risk management, and financial regulation.

Securitization Framework

Reducing the potential for regulatory capital arbitrage in securitization was a key target behind developing the new framework (see examples in [28]). For this purpose, regulators, in designing new rules, considered the level of capital held internally by banks for securitization exposures and set out to achieve capital neutrality between on-balance and off-balance sheet transactions as well as to ensure that capital charges for individual tranches capture the

relative distribution of risks across all tranches. At the same time, interfering in market segments where securitization was deemed a viable instrument of risk transfer needed to be avoided.

For banks following an IRBA, two methods are available in Basel II. If the tranches carry external ratings, they have to use the *ratings-based approach* (RBA). In this case, the risk weights are tied to ratings and account also for the granularity of the pool, measured by the effective number of exposures (or the inverse of the *Herfindahl—Hirschman Index*). If a specific tranche is unrated, its rating can be inferred from lower, more risky tranches.

Peretyatkin and Perraudin [35] compare the RBA capital charges with the model of Pykhtin and Dev [36, 37] on which they are based. They identify two simplified assumptions of the RBA that may have a significant effect on the capital charge: first, the missing differentiation between maturities and, second, not accounting for differences in correlation between the tranches exposures and exposures in the loan book.

The *supervisory formula approach* (SFA) provides an alternative for regulatory capital calculation if external ratings are not available and cannot be inferred. The SFA is based on a stylized model, similar to the ASRF model for traditional loans. If the required input data are unavailable, the exposure needs to be deducted from capital. In the special case of eligible liquidity facilities and credit enhancements in connection with ABCP (asset-backed commercial paper) programs, banks can determine the capital charge by their own internal assessment.

The conceptual underpinnings of the SFA are outlined in [23] and [22]. It strikes a balance between risk sensitivity and the tractability required for regulatory purposes. In the ASRF framework of the IRBA, the cumulative capital charge $K(\zeta)$ up to the enhancement level ζ is given by

$$K(\zeta) = \mathbb{E}[\min\{\zeta, L\} | X = q_X(\alpha)] \quad (11)$$

- ζ : share of the total amount of subordinate securitization exposures in the total
- : pool exposure
- L : book-value loss incurred in the pool as a share in the total pool exposure.

For a tranche of thickness Δ with enhancement level ζ , the capital in percent of the bank's exposure is

$$K(\zeta, \Delta) = \frac{K(\zeta + \Delta) - K(\zeta)}{\Delta} \quad (12)$$

By construction, the cumulative capital charge over all tranches equals K_{IRB} , which is defined as the sum of the UL capital from the ASRF model (6) and EL.

A direct application of the ASRF model to measure the risk of securitization tranches was not deemed as meaningful since the model implies relatively high capital charges for junior tranches, whereas at a certain level of protection the capital charge jumps to zero, depending also on the thickness of the tranche. This problem is solved by the *uncertainty in loss prioritization* assumption. By this assumption, the deterministic enhancement level ζ in equation (11) is replaced by the *effective* enhancement level $Z(\zeta)$. Specifying $Z(\zeta)$ as a Dirichlet process with parameter τ , the marginal distribution of $Z(\zeta)$ is $\text{Beta}(\tau \zeta, \tau(1 - \zeta))$. For $\tau \rightarrow \infty$, the model converges to the case of strict loss prioritization. The assumption that the exact level of protection of a tranche is uncertain can be motivated by the complexity of deal-specific contractual cash flows, often termed the *waterfall* of a securitization transaction, which cannot be explicitly modeled in the interest of tractability. In addition to the single risk factor assumption shared by the ASRF model, the SFA model assumes that the collateral pool is homogeneous and that LGDs are pairwise independent conditional on the systematic risk factor X and subject only to idiosyncratic risk.

Computing the SFA-based capital requirement for a tranche requires five inputs: two tranche-specific figures from the bank, namely, the nominal enhancement level ζ and the tranche thickness Δ , and three collateral pool-specific input values, namely, the *effective* number of exposures N , the exposure-weighted average expected *LGD*, and K_{IRB} . Since real portfolios are heterogeneous in exposure size, N is concentration-adjusted, that is, the inverse of the *Herfindahl—Hirschman Index*. To the extent that K_{IRB} contains maturity adjustments, maturity is also captured by the SFA. The parameter τ of the Beta distribution is set to 1000 in Basel II ([7], para. 623–634).

The model developed by Gordy and Jones assumes that a bank's exposure to every securitization is a negligible part of its total portfolio and that the collateral pool contains only loans or bonds but not that the collateral pool is infinitely fine grained. As an extension of the ASRF model, it does not capture the risk concentration if the same obligor is also part of other securitization transactions or of the traditional loan portfolio.

A level of conservatism is introduced by requiring deduction from capital for any protection level up to K_{IRB} . Furthermore, even for the most senior tranches, the capital charge cannot fall below the floor of 0.56%. Figure 2 shows the cumulative capital given by the SFA depending on the enhancement level ζ for collateral pools of 200, 50, and 5 exposures. Up to K_{IRB} , the capital charge increases almost linearly and afterward flattens out—the faster the flattening out, the less granular the collateral pool is.

Substantial losses were encountered from structured products in the wake of the 2007 subprime turbulence, particularly for previously highly rated tranches. This outcome supports allocating more capital for senior tranches than implied by direct application of the IRB model. In answer to the 2007 subprime turbulence, minimum capital requirements for securitizations were strengthened, particularly for *res securitizations*, that is, securitizations in

which at least one of the underlying exposures is already a securitization exposure itself, because of their higher systematic risk.

Capital Requirements for Other Risks in Pillar 1

Operational Risk

In Basel II, operational risk is, for the first time, a separate risk category subject to regulatory minimum capital requirements. It is defined as the risk of monetary losses “resulting from inadequate or failed internal processes, people and systems or from external events” ([7] para. 644.). It includes losses from failures in IT systems, fraud, natural catastrophes, and legal risk, but excludes losses from strategic and reputational risks. Ongoing developments in the financial industry, for example, the propagation of securitization and an increased reliance on rapidly evolving technology and complex financial products and strategies, support a new regulatory capital charge for operational risk.

Similar to other risk categories and consistent with the evolutionary concept of Basel II, different approaches characterized by increasing complexity and risk sensitivity are available to determine the regulatory capital charge for operational risk. The most advanced and technically most challenging method is

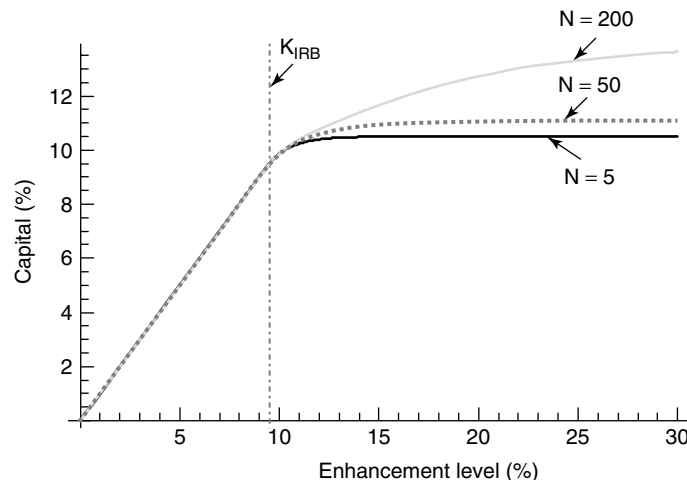


Figure 2 Capital requirements based on the supervisory formula. This figure shows the capital charge for securitizations under the supervisory formula approach for homogeneous collateral pools of $N \in \{200, 20, 5\}$ exposures. The default probability of every asset in the homogeneous pool is 2% and the loss given default 50%

the *advanced measurement approach*, which allows banks, subject to regulatory approval, to apply their own internal models for the calculation of regulatory capital (see **Operational Risk**). The resulting capital charge must capture potentially severe “tail” loss events. It needs to consider both EL and UL, with the UL defined for a one-year time horizon and a 99.9% confidence interval.

The loss distribution approach seems to have emerged as the most common modeling framework (see [4], as an example). Loss severity and loss frequency are typically modeled separately. Losses are aggregated over loss event types (e.g. external fraud or business disruptions and system failures) and business lines (e.g., corporate finance or retail banking). The Value at Risk figures of different business lines must be added, although banks can take into account correlation effects subject to regulatory approval.

Development of models for operational risk has been fast. *Extreme value theory* (EVT) is often applied, but these methods are still under development (see [20] for a critical examination of the use of EVT for operational risk modeling). The complex methods typically require an extensive set of historical loss data. The relative scarcity of high-impact, low-frequency events in individual banks has prompted the construction of external data bases. The outstanding methodological challenges and data restrictions in the modeling of operational risk are arguments in favor of complementing the quantitative risk measurement methodology with a self assessment, using, for example, scenario analyses and scorecards, a distinct pillar 2 review, and market discipline.

Market Risk

The 1996 amendment for market risk, which covered interest rate risk and equity risk in the trading book as well as foreign exchange and commodity risk, was integrated in Basel II. Since then, it has been possible to base the capital charge for market risk either on a standardized measurement method or – subject to regulatory approval – on the bank’s internal model for market risk (see **Market Risk**). Banks adopting the more risk-sensitive internal models approach are free to choose their own methodology, but to be recognized by the regulator, the internal

model has to meet various qualitative and quantitative requirements, including a regular backtesting program ([7], para. 718(LXX)–718(XCIX), and also see **Backtesting**).

The capital charge in the internal models approach is defined as follows:

$$MRC_t = \text{Max} \left\{ k \frac{1}{60} \sum_{i=1}^{60} VaR_{t-i}, VaR_{t-1} \right\} + SRC_t \quad (13)$$

where VaR_{t-1} is the Value at Risk of the portfolio over a 10-day horizon at a 99% confidence level ([7], para. 718). The second term, VaR_{t-1} , becomes binding if the portfolio composition has substantially changed.

The factor k in the first term of equation (13) is the sum of two components:

1. a multiplicative factor set by the local regulator with a floor of 3 that captures *inter alia* model risk, which stems, for example, from the assumptions of normality and stability of the distributions;
2. a *plus factor* that depends on the *ex post* performance, that is, the *backtesting* results, of the model.

The plus factor is based on mandatory quarterly backtesting with trading results of the last 250 days. The size of the plus factor depends on the number of Value at Risk exceptions and is given by a traffic-light table ([7], Annex 10, para. 27–61).

Various papers have advocated different backtesting methodologies (e.g., [31] or [13]). Accepting the imperfect nature of signals from backtesting results, supervisors have substantial leeway in how to assess the backtesting methodology of banks in their jurisdiction, for example, concerning the choice between a Value at Risk measure based on a “hypothetical profit and loss”, assuming a static portfolio, or on the “actual profit and loss”, which includes also commissions and intraday trading.

The *specific risk charge* SRC_t in equation (13) provides a buffer against idiosyncratic risk, including default risk and event risk. For banks that do not have methodologies in place to adequately capture event and default risks, a multiplicative factor of 4 instead of 3 applies. If the model captures only generic market risk, SRC_t is computed by the standardized approach.

The risk measurement approach of the 1996 amendment was challenged by financial innovation, which caused a shift of credit exposures from the banking book into the trading book when pools of traditional loans were packaged as tranches of collateralized debt obligations (CDO) and sold in the market. Although the risk of a credit exposure is not affected by transferring it between banking book and trading book, the risk measurement framework becomes different. Not only the different considered risk factors but also the shorter risk horizon and the lower confidence level drive a wedge between capital charges for the same exposure dependent on its assignment to the banking book or the trading book. As a consequence, making credit risk tradable by securitization created rich opportunities for regulatory capital arbitrage.

Recognizing the inherent risks of this development, the Basel Committee already decided in 2005 to reexamine the trading book treatment and issued two consultative documents with proposals for a revision in January 2009. An *incremental risk charge* is introduced for the internal models approach, which refines the treatment of specific risk and reduces the gap between regulatory capital charges in the banking book and in the trading book. It includes default risk and credit migration risk and also takes account of correlations between default events and migration events, which cause a clustering of both. Widening the definition beyond default risk alone reflects experience from the 2007 subprime turbulence that material credit-risk-related losses occurred without actual default events. Securitization positions cannot be incorporated into the incremental risk charge model on the grounds that a reliable quantification of diversification benefits is not possible given the current state of risk modeling. A standardized method applies instead.

Furthermore, a *stressed Value at Risk* is suggested that needs to be added to the Value at Risk in equation (13). For this purpose, a bank has to calibrate its model to historical data from a period of significant financial stress relevant to its portfolio. In summary, with the revised treatment of the trading book, the Basel Committee has addressed key limitations of previous capital requirements for the trading book and also incorporated lessons from the 2007 subprime turbulence.

Supervisory Review Process and Validation of Credit Risk Components

Whereas pillar 1 is concerned with regulatory minimum capital requirements, pillar 2 is about the internal capital adequacy assessment process. A bank has to ensure through its internal capital assessment that its capital basis is commensurate with its risk profile. The internal capital assessment has to address all material risks not captured in pillar 1, notably interest rate risk in the banking book, liquidity risk, and concentration risk. Banks' assessment of capital adequacy is evaluated in the supervisory review and evaluation process. This section is focused on two aspects that may require complex quantitative models, namely, the measurement of concentration risk in credit portfolios and stress testing.

Concentration Risk

Since lending is the primary activity of most banks, credit risk concentrations often present the most material risk concentrations ([7], para. 771). They are particularly important as they are not captured by the ASRF model, which is the basis of the regulatory capital requirements for credit risk in pillar 1. In a real bank portfolio with a finite number of assets, two of its assumptions, namely, that systematic risk is fully captured by a single factor and that the portfolio is infinitely granular, are not fulfilled. Deviations from this ideal case create *concentration risk*, which is not a different risk category but rather a key driver of credit risk. It can be decomposed into *single-name concentration* and *sectoral concentration*. Single-name concentration refers to an unbalanced exposure distribution across borrowers, whereas sectoral concentration refers to exposure concentration either in regional sectors or industry sectors ([9] gives an overview of various measurement tools for both kinds of risk concentrations).

Single-name concentration can generally be diversified away by increasing the number of exposures. A *granularity adjustment* for regulatory capital requirements to address single-name concentration was considered but not included in Basel II. Its computational burden was later significantly reduced by refined proposals in [34] and [24].

In the special case of a homogeneous portfolio (in terms of the risk components that enter the IRBA risk weight function), the granularity adjustment in

[24] computed from the loan volumes is linear in the Herfindahl—Hirschman Index. This gives a theoretical motivation for this popular but otherwise *ad hoc* concentration measure. *Ad hoc* measures generally use only information on exposure size in the portfolio. Model-based approaches consider also default probabilities and correlations and allow a risk quantification in terms of required capital. Therefore, model-based approaches have a higher informational value than *ad hoc* indicators.

Accounting for sectoral concentrations is technically more demanding than single-name concentrations since they drive the systematic risk through cross-sectoral default dependencies. At the same time, they are more relevant for typical credit portfolios of banks than name concentrations [19]. The asset-correlation parameter in the IRBA risk-weight functions was calibrated as an average correlation value, appropriate for portfolios that are very well diversified across sectors. Without further knowledge on this calibration procedure, it is, therefore, not possible to compute an adjustment of the ASRF model with as much rigor as the granularity adjustment in [24]. Contrary to the ASRF model, standard portfolio models for credit risk automatically account for risk concentrations. Therefore, the internal capital adequacy assessment process should be well suited to ensure that risk concentrations are adequately captured by capital.

Stress Tests

The inability to capture extreme events or “tail risk” is a commonly known shortcoming of standard risk models, which was confirmed by the market disruptions in the wake of the 2007 subprime turbulence. Therefore, the transition to model-based capital rules has increased the importance of stress tests (*see Stress Testing*). Initially introduced as compulsory in the 1996 amendment for internal market risk models, market risk is still the area in which stress testing methodologies are most advanced (see, e.g., [2, 12], and [32]).

Stress tests of credit portfolios are an essential element of the new framework. Stress scenarios should consider economic downturns, market-risk events, and liquidity conditions ([7], para. 434–437). They need to be integrated into the risk management framework of the institution and, therefore, they are typically based on institution-specific stress scenarios and

use internal risk models. At least one out of different components of a credit risk model are usually stressed: PD, LGD, volatility and correlation parameters, or the systematic risk factors. Feeding “stressed” PDs or LGDs into the ASRF model is a common approach that provides valuable information on the potential impact of the stress scenario on regulatory capital requirements. Its informational value for risk management decisions, however, is constrained by the stylized structure of the ASRF model, for example, its insensitivity to risk concentrations.

Stress tests of credit portfolios can serve three different purposes. First, they allow to assess the capital adequacy in selected stress scenarios. Second, they can help validate a model. Finding, for example, that a model does not react meaningfully to stressed conditions could be a warning signal. Third, they can help dampen the time variation of minimum capital requirements and to mitigate procyclical effects. This purpose can be achieved if stress tests are used to determine a dynamic capital buffer that is accumulated under benign and run down under adverse economic conditions.

Stress tests are typically either carried out for a broad economic scenario affecting the total portfolio or with stress scenarios confined to one or few industry sector(s). Particularly in the first case, the link between the credit risk drivers, modeled at the micro or portfolio level and the macroeconomic level, still poses a significant methodological and data challenge. Stress tests with sector-specific or borrower-specific scenarios can be useful to detect *hidden risk concentrations* in credit portfolios, which affect portfolio risk through elevated default correlations between certain (groups of) borrowers [14].

Outstanding Challenges to Future Regulation

Risk measurement in the context of future regulation faces at least three major challenges that have also gained in prominence in the 2007 subprime turbulence. The first is the need to broaden the scope of covered risks, in particular, by giving more attention to liquidity risk. The second is to further improve the risk assessment methodology, for example, by better taking into account linkages between different risk categories. The third is the need to strengthen a system-oriented *macroprudential* approach to bank regulation.

The first challenge pertains to liquidity risk [10], which played a prominent role in the 2007 subprime turbulence. An imminent shortage in funding liquidity due to unexpectedly high liquidity demands from liquidity facilities granted to special purpose vehicles is a case in point.

As a second challenge, the 2007 subprime turbulence uncovered weaknesses in modeling the economic risk of specific financial instruments as well as between different risk categories. One example is the collective increase of market risk, credit risk, and liquidity risk, which suggests a more integrated treatment in internal capital modeling [11].

The third major challenge is a more prominent role of a macroprudential perspective in regulation. Strategic interactions among market participants, for example, herding behavior, played an important role when the 2007 subprime turbulence unfolded into a major financial crisis. Their consequences and the resulting endogeneity of risk are not well captured either by regulatory minimum capital requirements or by statistical risk models currently used in financial institutions. Potential procyclical effects of risk-sensitive capital rules can provide another example that the micro- and the macroprudential perspectives do not necessarily coincide. Hence, supplementing regulation with a stronger macroprudential dimension still poses a formidable challenge.

End Notes

^aThe views expressed herein are my own and do not necessarily reflect those of the Deutsche Bundesbank.

References

- [1] Acharya, V.V., Bharath, S.T. & Srinivasan, A. (2007). Does industry-wide distress affect defaulted firms?: Evidence from creditor recoveries, *Journal of Financial Economics* **85**(3), 787–821.
- [2] Alexander, C. & Sheedy, E. (2008). Developing a stress testing framework based on market risk models, *Journal of Banking and Finance* **32**(10), 2220–2236.
- [3] Altman, E.I., Resti, A. & Sironi, A. (2005). The link between default and recovery rates: theory, empirical evidence and implications, *Journal of Business* **78**(11), 2203–2228.
- [4] Aue, F. & Kalkbrener, M. (2006). LDA at work: Deutsche Bank's approach to quantifying operational risk, *Journal of Operational Risk* **1**(4), 49–93.
- [5] Basel Committee on Banking Supervision (1999). *Credit Risk Modelling: Current Practices and Applications*, technical report, <http://www.bis.org/publ/bcbs49.htm>
- [6] Basel Committee on Banking Supervision (2005). *Studies on the Validation of Internal Rating Systems*, working paper, No. 14, http://www.bis.org/publ/bcbs_wp14.htm
- [7] Basel Committee on Banking Supervision (2006). *International Convergence of Capital Measurement and Capital Standards*, A revised framework, Comprehensive version, www.bis.org/publ/bcbs128.pdf
- [8] Basel Committee on Banking Supervision (2006). *Results of the Fifth Quantitative Impact Study (QIS 5)*, <http://www.bis.org/bcbs/qis/qis5.htm>
- [9] Basel Committee on Banking Supervision (2006). *Studies on Credit Risk Concentration: An Overview of the Issues and a Synopsis of the Results from the Research Task Force project*, Working Paper, No. 15, http://www.bis.org/publ/bcbs_wp15.htm
- [10] Basel Committee on Banking Supervision (2008). *Principles for Sound Liquidity Risk Management and Supervision*, June, <http://www.bis.org/publ/bcbs138.htm>
- [11] Basel Committee on Banking Supervision (2009). *Findings on the interaction of market and credit risk*, Working paper, No. 16, http://www.bis.org/publ/bcbs_wp16.htm
- [12] Berkowitz, J. (1999). A coherent framework for stress-testing, *Journal of Risk* **2**(2), 1–11.
- [13] Berkowitz, J. (2001). Testing density forecasts, With applications to risk management, *Journal of Business and Economic Statistics* **19**(4), 465–474.
- [14] Bonti, G., Kalkbrener, M., Lotz, C. & Stahl, G. (2006). Credit risk concentrations under stress, *Journal of Credit Risk* **2**(3), 115–136.
- [15] Canabarro, E., Picoult, E. & Wilde, T. (2003). Analysing counterparty risk, *Risk Magazine* **16**(9), 117–122.
- [16] Catarineu-Rabell, E., Jackson, P. & Tsomocos, D.P. (2005). Procyclicality and the new Basel Accord—Banks' choice of loan rating system, *Economic Theory* **26**(3), 537–557.
- [17] Danielsson, J. (2002). The emperor has no clothes: Limits to risk modelling, *Journal of Banking and Finance* **26**(7), 1273–1296.
- [18] Dietsch, M. & Petey, J. (2004). Should SME exposures be treated as retail or corporate exposures? A comparative analysis of probabilities of default and asset correlations in French and German SMEs, *Journal of Banking and Finance* **28**(4), 773–788.
- [19] Duellmann, K. & Masschelein, N. (2007). A tractable model to measure sector concentration risk in credit portfolios, *Journal of Financial Services Research* **32**(1–2), 55–79.
- [20] Dutta, K. & Perry, J. (2006). *A Tale of Tails: An Empirical Analysis of Loss Distribution Models for Estimating Operational Risk Capital*, Federal Reserve Bank of Boston working paper, No. 06/13 <http://www.bos.frb.org/economic/wp/wp2006/wp0613.pdf>
- [21] Gordy, M. (2003). A risk-factor model foundation for ratings-based bank capital rules, *Journal of Financial Intermediation* **12**(3), 199–232.

- [22] Gordy, M. (2004). Model foundations for the supervisory formula approach, in *Structured Credit Products: Pricing, Rating, Risk Management and Basel II*, W. Perraudin, ed, Risk Books, pp. 307–328.
- [23] Gordy, M. & Jones, D. (2003). Random tranches, *Risk Magazine* **16**(3), 78–81.
- [24] Gordy, M. & Lütkebohmert, E. (2007). *Granularity Adjustment for Basel II*, Deutsche Bundesbank discussion paper, Series 2, No. 1.
- [25] Grundke, P. (2008). Regulatory treatment of the double default effect under the New Basel Accord: How conservative is it? *Review of Managerial Science* **2**(1), 37–59.
- [26] Hirtle, B.J., Levonian, M., Saidenberg, M., Walter, S. & Wright, D. (2001). Using credit risk models for regulatory capital: issues and options, *Federal Reserve Bank of New York Economic Policy Review* **7**(1), 19–36. Research Paper Series - Staff Report.
- [27] Jackson, P., Perraudin, W. & Saporta, V. (2002). Regulatory and “economic” solvency standards for internationally active banks, *Journal of Banking and Finance* **26**(5), 953–976.
- [28] Jones, D. (2000). Emerging problems with the Accord: Regulatory capital arbitrage and related issues, *Journal of Banking and Finance* **24**(1), 35–58.
- [29] Kalkbrener, M. & Overbeck, L. (2002). The maturity effect on credit risk capital, *Risk Magazine* **15**(7), 59–63.
- [30] Kashyap, A.K. & Stein, J.C. (2003). *Cyclical Implications of the Basel II Capital Standards*, Federal Reserve Bank of Chicago, Economic Perspectives, 1Q, pp. 18–31.
- [31] Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models, *Journal of Derivatives* **2**, 73–84.
- [32] Kupiec, P. (1998). Stress testing in a value at risk framework, *Journal of Derivatives* **6**(1), 7–24.
- [33] Lopez, J. (2004). The empirical relationship between average asset correlation, firm probability of default and asset size, *Journal of Financial Intermediation* **13**(2), 265–283.
- [34] Martin, R. & Wilde, T. (2002). Unsystematic credit risk, *Risk Magazine* **15**(11), 123–128.
- [35] Peretyatkin, V. & Perraudin, W. (2004). *Structured Credit Products*, in *Capital for Structured Products*, W. Perraudin, ed, Risk Books, London, pp. 329–362.
- [36] Pykhtin, M. & Dev, A. (2002). Credit risk in asset securitizations: analytical model, *Risk Magazine* **15**(5), 16–20.
- [37] Pykhtin, M. & Dev, A. (2003). Coarse-grained CDOs, *Risk Magazine* **16**(1), 113–116.
- [38] Schönbucher, P.J. (2001). Factor models: portfolio credit risks when defaults are correlated, *Journal of Risk Finance* **3**(1), 45–56.
- [39] Vasicek, O.A. (1997). *The Loan Loss Distribution*, technical report, KMV Corporation.
- [40] Wilde, T. (2005). Analytic methods for portfolio counterparty risk, in *Counterparty Credit Risk Modelling*, M. Pykhtin, ed, Risk Books, pp. 189–225.

Related Articles

**Backtesting; Market Risk; Operational Risk;
Risk Exposures; Stress Testing; Value-at-Risk.**

KLAUS DUELLMANN

Economic Capital

Role of Capital in Financial Institutions

In contrast to a typical corporation, the key role of capital in a financial institution is not primarily one of providing a source of funding for the organization.^a For example, banks usually have ready access to funding through their deposit-taking activities, which can be increased fairly fluidly. Instead, the primary role of capital in a bank, apart from the transfer of ownership, is to act as a buffer to

- absorb large unexpected losses;
- protect depositors and other claim holders; and
- provide confidence to external investors and rating agencies on the financial health and viability of the firm.

Typically, the sources of financial risk are classified as (*see Risk Exposures*)

- *credit risk*—losses associated with the default (or credit downgrades and upgrades) of an obligor (a counterparty, borrower, or debt issuer);
- *market risk*—losses associated with changes in market values; and
- *operational risk*—losses associated with operating failures.

Other types of risk faced by a financial institution include *liquidity risk*, *business* (or *strategic*) *risk*, *legal risk*, *insurance* and *liability risks*, etc. Capital provides a metric for aggregating risks across both different asset classes and across different risk types.

We can broadly classify capital into three types:

- *Book capital*—the actual physical capital held. While in its strictest definition this should be simply *equity capital*, more generally this might also include other assets such as liquid debt or hybrid instruments.
- *Regulatory capital*—the capital that a bank is required to hold by regulators in order to operate; this is defined by the regulatory authorities to act as a proxy for economic capital (EC).
- *Economic capital*—an estimate of the minimum level of capital that a firm requires to operate its business with a desired target solvency level. Sometimes this is also referred to as *risk capital*.^b

EC is meant to reflect the true “fair market” value differential between assets and liabilities. In contrast, regulatory capital may be commonly defined in terms of accounting book value measures rather than market values.^c The objectives of regulatory capital requirements are to safeguard the security of the financial system as a whole and to ensure its ongoing viability, as well as to create a level playing field across all institutions.

Economic Capital Definitions

Economic Capital as a Buffer for Unexpected Losses

EC represents the minimum buffer that provides protection against all risks faced by an institution at a target confidence level. Since it would be too costly to cover losses at the 100% level, EC is set at a lower confidence level (e.g., 99.9%). The choice of this level involves a trade-off between providing high returns on capital for shareholders and providing protection to the debt holders and other claim holders, such as depositors, while achieving a desired rating. For example, a firm’s probability of default over one year should be at most 0.5% to achieve a BBB rating by S&P. Thus, it must have enough capital to sustain a “0.5% worst-case loss” over a one-year time horizon.

The approach commonly taken by practitioners is to define EC to absorb only *unexpected losses* up to a certain confidence level, with *reserves* set aside to absorb *expected losses* (EL). Thus, EC is estimated as the *Value at Risk* (VaR) (*see Value-at-Risk; Market Risk*) at an $\alpha\%$ confidence level less the expected loss (*see Figure 1*):

$$EC_{\alpha} = VaR_{\alpha} - EL \quad (1)$$

Subtracting *EL* from the worse-case losses represents a simplifying approximation to estimate *EC* and generally leads to conservative estimates. More precisely, *EC* should measure loss relative to the portfolio’s initial mark-to-market (MtM) value and *not* relative to the *EL* of its end-of-period distribution, and should account for the interest payments that must be made on the funding debt:

$$EC = A_0 - \frac{VaR_{\alpha}(A_1)}{(1 + r_D)} \quad (2)$$

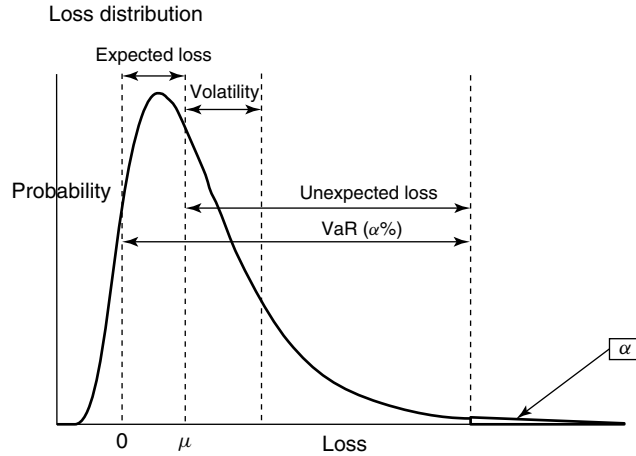


Figure 1 Loss distribution and capital

where A_0 is the market value of the assets at time 0, and r_D the nominal return on the liabilities. The estimation of the interest compensation factor and of the MtM value of the portfolio generally requires the use of an asset pricing model.^d

Economic Capital or “Risk Capital” as Insurance for the Value of the Firm

Some economists have used the term *risk capital* to refer to the smallest amount that can be invested to insure the value of the firm’s net assets against loss in value relative to a risk-free investment [4, 5]. For example, the risk capital of a long US Treasury bond position is the value of a *put option* with strike equal to the forward price of the bond.

Economic Capital as a Management Tool

EC provides a consistent metric to determine risk aggregation, asset and business allocation, as well as performance measurement.

In addition to computing the total EC for a portfolio, it is important to *attribute* this capital *a posteriori* to various subportfolios such as the firm’s activities, business units, counterparties or even individual transactions, and to *allocate* it *a priori* in an optimal manner, to maximize risk-adjusted returns (see **Economic Capital Allocation**). EC provides a useful basis for *risk-adjusted performance measurement* (RAPM), also referred to as *risk-adjusted return on*

capital (RAROC) (see **Risk-adjusted Return on Capital (RAROC)**) since it defines a consistent metric that spans all asset and risk classes, thereby providing an “apples-to-apples” benchmark for evaluating the performance of alternative business opportunities. Each asset can be assessed on a consistent basis, with returns adjusted appropriately in the context of the amount of risk assumed.

Economic Capital Calculation

Methods for calculating EC can be classified as *top-down* or *bottom-up* approaches. In addition to statistical methods, practitioners also commonly use *stress-testing* methods (see **Stress Testing**).

Top-down Approaches

EC can be estimated based on the aggregate information of the firm’s performance. Such “top-down approaches” generally use information on either *earnings* or *stock prices*.

The earnings-based approach typically makes the simplifying assumption that the market value of capital is equal to the value of a perpetual stream of expected earnings. As the determination of EC is based on the ability to sustain a worst-case loss at a given confidence level, this leads to

$$EC_0 = EaR/k \quad (3)$$

where EaR is the difference between expected earnings and the earnings under the worst-case scenario at a confidence level, and k is the required return associated with the riskiness of the earnings. Often, this approach is used with the additional assumption that earnings are normally distributed; thus, the confidence level is determined as a multiple of the volatility (see [7]).

Limitations of the earnings approach include its requirement of reliable historical performance data, and its inability to link EC directly to the sources of risk.

An alternative top-down approach is based on the *Black-Scholes-Merton* option pricing framework (see **Structural Default Risk Models**). It assumes that the market value of capital can be modeled as a *call option* on the value of the firm's assets where the strike is the notional value of the debt. An advantage of this approach is the availability of stock market data. However, several simplifications regarding the capital structure and model assumptions must be made to apply this tool in practice. A key limitation is that it does not allow the allocation of capital into different risks or business lines.

Bottom-up Approaches

EC is estimated by modeling the individual transactions and businesses and then aggregating the risks using advanced statistical portfolio models and stress testing. The bottom-up approaches have now become the best practice and provide greater transparency in isolating risks and allocating capital.

The estimation of EC at the firm level requires consolidation of risks at two levels.

- First, one computes the market risk, credit risk, operational risk, and so on, at the enterprise level. To achieve this, a firm might use separate VaR models for market risk, credit risk, and operational risk.
- Then, at the second level, the firm consolidates the capital across these risks.

A simple method to aggregate capital in the second step is to add up the credit risk capital, market risk capital, and operational risk capital. This produces a conservative capital measure (essentially, assuming that the risks are perfectly correlated). More advanced approaches apply *copula functions* (see **Copulas**:

Estimation) to aggregate risks, and many firms devote considerable effort to measure the correlations between these risks.

Stress Testing and Economic Capital

In addition to the VaR-based approaches, practitioners usually use stress testing as an important part of their EC methodology. Commonly, this involves the development of several specific adverse extreme scenarios and the assessment of portfolio losses against them. Stress scenarios may be based on historical experience or management judgment.

The combination of stress testing and statistical measures to develop EC measures is generally *ad hoc*, largely based on management's judgment. For example, an institution might assign the EC for market risk as 50% times the 99% VaR plus 50% the loss outcome from some stress scenarios (see **Stress Testing**).

End Notes

^aFor a more detailed discussion of economic capital, see [1, 6] and the references cited therein.

^bAlternatively, Matten [3] defines $EC = \text{risk capital} + \text{goodwill}$.

^cIn some cases, accounting practice allows items to be reported on a market value basis.

^dThis equation is derived by assuming that assets A_0 are funded by debt, L_0 , and equity, E_0 : $A_0 = L_0 + E_0$. The amount owed at maturity will be $L_0 \times (1 + r_D) = (A_0 - E_0) \times (1 + r_D)$. To match a given default probability α , EC is E_0 chosen so that this expression matches $\text{VaR}_\alpha(A_1)$, the α percentile of the final asset value. For a detailed discussion, see [1, 2].

References

- [1] Aziz, A. & Rosen, D. (2004). Capital allocation and RAPM, *Professional Risk Manager (PRM) Handbook*, Chapter III.0, PRMIA Publications.
- [2] Kupiec, P. (2002). *Calibrating Your Intuition: Capital Allocation for Market and Credit Risk*, Working paper 99/02, IMF, at <http://www.imf.org/external/pubs/ft/wp/2002/wp0299.pdf>.
- [3] Matten, C. (2000). *Managing Bank Capital*, Wiley, Chichester.
- [4] Merton, R.C. & Perold, A.F. (1993). Theory of risk capital in financial firms, *Journal of Applied Corporate Finance* 6(Fall), 16–32.

- [5] Perold, A.F. (2005). Capital allocation in financial firms, *Journal of Applied Corporate Finance* **17**(3), 110–118.
- [6] Rosen, D. (2004). Credit risk capital calculation, *Professional Risk Manager (PRM) Handbook*, Chapter III.B.6, PRMIA Publications.
- [7] Saita, F. (2004). Measuring risk-adjusted performances for credit risk, in *Frontiers of Performance Measurement in Banking and Finance*, G. de Laurentis, ed, Egea.

Related Articles

Copulas; Estimation; Economic Capital Allocation; Structural Default Risk Models; Risk-adjusted Return on Capital (RAROC); Stress Testing; Value-at-Risk.

DAN ROSEN & DAVID SAUNDERS

Risk-adjusted Return on Capital (RAROC)

Maintaining credit worthiness is a cost of doing business for a bank. In a multibusiness bank, senior management faces the problem of allocating capital to the different businesses in order to maximize the market value of the bank to existing shareholders, subject to the constraint that it maintain its desired level of creditworthiness. The capital structure of the bank is set so that the bank's equity provides a buffer for the bank to maintain its desired level of creditworthiness. We refer to this buffer as the *economic capital* (see **Economic Capital**) of the bank. Traditional approaches to capital budgeting define the risk of a project in terms of a project's systematic risk with market factors. Such a measure does not address the concerns of senior management about how the project affects the overall risk of the bank. If a project increases the probability of default, to restore the bank's creditworthiness, the bank must increase its economic capital. The cost of the marginal increase in the economic capital should be allocated to the project and an adjusted net present value determined. The value additivity principle will not hold, as economic capital is not additive, as noted by Merton [8] (see **Convex Risk Measures; Expected Shortfall**).

In determining whether to add or eliminate a business from a bank's portfolio, a marginal approach is often advocated for determining the economic capital. The standard assumption is that the business is "small" in comparison with the existing portfolio. If this is not the case, the whole nature of the bank and its capital structure are affected, implying that the determination of the marginal effects is problematic. Merton and Perold [9] show that simply aggregating the marginal economic capital of individual businesses underestimates the total economic capital and the error can be substantial. Apart from these theoretical concerns, the actual determination of the marginal capital is fraught with many practical difficulties.

Each business within the bank contributes to the overall credit risk of the bank. The economic capital of the bank is less than the aggregate economic capital of the individual businesses on a stand-alone basis, owing to the fact that businesses are correlated.^a We

refer to the difference between the economic capital of the bank and the aggregate stand-alone value as the *portfolio effect*. This raises a number of challenging issues. The first is the measurement of economic capital. If economic capital is estimated on a stand-alone basis, does the bank own the portfolio effect or individual businesses? If businesses, how should economic capital be allocated? Given the allocation of capital to each business within the bank, how should it measure the performance of each business, recognizing that each business affects the overall creditworthiness of the bank and the cost of funding? What benchmark should be used to judge the performance of a business? How should the performance of a business be measured, if it does not fully utilize its allocated capital? These questions hint at some of the difficulties in measuring the performances of the businesses within the bank.

Practitioners have addressed the issue of an appropriate performance metric with the use of a risk-adjusted rate of return on capital (RAROC).^b A generic RAROC takes the form

$$RAROC = \frac{ER - EL - Costs}{EC} \quad (1)$$

where ER denotes the expected revenue, EL the expected losses, $Costs$ the funding cost associated with the project, and EC the economic capital attributed to the project.^c The economic capital is usually defined as the capital necessary to cushion against unexpected losses, operational, and market risks, and is often referred to as a *Value at Risk* (see **Value-at-Risk; Market Risk**). There are both practical and theoretical issues with the use of this type of metric. The first is the myopic nature of the measure. The second is the issue of how to measure economic capital for a business. The third is what benchmark to use to judge whether the risk-adjusted return is acceptable. The fourth is that economic capital is not a sufficient statistic for the measurement of risk. It is known from the work of Wilson [17], Froot and Stein [4], and Crouhy *et al.* [2] that economic capital does not incorporate the systematic risk of a business and serious errors can occur if projects are solely ranked on the basis of RAROC.^d

Froot and Stein (FS) [4] assume that there are no agency issues—the management acts in the best interests of existing shareholders. In their three-period model, risk management arises endogenously from

the need to avoid an adverse selection problem associated with costly external financing in the last period. They show that the expected excess rate of return for a new project depends on two terms. The first term depends on the covariance of the cash flow with the market portfolio and the second depends on the covariance of the project's cash flow with the nontradable cash flows of the bank multiplied by the price of nontradable risk. As FS observe, it is not clear how to estimate this term.

Banks are opaque to outsiders. Any outside assessment of the creditworthiness of a bank is out of date even when issued. Banks are translucent if not opaque to insiders, with information flows restricted across businesses inside the bank and even within businesses.^c Perold [10] describes a single-period model that incorporates the deadweight costs generated by the opaqueness of banks. Stoughton and Zechner [14] examine the question of capital allocation among businesses of a bank, assuming information asymmetry between the different business managers and senior management in a single-period model. The investment in risky technologies is financed with debt. Equity is used to satisfy a Value-at-Risk constraint. They derive a type of RAROC performance metric for each business, the hurdle rate being driven by the cost of debt, the distribution associated with the information asymmetry, and the outside employment opportunities for managers. Implication and measurement issues are not addressed.

None of these papers attempt to address the myriad of practical issues facing banks. A start is made in [15]. Each business within the bank is assumed to have its own balance sheet. This necessitates making some assumptions about the capital structure of the business. The risky assets of the business are financed with debt. The bank lends the business an amount of debt and charges the business an interest rate reflecting the expected maturity of the project and the credit risk of the bank. The required economic capital depends on the nature of the risks associated with project over its expected duration, giving rise to a *term structure of required economic capital*.^f This is taken as exogenous. The economic capital is assumed to be financed *via* equity. The business invests the economic capital in a short-term default-free asset. The costs associated with economic capital arise from taxation. The resulting hurdle rate depends on the systematic risk associated with the risky

investment and a second factor due to the requirement to hold costly economic capital. The weights of the two factors can be estimated, implying that it is possible to establish a required benchmark for performance measurement. If the term structure of economic capital allocated to the business is done on a marginal basis, the required rate of return on the business depends on the variance of the return on the business, the variance of the return on the remaining part of the bank, and the covariance between the two returns. If done on a stand-alone basis, it depends on the variance of the return on the business.

End Notes

^aMany different methods, from bottom-up to top-down, have been advocated for the allocation of capital. See [6, 16, Chapter 6] for brief descriptions.

^bMatten [5, Chapter 7], describes many different types of performance metrics used by practitioners, as well as references to the history of RAROC. In a recent survey of chief risk officers of financial firms—see [11]—nearly 90% stated that they used risk-adjusted performance measures, with RAROC being the dominant measure.

^cSee [1, Chapter 14; 12, Chapter 7] for a detailed description of how to apply RAROC.

^dSee [3] for a discussion on the pitfalls of RAROC.

^eThe issue of opaqueness in financial institutions is discussed in [13, 7].

^fA simple example would be a foreign currency swap. Principal is exchanged in this type of swap and this increases the credit risk. Initially, this is low risk, though as the swap approaches maturity, more economic capital is required.

References

- [1] Crouhy, M., Galai, D. & Mark, R. (2001). *Risk Management*, McGraw-Hill, New York.
- [2] Crouhy, M., Turnbull, S.M. & Wakeman, L. (1999). Measuring risk adjusted performance, *Journal of Risk* 2(1), 5–35.
- [3] Demine, J. (1998). Pitfalls in the application of RAROC in loan management, *Arbitrageur* 1(1), 21–27.
- [4] Froot, K.A. & Stein, J.C. (1998). Risk management, capital budgeting and capital structure policy for financial institutions: an integrated approach, *Journal of Finance* 48, 1629–1658.
- [5] Matten, C. (2000). *Managing Bank Capital*. John Wiley & Sons, New York.
- [6] McNeil, A. Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management*, Princeton University Press, NJ.

-
- [7] Merton, R.C. (1993). Operation and regulation in financial intermediation: a functional perspective, in *Operational and Regulation of Financial Markets*, P. Englund, ed, The Economic Council, Stockholm.
 - [8] Merton, R.C. (1997). A model of financial guarantees for credit sensitive, opaque financial intermediaries, *European Finance Review* **1**, 1–13.
 - [9] Merton, R.C. & Perold, A.F. (1993). Management of risk capital in financial firms, in *Financial Services*, S.L. Hayes, ed, Harvard Business School Press, Boston.
 - [10] Perold, A.F. (2001). *Capital Allocation in Financial Firms*, Working paper, Harvard Business School.
 - [11] Prmia (2007). *Risk Adjusted Performance Measurement*, surveys@prmia.org.
 - [12] Ranson, B.J. (2003). *Credit Risk Management*, Thomson, Sheshunoff, TX.
 - [13] Ross, S.A. (1989). Institution markets, financial marketing, and financial innovation, *Journal of Finance* **44**, 541–556.
 - [14] Stoughton, N.M. & Zechner, J. (2006). *Optimal Capital Allocation Using RAROC and EVA*, Working paper, University of Calgary.
 - [15] Turnbull, S.M. (2000). Capital allocation and risk performance measurement in a financial institution, *Financial Markets, Institutions & Instruments*, NYU **9**(5), 325–357.
 - [16] Turnbull, S.M. (2002). Bank and business performance measurement, *Journal of Economic Notes* **31**(2), 215–236.
 - [17] Wilson, T. (1992). RAROC remodelling, *Risk* **5**(8), 112–119.

Related Articles

Economic Capital Allocation; Economic Capital; Market Risk; Value-at-Risk.

STUART M. TURNBULL

Stress Testing

It is widely recognized that the most popular statistical financial risk measure, Value at Risk (VaR), is not, except under highly restrictive conditions, a sufficient statistic for describing risk in portfolios of financial assets. It is slightly ironic that nonstatistical techniques increasingly have come to be relied upon as tools for overcoming the deficiencies in VaR, signaling an (implicit) rejection of a risk epistemology based on objective principles and deductive reasoning in favor of a subjective alternative reliant on inductive reasoning. The prescriptions of regulatory authorities have added significant impetus to this trend. Partly as a result of this exogenous motivating force, these tools have experienced an incomplete ontogeny. Although various mathematical techniques of stress testing and scenario analysis have been developed, an axiomatic treatment justifying the incorporation of these tools into economic decision making is still missing.

Stress testing and scenario analysis aim to elicit information about how risk assessments are affected by extreme changes in either valuation-driving factors or exogenously determined parameters. Scenario analysis generally involves a more holistic approach, in which many factors are assumed to change simultaneously. Stress testing usually refers to techniques in which a single factor or parameter is changed, possibly repeatedly. Unfortunately, the two sobriquets are frequently used interchangeably.

These techniques are widely used in banks and other financial institutions. Some financial regulation requires regulated entities to use these techniques in risk measurement as a condition for allowing the use of (more favorable) risk-based capital calculations.

Historical Origins

Following several, high profile, financial derivatives-related trading losses incurred by financial and nonfinancial firms in the early 1990s, various government-sponsored studies seeking ways to ameliorate future adverse events called for improved risk assessment. An increased focus on stress testing and scenario analysis was among the specific recommendations therein.

The earliest of these recommendations came from the Group of Thirty's Report of the Global Derivatives Study Group (1993). Recommendation six of the report states, "Dealers should regularly perform simulations to determine how their portfolios would perform under stress conditions." The most detailed prescriptions came from the Basel Committee for Banking Supervision in its (1996) "Amendment to the Capital Accord to Incorporate Market Risks". For example, the document states, "Banks should subject their portfolios to a series of simulated stress scenarios . . . for example, the 1987 equity crash, the ERM crises of 1992 and 1993, or the fall in the bond markets in the first quarter of 1994." The perceived importance of these techniques continues to wax large in the regulatory agenda. For example, in 2006 the Committee of European Banking Supervisors issued a consultative paper on stress testing in which it opined, "Stress testing is a valuable risk management technique whose potential applications are quite varied . . ." A comprehensive system of stress testing and scenario analysis is generally a required concomitant to risk-based methods for regulatory capital adequacy determination.

Stress Testing

Stress testing, as the name suggests, seeks to gauge risk by examination of extreme changes in the underlying random factors that affect valuation. By factor, we mean a source of random variation in the price of a financial instrument. For example, two random factors affect the value of an exchange traded equity call (or put) option, variation in the market price of the underlying equity and variation in the volatility of the underlying equity.

In its basic form, stress testing involves changing one factor at a time only, holding all others constant. This technique can be applied to a portfolio of instruments or to a single position.

Factor-push

One common flavor of this sort of stress test is called *factor-push*, because the stress test outcome is the result of "pushing" each factor in the direction resulting in portfolio loss. This approach can be implemented mechanically as follows.

First, a "push" size must be selected. This push size will be applied sequentially to each factor, i ,

where f_i refers to the current value of the i th factor. The push size is subjective. One way it may be given a heuristic justification is if, for each factor, it is taken to be a constant, c , multiplied by the factor's standard deviation, s_i .

Second, every position under consideration is valued twice, using the following two alternative values for the factor:

$$f_i^+ = f_i(1 + cs_i), \quad f_i^- = f_i(1 - cs_i) \quad (1)$$

The value for the factor that results in the larger total portfolio reduction in value is retained, whereas the other is discarded. This process is then repeated for the remaining factors.

Third, every position is valued once more, this time simultaneously using all the factor changes that were retained during the previous step.

As noted earlier, there is no rigorously defensible rationale for the selection of any particular size for the push factor. This method can introduce other undesirable effects into the final portfolio stress loss estimate as well.

Because the ultimate step uses factor changes that are each chosen independently, the set of factor changes may make no economic sense, if viewed together as a single economic scenario. Also, to the extent that the portfolio contains instruments that have nonzero second-order cross-factor sensitivities, these will be ignored in developing the final set of factor changes.

If the portfolio contains some instruments whose values are either nonlinear or discontinuous functions of some of the factors, then for any given scheme for determining the push size, it can be shown that the resulting portfolio stress loss may not be greater than alternative schemes that have a strictly smaller push size. Such would be the case, for example, if the portfolio consisted of only a long position in an option straddle.

Another variation of the factor-push method is used by some derivatives exchanges and over the counter (OTC) market dealers for the determination of the customers' portfolio margin requirements. In this method, each factor is allowed to take on a range of values. For example, the current factor level may be scaled up or down in increments of 1 standard deviation until the increment reaches c standard deviations. A zero factor change is included (for a reason which will shortly be obvious), giving in this case $2c$ alternative values for each factor.

In this method, all factors are allowed to change simultaneously. Therefore, if there are n factors and $2c + 1$ possible values for each, then $(2c + 1)^n$ portfolio revaluations are performed, and the one yielding the greatest portfolio loss is taken to be the stress loss.

The factor-push method is also related to a common method for managing trading portfolio risk when the portfolio includes instruments whose values are a nonlinear function of the factors driving pricing. A *sensitivity ladder* is a set of possible changes specified for each factor, where the instruments are revalued (and typically first-order sensitivities recalculated as well) for each specified change in a factor.

Maximum Loss

A stress testing technique that attempts to account for the deficiencies in the factor-push type of stress test is called *maximum loss* and was first developed by Studer [8].

Maximum loss is defined as the n -tuple of factor changes for which the corresponding portfolio loss is a maximum, subject to the factor changes being a member of a feasible set. Studer defines the feasible set of factor changes with respect to a (continuous) joint distribution of factor returns. Specifically, he requires that the likelihood of any feasible n -tuple of factor returns must exceed some given probability.

This approach is difficult to implement. For an arbitrary portfolio of financial instruments, this approach requires evaluating every feasible point, both boundary and interior points, in the space of factor returns. Random or quasirandom methods might be employed to create a practical implementation of this method. Any sampling technique would potentially run into problems with portfolios that exhibit nonlinear sensitivities to the factors. With detailed knowledge of the instruments in the portfolio, it may be possible to construct a computationally efficient search algorithm that samples only from areas of the factor return distribution that are likely to result in the maximum loss.

Stress Testing of Correlation Matrices

This technique focuses on the impact of changing the joint dynamics of factors affecting portfolio valuation. It is common belief that volatile market environments are characterized by different correlation

dynamics. Several authors have argued inconclusively for a mixture of normals as description of factor returns over time.

To implement stress testing for changes in a correlation matrix, a “shocked” matrix is used to calculate the change in portfolio value. Typically, the shocked matrix is intended to represent a “high correlation” environment. The shocks are typically specified based on subjective rules similar to those used in the techniques discussed earlier. However, because the covariance matrix must be positive definite, shocks that can be validly applied to the correlations are constrained.

The usual technique to deal with this issue is to apply the subjectively determined shocks, then manipulate the resulting covariance matrix to ensure positive definiteness with the goal of departing minimally from the intended shocks. Much has been written on the problem of determining the closest positive definite matrix to a given matrix by elimination of the negative eigenvalues from the stressed correlation matrix, for example, see [7].

Scenario Analysis

The intent of scenario analysis is to analyze the performance of a portfolio when subjected to an ensemble of factor changes intended to represent a specific market event, one typically characterized as extreme but plausible and relevant.

The “extremeness” of a scenario may be defined in two ways. First, it may be measured with respect to the magnitude of factor changes considered. In this case, there is no guarantee that the scenario will correspond to a significant loss in portfolio value. Alternatively, it may be measured with respect to the magnitude of the change in portfolio value resulting from factor changes, that is, an ensemble of factor changes that corresponds to a portfolio loss greater than a threshold level.

The concept of plausibility is usually judged in reference to the set of factor changes, not the portfolio positions. No standards of plausibility exist. However, Breuer and Krenn [5] propose to define the plausibility of a given scenario as the cumulative probability, under the assumption that factor returns are distributed multivariate normal, of all the scenarios with probability less than or equal to the given scenario. To justify this criterion, Basel Committee on Banking Supervision [3] refers to its intuitive appeal.

Without an arbitrary definition of plausibility, it may not be possible to establish necessary conditions for a scenario to be deemed plausible. For example, no-arbitrage and nonnegative price conditions may be violated under severe market stresses. As illustrations, note that during the equity market crash of 1987, the arbitrage relationship between the US equity index cash market and index futures prices did not hold at all times. Similarly, negative yields for discount bills of the US Treasury were observed at times during the market crisis of 2007–2008.

Relevance is defined with respect to the positions in the portfolio and the current market environment. For example, an “oil embargo” scenario might be thought not relevant to a portfolio containing only interest rate derivatives. There is no objective method for evaluating (or ranking) relevance. Therefore, relevance is incorporated into scenario selection by judgment.

The selection of factor changes used in the scenario selection is, therefore, in the final analysis, a subjective exercise. Scenarios are typically modeled in one of the two ways. First, an actual historical period is used as the basis for the scenario. Historical scenarios may be generally agreed “events” or may be selected mechanically, for example, by “data-snooping” histories of factor changes, selecting as candidate scenarios those dates where more than a minimum number of factor changes exceeds a given threshold. Also, financial regulators have identified certain historical events, which they have deemed to be suitable stress scenarios, such as the equity market crash of 1987 (see, e.g. [3]).

Given the chosen period, factor returns from that period are collected and applied to the current levels of market factors, yielding “shocked” factor values. For market factors that did not trade during the chosen period, a proxy must be used to create an estimated shock. Changes in the dynamical properties of factors, arising, for example, from market microstructure changes between the historical period chosen and the present, are typically ignored, and the factor shocks applied as if those changes would have no effect in a replay of the scenario. Finally, the shocked factor values are used to revalue the portfolio and a stress loss (or profit) is calculated.

Second, an ensemble of factor changes is created by employing (explicitly or implicitly) a structural or reduced form model of market dynamics to generate a hypothetical scenario. The result may yield

quantitative shocks for only a subset of the market factors relevant to a portfolio. The remaining factor shocks are set based on judgment. A variation on this approach employs scenarios in which a subset of factors are shocked in highly stylized ways (e.g., see [6]), such as by a parallel move in yields or a change in all equity market prices of a single percentage amount. Historical scenarios claim general assent to plausibility because they have actually happened. However, the claim may sometimes be questioned, for, as noted above, the passage of time may render an historical scenario increasingly artificial, as reliance on proxies for shocks and the extent of changes in market dynamics increase.

It has been documented in the experimental psychology literature that individuals systematically underestimate the probability of rare events, and they overestimate probabilities for events that have occurred more recently. The subjective nature of scenario selection may therefore *ipso facto* lead to systematic biases in scenario selection.

It is claimed by some that scenario analysis will provide incremental knowledge about the risk of a portfolio beyond which it might be available using other risk measures, such as VaR. The arguments in support of this claim are not rigorous. In the absence of a probability model, what may be extreme and plausible can only be evaluated subjectively. Thus, if one individual claims some quantum of knowledge derived from evaluating a scenario, then a general assent as to the claim may be difficult to achieve (beyond the trivial sense in which all would agree with the statement, “if these shocks are applied in this way to this set of mathematical relations describing portfolio value, then this change in portfolio value is observed”).

Even if the problem of intersubjectivity is put aside, it is not obvious how to measure the increment to knowledge from scenario analysis. A scenario maps an n -tuple of factor changes to a single value for the portfolio return. If the space of factor changes is continuous and the set of possible (extreme) values for portfolio return is continuous, it can be argued that the measure of the information from a scenario is zero, unless a not-small (subjective) probability is assigned to the specific set of factor changes considered. This conclusion might be modified if it can be shown that (i) small perturbations in the n -tuple of factor changes from the base case scenario result in portfolio returns that are close the return

in the base case and (ii) small perturbations in the portfolio return map to only n -tuples of factor changes that are “close” to the base case set of factor changes. For a large and complex portfolio containing instruments, whose price is nonlinear or discontinuous in terms of factor changes, neither of these conditions can be guaranteed.

It has been shown in [1, 2] and related research that a scenario can be interpreted as a worst-case loss on a countable, finite event space, and as such satisfies the axioms of a coherent risk measure. It has also been shown that a coherent risk measure can be obtained from scenarios. These results place scenario analysis squarely within the framework of risk theory. In limited contexts, it is possible to construct decision-theoretic justifications for rules based on stress scenarios. The key element in so doing is the ability to completely specify the event space of portfolio outcomes. This can be accomplished self-referentially, as when a decision maker, regulator, and so on, defines the set of relevant outcomes, where any other outcome not included in (or derivable from) this set of predefined outcomes is ignored or declared irrelevant. A widely cited example is the Standard Portfolio Analysis of Risk (SPAN) margining system used by many derivatives exchanges in which a given set of scenarios is deemed to be relevant for determining capital adequacy for a market participant.

Historical scenarios are typically constructed to mimic factor moves that span multiple trading days in order to encompass an entire period of market disturbance. The identification of either a beginning or ending date may be ambiguous, if the goal is to capture peak-to-trough factor changes, as these may not be synchronous across all markets. A subjective choice of dates to be employed must be made in such cases.

In general, no allowances are made for time-related valuation effects in a scenario, such as cash flows, time decay (“theta”), expiration, and reinvestment of net cash inflows. However, long horizon (e.g., several months to over one year) scenario analyses, such as the one performed by some financial firms for nontrading books, do include these effects.

The static portfolio assumption is most often justified by the argument that illiquidity would make it very difficult and costly either to trade out of any positions or to put on hedge positions during the market event. To some extent, this can be observed in markets undergoing stress.

This static approach to scenario-based stress testing does not allow for an active assessment of liquidity risk, which is greatest during a crisis. Only a dynamic scenario can capture the losses resulting from portfolio adjustment subsequent to a change in market liquidity, as manifest in widening of bid-ask spreads, and an increase in the price impact of trades.

The potential usefulness of a given historical scenario will fall over time, making scenario analysis a dynamic process; over time it is necessary to replace or revise scenarios. The deterioration has several origins. Market microstructure changes will affect factor price dynamics. Examples are the switch from eighths to decimals on the US equity exchanges in 2001 and the replacement of floor execution with electronic trading systems. Also, the set of traded instruments changes as a result of bankruptcy, merger, financial innovation, and so on. Changes in market liquidity (perhaps arising from the entrance or exit of new market participants) will affect price dynamics as well. Even large changes in price levels can adversely affect the usability of an historical scenario, potentially causing the historical factor changes to defy common sense by being large or too small.

Since 2001 the International Monetary Fund, in cooperation with member central banks, has conducted model-based scenario analysis for certain member countries. In each country, *macro stress testing* employs a central-bank model of the entire banking system and focuses on factor changes and their impacts on balance sheets, which might result in a system-wide event, for example, contagious default

and systemic contraction of lending to preserve capital. These factors might include the change in corporate defaults, changes in counterparty default, or changes in foreign exchange (FX) rates, trade balance or FX reserves, or shocks to asset prices, such as real estate. For a survey of these methods, see [4].

References

- [1] Artzner, P. (1999). Application of coherent risk measures to capital requirements in insurance, *North American Actuarial Journal* **3**, 11–25.
- [2] Artzner, P., Delbaen, F., Eber, J.-M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**, 203–228.
- [3] Basel Committee on Banking Supervision (1996). *Amendment to the Capital Accord to Incorporate Market Risks*, Mimeo, 46.
- [4] Blaschke, W., Jones, M., Majnoni, G., Martinez, P. & Soledad, M. (2001). *Stress Testing of Financial Systems*, International Monetary Fund, Mimeo.
- [5] Breuer, T. & Krenn, G. (1999). Reliving past crises: identifying stress scenarios from historical data, *Third International Stockholm Seminar on Risk Behaviour and Risk Management*, Complementary Papers, Stockholm, pp. 111–119.
- [6] Derivatives Policy Group (1995). *Framework for Voluntary Oversight*, Mimeo.
- [7] Rebonato, R. & Jaeckel, P. (1999). The most general methodology to create a valid correlation matrix for risk management and option pricing purposes, *Journal of Risk* **2**(Winter), 1–16.
- [8] Studer, G. (1997). *Maximum Loss Measurement for Market Risk*, Doctoral Thesis, Swiss Federal Institute of Technology.

BARRY SCHACHTER

Model Validation

One important way to mitigate model risk (*see Models*; for a more detailed discussion of model risk and how to mitigate model risk, see Crouhy *et al.* [1], chapter 14) is to invest in research to improve models and to develop better statistical tools, either internally at the bank or externally at a university (or at an analytically oriented consulting organization).

An even more vital way of reducing model risk is to establish a process for the independent vetting of how models are both selected and constructed. This should be complemented by independent oversight of the profit and loss (P&L) calculation. In this article, we address the *ex ante* validation of pricing models developed for the front office of a trading institution. *Ex post* validation of risk models, or back testing, is addressed in **Backtesting**.

The role of vetting is to offer assurance to the firm's management that any model proposed by, say, a trading desk for the valuation of a given security, is reasonable. In other words, the model offers a reasonable representation of how the market itself values the instrument, and that the model has been implemented correctly. Vetting should consist of the following phases:

1. Documentation

The vetting team should ask for a full documentation of the model, including both the assumptions underlying the model and its mathematical expression. This should be independent of any particular implementation, such as a spreadsheet or a C++ computer code, and should include the following:

- The term sheet or, equivalently, a complete description of the transaction.
- A mathematical statement of the model, which should include the following:
 - An explicit statement of all the components of the model: stochastic variables and their processes, parameters, equations, and so on.
 - The payoff function and/or any pricing algorithm for complex structured deals.
 - The calibration procedure for the model parameters.
 - The hedge ratios/sensitivities.

- Implementation features, that is, inputs, outputs, and numerical methods employed.
- A working version of the implementation.

2. Soundness of model

An independent model vetter needs to verify that the mathematical model is a reasonable representation of the instrument that is being valued. For example, the manager might reasonably accept the use of a particular model (e.g., the Black model) for valuing a short-term option on a long maturity bond, but reject (without even looking at the computer code) the use of the same model to value a two-year option on a three-year bond. At this stage, the risk manager should concentrate on the finance aspects, and not become overly focused on the mathematics.

3. Independent access to financial rates

The model vetter should check that the bank's middle office has independent access to an independent market risk management financial rates database (to facilitate independent parameter estimation).

4. Benchmark modeling

The model vetter should develop a benchmark model based on the assumptions that are being made and on the specifications of the deal. Here, the model vetter may use a different implementation from the implementation that is being proposed. A proposed analytical model can be tested against a numerical approximation technique or against a simulation approach. (For example, if the model to be vetted is based on a "tree" implementation, one may instead rely on the partial differential equation approach and use the finite element technique to derive the numerical results.) The results of the benchmark test can be then compared with those of the proposed model.

5. Health check and stress test the model

Also, make sure that the model possesses the basic properties that all derivatives models should possess, such as put-call parity and other nonarbitrage conditions. Finally, the vetter should stress-test the model. The model can be stress-tested by looking at some limit scenario, in order to identify the range of parameter values for which the model provides accurate pricing. This is especially important for implementations that rely on numerical techniques.

6. Build a formal treatment of model risk into the overall risk management procedures, and periodically reevaluate models

Also, reestimate parameters using best-practice statistical procedures. Experience shows that simple, but robust models tend to work better than more ambitious, but fragile models. It is essential to monitor and control model performance over time.

Reference

- [1] Crouhy, M., Galai, D. & Mark, R. (2006). *The Essentials of Risk Management*, McGraw-Hill.

Related Articles

Backtesting; Models.

MICHEL CROUHY, DAN GALAI & ROBERT
MARK

Backtesting

The term *backtesting* is used in several ways in finance. Most commonly, backtesting denotes either (i) an assessment of the hypothetical historical performance of a suggested trading strategy or (ii) the evaluation of financial risk models using historical data on risk forecasts and profit and loss realizations. The following discussion is about the evaluation of risk models.

The procedure is to consider the daily *ex ante* risk measure forecasts from a model and test it against the daily *ex post* realized portfolio loss. The risk measure forecast could take the form of a Value-at-Risk (VaR), an expected shortfall, or a distributional forecast. The goal is to be able to backtest any of these risk measures of interest. The backtesting procedures can be seen as a final diagnostic check on the risk model carried out by the risk management team that constructed the risk model, or they can be used by external model evaluators such as bank supervisors.

Evidence on actual bank VaRs and their backtesting performance can be found in [7, 22, 28, 29, 31]. The regulatory considerations involved in backtesting are detailed by the Basel Committee on Banking Supervision [2–4]. Lopez [24] analyzes the regulatory approach to backtesting and Campbell [9] provides a survey of backtesting that includes a discussion of regulatory considerations. Backtesting of credit risk models is investigated in [25].

First, we discuss procedures for backtesting the VaR metric (*see Value-at-Risk; Market Risk*). Second, we consider increasing backtesting power by using explanatory variables to backtest the VaR. Third, we consider the increasing power by backtesting risk measures other than VaR, and we discuss various other issues in backtesting.

Backtesting VaR

By now, VaR is by far the most popular portfolio risk measure used by risk management practitioners. The VaR revolution in risk management was triggered by JP Morgans RiskMetrics approach launched in 1994. Supervisory authorities immediately recognized the need for methods to backtest VaR and the first research on backtesting was published soon after in

[21, 23]. Christoffersen [10] extended Kupiec’s test of unconditional VaR coverage to tests of conditional VaR coverage. These concepts are defined later.

Defining the Hit Sequence

First, define VaR_{t+1}^p to be a number constructed on day t such that the portfolio losses on day $t + 1$ will only be larger than the VaR_{t+1}^p forecast with probability p . If we observe a time series of past *ex ante* VaR forecasts and past *ex post* losses, PL , we can define the “hit sequence” of VaR violations as

$$I_{t+1} = 1 \quad \text{if} \quad PL_{t+1} > VaR_{t+1}^p \quad (1)$$

$$= 0 \quad \text{if} \quad PL_{t+1} < VaR_{t+1}^p \quad (2)$$

The hit sequence returns a 1 on day $t + 1$ if the loss on that day is larger than the VaR number predicted in advance for that day. If the VaR is not exceeded (or violated), then the hit sequence returns a 0. When backtesting the risk model, we construct a sequence $\{I_{t+1}\}_{t=1}^T$ across T days, indicating when the past violations occurred.

The Null Hypothesis

If we use the perfect VaR model, then given all the information available to us at the time the VaR forecast is made, we should not be able to predict whether the VaR will be violated. We should be expecting a 1 in the hit sequence with probability p and we should be expecting a 0 with probability $1 - p$ and the occurrences of the hits should be random over time.

We say that a risk model has correct *unconditional* VaR coverage if $\Pr(I_{t+1} = 1) = p$ and we say that a risk model has correct *conditional* VaR coverage if $\Pr_t(I_{t+1} = 1) = p$. Roughly speaking, correct *unconditional* VaR coverage just means that the risk model delivers VaR hits with probability p on average across the days. Correct *conditional* VaR coverage means that the risk model gives a VaR hit with probability p on every day given all the information available on the day before. Note that correct conditional coverage implies correct unconditional coverage but not *vice versa*.

We can think of VaR backtesting as testing the hypothesis:

$$H_0 : I_{t+1} \sim \text{i.i.d. Bernoulli}(p) \quad (3)$$

If p is one half, then the i.i.d. Bernoulli distribution describes the distribution of getting a head when tossing a fair coin. When backtesting risk models, p will not be one half but instead on the order of 0.01 or 0.05 depending on the coverage rate of the VaR. The hit sequence from a correctly specified risk model should look like a sequence of random tosses of a coin that comes up heads 1 or 5% of the time depending on the VaR coverage rate.

Note that, in general, the expected value of a binary sequence is simply the probability of getting a 1:

$$\begin{aligned} E_t [I_{t+1}] &= \Pr_t (I_{t+1} = 1) 1 + \Pr_t (I_{t+1} = 0) 0 \\ &= \Pr_t (I_{t+1} = 1) \equiv \pi_{t+1|t} \end{aligned} \quad (4)$$

We denote this conditional hit probability by $\pi_{t+1|t}$ and its unconditional counterpart is defined $E [I_{t+1}] \equiv \pi$.

We can therefore construct the following null hypothesis of correct conditional coverage for the hit sequence from a VaR model

$$H_0 : E_t [I_{t+1}] = \pi_{t+1|t} = p \quad (5)$$

namely that the conditionally expected value of the hit sequence at time $t + 1$, given all the information available at time t is the promised VaR coverage rate p .

Unconditional Coverage Testing

First, we want to test if the unconditional probability of a violation in the risk model, π , is significantly different from the promised probability, p . We call this the *unconditional coverage hypothesis*. We can write $H_0 : E [I_t] \equiv \pi = p$.

The expected value of the hit sequence can be estimated by the sample average, $\hat{\pi} = \frac{1}{T} \sum_{t=1}^T I_t = T_1/T$, where T_1 is the number of 1s in the sample. Note that if the null hypothesis is true, we have that $E [\hat{\pi}] = E \left[\frac{1}{T} \sum_{t=1}^T I_t \right] = p$. Assuming that the observations on the hit sequence are independent over time, the variance of the $\hat{\pi}$ estimate is $1/T \text{Var} (I_t)$ where $\text{Var} (I_t)$ can be estimated as the sample variance of the hit sequence. The hypothesis that $E [I_t] = p$ can therefore be tested in a simple means test

$$MT = \sqrt{T} \frac{\hat{\pi} - p}{\sqrt{\text{Var} (I_t)}} \sim N(0, 1) \quad (6)$$

We can also implement the unconditional coverage test as a likelihood ratio test. For this, we write the likelihood of an i.i.d. Bernoulli(π) hit sequence:

$$L(\pi) = \prod_{t=1}^T (1 - \pi)^{1-I_{t+1}} \pi^{I_{t+1}} = (1 - \pi)^{T_0} \pi^{T_1} \quad (7)$$

where T_0 and T_1 are the number of 0s and 1s in the sample. We can easily estimate π from $\hat{\pi} = T_1/T$, that is, the observed fraction of violations in the sequence. Plugging the ML estimates back into the likelihood function gives the optimized likelihood as

$$L(\hat{\pi}) = \left(1 - \frac{T_1}{T}\right)^{T_0} \left(\frac{T_1}{T}\right)^{T_1} \quad (8)$$

Under the unconditional coverage null hypothesis that $\pi = p$, where p is the known VaR coverage rate, we have the likelihood

$$L(p) = \prod_{t=1}^T (1 - p)^{1-I_{t+1}} p^{I_{t+1}} = (1 - p)^{T_0} p^{T_1} \quad (9)$$

And we can check the unconditional coverage hypothesis using a likelihood ratio test

$$LR_{uc} = -2 \ln \left[\frac{L(p)}{L(\hat{\pi})} \right] \sim \chi_1^2 \quad (10)$$

Asymptotically, that is, as the number of observations, T , goes to infinity, the test will be distributed as a χ^2 with one degree of freedom.

The choice of significance level comes down to an assessment of the costs of making two types of mistakes: we could reject a correct model (type I error) or we could fail to reject (that is accept) an incorrect model (type II error). Increasing the significance level implies larger type I errors but smaller type II errors and *vice versa*. In academic work, a significant level of 1, 5, or 10% is typically used. In risk management, type II errors may be very costly so that a significance level of 10% may be appropriate.

Often, we do not have a large number of observations available for backtesting, and we certainly will typically not have a large number of violations, T_1 , which are the informative observations. It is, therefore, often better to rely on Monte-Carlo-simulated P values rather than those from the χ^2 distribution. Christoffersen and Pelletier [14] discuss how

to implement the [17] Monte Carlo P -values in a backtesting setting.

Independence Testing

There is strong evidence of time-varying volatility in daily asset returns as surveyed in [1]. If the risk model ignores such dynamics, then the VaR will react slowly to the changing market conditions and VaR violations will appear clustered in time. Pritsker [30] illustrates this problem using VaRs computed from historical simulation. If the VaR violations are clustered, then the risk manager can essentially predict that if today has a VaR hit, then there is a probability larger than p of getting a hit tomorrow, which violates that the VaR is based on an adequate model. This is clearly not satisfactory. In such a situation, the risk manager should increase the VaR in order to lower the conditional probability of a violation to the promised p .

The most common way to test for dynamics in time series analysis is to rely on the autocorrelation function and the associated portmanteau or Ljung–Box-type tests. We can implement this approach for backtesting as well. Let γ_k be the autocorrelation at lag k for the hit sequence. Plotting γ_k against k for $k = 1, \dots, m$ will give a visual impression of the correlation of between a hit in one of the last m trading days and a hit today. The Ljung–Box test provides a formal check of the null hypothesis that all of the first m autocorrelations are zero against the alternative that any of them is not. The test is easily constructed as

$$LB(m) = T(T+2) \sum_{k=1}^m \frac{\gamma_k^2}{T-k} \sim \chi_m^2 \quad (11)$$

where χ_m^2 denotes the chi-squared distribution with m degrees of freedom. Berkowitz *et al.* [6] find that setting $m = 5$ gives good testing power in a realistic daily VaR backtesting setting.

Independence testing can also be done using the likelihood approach. Assume that the hit sequence is dependent over time and that it can be described as a first-order Markov sequence with transition probability matrix:

$$\Pi_1 = \begin{bmatrix} 1 - \pi_{01} & \pi_{01} \\ 1 - \pi_{11} & \pi_{11} \end{bmatrix} \quad (12)$$

These transition probabilities simply mean that conditional on today being a nonviolation (i.e., $I_t = 0$) then the probability of tomorrow being a violation (i.e., $I_{t+1} = 1$) is π_{01} . The probability of tomorrow being a violation given today is also a violation is $\pi_{11} = \Pr(I_t = 1 \text{ and } I_{t+1} = 1)$. The probability of a nonviolation following a nonviolation is $1 - \pi_{01}$ and the probability of a nonviolation following a violation is $1 - \pi_{11}$.

If we observe a sample of T observations, then we can write the likelihood function of the first-order Markov process as

$$L(\Pi_1) = (1 - \pi_{01})^{T_{00}} \pi_{01}^{T_{01}} (1 - \pi_{11})^{T_{10}} \pi_{11}^{T_{11}} \quad (13)$$

where T_{ij} , $i, j = 0, 1$, is the number of observations with a j following an i . Taking first derivatives with respect to π_{01} and π_{11} and setting these derivatives to zero, one can solve for the maximum likelihood estimates:

$$\hat{\pi}_{01} = \frac{T_{01}}{T_{00} + T_{01}}, \quad \hat{\pi}_{11} = \frac{T_{11}}{T_{10} + T_{11}} \quad (14)$$

Then using the fact that the probabilities have to sum to one, we have $\hat{\pi}_{00} = 1 - \hat{\pi}_{01}$, $\hat{\pi}_{10} = 1 - \hat{\pi}_{11}$.

Allowing for dependence in the hit sequence corresponds to allowing π_{01} to be different from π_{11} . Positive dependence is a matter of concern, which amounts to the probability of a violation following a violation (π_{11}) being larger than the probability of a violation following a nonviolation (π_{01}). If, on the other hand, the hits are independent over time, then the probability of a violation tomorrow does not depend on today being a violation or not and we write $\pi_{01} = \pi_{11} = \pi$. We can test the independence hypothesis that $\pi_{01} = \pi_{11}$ using a likelihood ratio test

$$LR_{\text{ind}} = -2 \ln \left[\frac{L(\hat{\pi})}{L(\hat{\Pi}_1)} \right] \sim \chi_1^2 \quad (15)$$

where $L(\hat{\pi})$ is the likelihood under the alternative hypothesis from the LR_{uc} test.

Other methods for independence testing based on the duration of time between hits can be found in [14].

Backtesting with Information Variables

The preceding tests are quick and easy to implement. However, as they only use information on past VaR

hits, they might not have much power to detect misspecified risk models. To increase the testing power, we consider using the information in past market variables, such as interest rate spreads or volatility measures. The basic idea is to test the model using information that may explain when violations occur. The advantage of increasing the information set is not only to increase power but also to help us understand the areas in which the risk model is misspecified. This understanding is key in improving the risk models further.

If we define the q -dimensional vector of information variables available to the backtester at time t as X_t , then the null hypothesis of a correct risk model can be written as

$$H_0 : \Pr(I_{t+1} = 1|X_t) = p \Leftrightarrow E[I_{t+1} - p|X_t] = 0 \quad (16)$$

The hypothesis says that the conditional probability of getting a VaR violation on day $t + 1$ should be independent of any variable observed at time t and it should simply be equal to the promised VaR coverage rate, p . This hypothesis is equivalent to the conditional expectation of the hit sequence less p being equal to 0.

Engle and Manganelli [18] develop a conditional autoregressive VaR by regression quantiles approach. For backtesting, they suggest the following dynamic quantile test:

$$DQ = \frac{(I - p)' X (X' X)^{-1} X' (I - p)}{Tp(1 - p)} \sim \chi_q^2 \quad (17)$$

where X is the $T \times q$ matrix of information variables, and $(I - p)$ is the $T \times 1$ vector of hits where each element has been subtracted by the desired covered rate p .

Berkowitz *et al.* [6] find that implementing the DQ test using simply the lagged VaR from a GARCH model as well as the lagged hit gives good power in a realistic daily VaR experiment. Smith [31] also finds good power when applying the DQ test. A regression-based approach to backtesting is also used in [11, 12]. Christoffersen *et al.* [13] develop tests for comparing different VaR models.

Smith [31] further suggests a probit approach where the potentially time-varying probability of a hit is modeled using the normal cumulative density

function and where Lagrange multiplier tests are used.

When backtesting with information variables, the question of which variables to include in the vector X_t immediately arises. It is difficult to give a general answer as it depends on the particular portfolio at hand. It is likely that variables that are thought to be correlated with the future volatility in the portfolio are good candidates. In equity portfolios option, implied volatility measures such as the VIX would be an obvious candidate. In FX portfolios option, implied volatility measures could be constructed as well. In bond portfolios, variables such as term spreads and credit spreads are likely candidates as are variables capturing the level, slope, and curvature of the term structure.

Other Issues in Backtesting

When correctly specified, the one-day VaR measure tells the user that there is a probability p of losing more than the VaR over the next day. Importantly, it does not, for example, say how much one can expect to lose on the days where the VaR is violated. Thus, the VaR only contains partial information about the distribution of losses. This limitation of the VaR is reflected in the definition of the hit variable in equation (1), which in turn limits the possibilities for backtesting in the VaR setting. We can only backtest on the hit occurrences and not, for example, on the magnitude of the losses when the hits occur because the VaR does not promise hits of a certain magnitude.

Backtesting Expected Shortfall

The limitations of the VaR as a risk measure suggest alternative risk measures, most prominently expected shortfall (ES), also referred to as *conditional VaR* ($CVAR$) (see **Value-at-Risk; Market Risk**), which denotes the expected loss on the days where the VaR is violated. We can define it formally as

$$ES_{t+1}^p = E_t[PL_{t+1}|PL_{t+1} > VaR_{t+1}^p] \quad (18)$$

Note that ES provides information on the expected magnitude of the loss whenever the VaR is violated. We now consider ways to backtest the ES risk measure.

Consider again a vector of variables, X_t , which are known to the risk manager and which may help explain potential portfolio losses beyond what is explained by the risk model. The ES risk measure promises that whenever we violate the VaR, the expected value of the violation will be equal to ES_{t+1}^p . We can therefore test the ES measure by checking if the vector X_t has any ability to explain the deviation of the observed shortfall or loss, PL_{t+1} , from the expected shortfall on the days where the VaR was violated. Mathematically, we can write

$$PL_{t+1} - ES_{t+1}^p = b_0 + b_1' X_t + e_{t+1} \quad \text{for} \\ t + 1 \text{ where } PL_{t+1} > VaR_{t+1}^p \quad (19)$$

where $t + 1$ now refers only to days where the VaR was violated. The observations where the VaR was not violated are simply removed from the sample. The error term e_{t+1} is assumed to be independent of the regressor, X_t .

To test the null hypothesis that the risk model from which the ES forecasts were made uses all information optimally ($b_1 = 0$) and that it is not biased ($b_0 = 0$), we can jointly test that $b_0 = b_1 = 0$.

Note that now the magnitude of the violation shows up on the left-hand side of the regression. However, note that we can still only use information in the tail to backtest. The ES measure does not reveal any particular properties about the remainder of the distribution, and therefore we only use the observations where the losses were larger than the VaR.

Further, discussion of ES backtesting can be found in [26].

Backtesting the Entire Distribution

Rather than focusing on particular risk measures from the loss distribution such as the VaR or the expected shortfall, we could instead decide to backtest the entire loss distribution from the risk model. This would have the benefit of potentially increasing further the power to reject bad risk models.

Assuming that the risk model produces a cumulative distribution of portfolio losses, call it $F_t(*)$. Then at the end of every day, after having observed the actual portfolio loss (or profit), we can calculate the risk model's probability of observing a

loss below the actual. We denote this probability by $p_{t+1} \equiv F_t(PL_{t+1})$.

If we use the correct risk model to forecast the loss distribution, then we should not be able to forecast the risk model's probability of falling below the actual return. In other words, the time series of observed probabilities p_{t+1} should be distributed independently over time as a Uniform(0,1) variable. We therefore want to consider tests of the null hypothesis:

$$H_0 : p_{t+1} \sim \text{i.i.d. Uniform}(0, 1) \quad (20)$$

The Uniform(0,1) distribution function is flat on the interval 0–1 and 0 everywhere else. As the p_{t+1} variable is a probability, it is a must that it should lie in the 0–1 interval. Constructing a histogram and checking if it looks reasonably flat provide a useful visual diagnostic. If systematic deviations from a flat line appear in the histogram, then we would conclude that the distribution from the risk model is misspecified. A detailed discussion of this approach is found in [16].

Unfortunately, testing the i.i.d. uniform distribution hypothesis is cumbersome due to the restricted support of the uniform distribution. However, we can transform the i.i.d. Uniform p_{t+1} to an i.i.d. standard normal variable, z_{t+1} , using the inverse cumulative distribution function, $z_{t+1} = \Phi^{-1}(p_{t+1})$. We are then left with a test of a variable conforming to the standard normal distribution, which can easily be implemented. The i.i.d. property of z_{t+1} can be assessed *via* the autocorrelation functions and by using the $LB(m)$ test in equation (4) on z_{t+1} and $|z_{t+1}|$, for example. The normal distribution property can be tested using the method of moments approach in [8]. Regression-based tests using information variables can also be constructed. Further analysis of distribution backtesting can be found in [5, 15].

Dirty Profits and Losses

The backtesting procedures compare the *ex ante* forecast from a risk model to the *ex post* profit and losses (P/L). It is therefore obviously essential, but unfortunately not always the case, that the *ex post* recorded P/L arise directly from the portfolio used to make the *ex ante* model predictions.

The risk model will typically produce a risk forecast for the loss distributions of a particular portfolio of assets. The *ex post* profits and losses should only contain cash flows directly related to the portfolio of assets assumed in the risk models. However, the total *P/L* of a trading desk may contain trading commission revenues as well as costs that are not directly related to holding the portfolio of assets assumed in the risk model. This extraneous cash flows should be stripped from the *P/L* before backtesting.

The daily *P/L* may also include cash flows from intraday transactions, that is, the *P/L* from selling an asset that was bought the same day. Such cash flows are not directly related to the end-of-day portfolio entered into the risk model and should ideally be stripped away as well. A detailed discussion of these issues can be found in [27].

Multiple-day Horizons and Changing Portfolio Weights

The backtesting procedures described earlier can be relatively easily adopted to the multiday risk-horizon setting. If, for example, a 10-day VaR forecast is observed daily along with the *ex post* 10-day *P/L*, then the hit sequence can be constructed daily as in equation (1). The 10-day VaR horizons implies that 9-day autocorrelation is to be expected in the hit sequence and that this should be allowed for when constructing its variance as in equation (6). Similarly, in the *DQ* test as in equation (17), one should use information variables available 10 days before the observed *P/L*. Smith [31] discusses backtesting with multiday VaRs.

Allowing for Risk Model Parameter Estimation Error

In the discussion so far, we have abstracted from the fact that the risk models in use most likely contain parameters that are estimated with errors. This parameter estimation error will in turn render the hit sequence observed with error. As a consequence when backtesting the risk model, we may reject the “true” risk model just because its parameters were estimated with error. As discussed in [18], this is mainly an issue when the risk model is evaluated in-sample, that is, when the model is evaluated on the same data it was estimated on. In typical external

backtesting applications, the backtesting procedures are used in a more realistic out-of-sample manner where the model is estimated on data until day t and then used to forecast risk for day $t + 1$. In this setting, parameter estimation error is less likely to be critical. Parameter estimation issues in relation to VaR modeling has been analyzed in detail in [19, 20].

Acknowledgments

The author is grateful for the financial support from FQRSC, IFM², and SSHRC and from the Center for Research in Econometric Analysis of Time Series, CREATES, funded by the Danish National Research Foundation.

References

- [1] Andersen, T., Bollerslev, T., Christoffersen, P. & Diebold, F. (2005). Practical volatility and correlation modeling for financial market risk management, in *Risks of Financial Institutions*, M. Carey & R. Stulz, eds, University of Chicago Press for NBER, pp. 513–548.
- [2] Basel Committee on Banking Supervision (1996). *Amendment to the Capital Accord to Incorporate Market Risks*, Bank for International Settlements.
- [3] Basel Committee on Banking Supervision (1996). *Supervisory Framework for the Use of ‘Backtesting’ in Conjunction With the Internal Models Approach to Market Risk Capital Requirements*, Bank for International Settlements.
- [4] Basel Committee on Banking Supervision (2004). *International Convergence of Capital Measurement and Capital Standards: a Revised Framework*.
- [5] Berkowitz, J. (2001). Testing density forecasts, applications to risk management, *Journal of Business and Economic Statistics* **19**, 465–474.
- [6] Berkowitz, J., Christoffersen, P. & Pelletier, D. (2007). *Evaluating Value-at-Risk Models with Desk-level Data*, Management Science, forthcoming.
- [7] Berkowitz, J. & O’Brien, J. (2002). How accurate are the Value-at-Risk models at commercial banks? *Journal of Finance* **57**, 1093–1112.
- [8] Bontemps, C. & Meddahi, N. (2005). Testing normality: a GMM approach, *Journal of Econometrics* **124**, 149–186.
- [9] Campbell, S. (2007). A review of backtesting and backtesting procedures, *Journal of Risk* **9**, 1–17.
- [10] Christoffersen, P. (1998). Evaluating interval forecasts, *International Economic Review* **39**, 841–862.
- [11] Christoffersen, P. (2003). *Elements of Financial Risk Management*, Academic Press.

-
- [12] Christoffersen, P. & Diebold, F. (2000). How relevant is volatility forecasting for financial risk management? *Review of Economics and Statistics* **82**, 1–11.
 - [13] Christoffersen, P., Hahn, J. & Inoue, A. (2001). Testing and comparing Value-at-Risk measures, *Journal of Empirical Finance* **8**, 325–342.
 - [14] Christoffersen, P.F. & Pelletier, D. (2004). Backtesting Value-at-Risk: a duration-based approach, *Journal of Financial Econometrics* **2**, 84–108.
 - [15] Crnkovic, C. & Drachman, J. (1996). Quality control, *Risk* **9**, 138–143.
 - [16] Diebold, F.X., Gunther, T. & Tay, A. (1998). Evaluating density forecasts, with applications to financial risk management, *International Economic Review* **39**, 863–883.
 - [17] Dufour, J.-M. (2006). Monte Carlo tests with nuisance parameters: a general approach to finite-sample inference and nonstandard asymptotics in econometrics, *Journal of Econometrics* **133**, 443–477.
 - [18] Engle, R.F. & Manganelli, S. (2004). CAViaR: conditional autoregressive Value-at-Risk by regression quantiles, *Journal of Business and Economic Statistics* **22**, 367–381.
 - [19] Escanciano, J. & Olmo, J. (2007). *Estimation Risk Effects on Backtesting For Parametric Value-at-Risk Models*, City University London (Manuscript).
 - [20] Giacomini, R. & Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts, *Journal of Business and Economic Statistics* **23**, 416–431.
 - [21] Hendricks, D. (1996). Evaluation of Value-at-Risk models using historical data, *Economic Policy Review*, Federal Reserve Bank of New York, 39–69.
 - [22] Jorion, P. (2002). How informative are Value-at-Risk disclosures? *Accounting Review* **77**, 911–931.
 - [23] Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models, *Journal of Derivatives* **3**, 73–84.
 - [24] Lopez, J. (1999). Regulatory evaluation of Value-at-Risk models, *Journal of Risk* **1**, 37–64.
 - [25] Lopez, J. & Saidenberg, M. (2000). Evaluating credit risk models, *Journal of Banking and Finance* **24**, 151–165.
 - [26] McNeil, A. & Frey, R. (2000). Estimation of tail-related risk measures for heteroskedastic financial time series: an extreme value approach, *Journal of Empirical Finance* **7**, 271–300.
 - [27] O'Brien, J. & Berkowitz, J. (2005). Bank trading revenues, VaR and market risk, in *Risks of Financial Institutions*, M. Carey & R. Stulz, eds, University of Chicago Press for NBER.
 - [28] Perignon, C., Deng, Z. & Wang, Z. (2006). *Do Banks Overstate their Value-at-Risk?*, HEC, Paris (Manuscript).
 - [29] Perignon, C. & Smith, D. (2006). *The Level and Quality of Value-at-Risk Disclosure by Commercial Banks*, Simon Fraser University (Manuscript).
 - [30] Pritsker, M. (2001). *The Hidden Dangers of Historical Simulation*, Finance and Economics Discussion Series 2001–27, Board of Governors of the Federal Reserve System.
 - [31] Smith, D. (2007). *Conditional Backtesting of Value-at-Risk Models*, Simon Fraser University (Manuscript).

Related Articles

Expected Shortfall; Market Risk; Model Validation; Value-at-Risk.

PETER CHRISTOFFERSEN

Copulas: Estimation

Copulas are a tool for modeling and capturing the dependence of two or more random variables (rv's). In the work of Sklar [22], the term *copula* was used for the first time; it is derived from the Latin word *copulare*, which means to connect or to join. Similarly, Hoeffding had already studied distributions under “arbitrary changes of scale” in the 1940s; see [7].

The main purpose of a copula is to disentangle the dependence structure of a random vector from its marginals. A d -dimensional copula is defined as a function $C : [0, 1]^d \rightarrow [0, 1]$, which is a cumulative distribution function (cdf) with uniform^a marginals. On one hand, this leads to the following properties:

1. $C(u_1, \dots, u_d)$ is increasing in each component $u_i, i \in \{1, \dots, d\}$.
2. $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$ for all $1 \leq i \leq d$.
3. For $a_i \leq b_i, 1 \leq i \leq d$, C satisfies the *rectangle inequality*

$$\sum_{i_1=1}^2 \dots \sum_{i_d=1}^2 (-1)^{i_1+\dots+i_d} C(u_{1,i_1}, \dots, u_{d,i_d}) \geq 0,$$

where $u_{j,1} = a_j$ and $u_{j,2} = b_j$.

On the other hand, every function satisfying (1)–(3) is a copula. Furthermore, $C(1, u_1, \dots, u_{d-1})$ is again a copula and so are all k -dimensional marginals with $2 \leq k < d$.

The construction of *multivariate* copulas is difficult. There is a rich literature on uni- and bivariate distributions but many of these families do not have obvious multivariate generalizations.^b Similarly, it is not at all straightforward to generalize a two-dimensional copula to higher dimensions. For example, consider the construction of three-dimensional copulas. A possible attempt is to try $C_1(C_2(u_1, u_2), u_3)$ where C_1, C_2 are bivariate copulas. However, already for $C_1 = C_2 = \max\{u_1 + u_2 - 1, 0\}$ (the countermonotonicity copula, introduced in the section *Important Copulas*) this procedure fails. See [18, Section 3.4] for further details. Also Chapter 4 in [11] gives an overview of construction of multivariate copulas with different concepts. In particular, it discusses the construction of a d -dimensional copula given the set of $d(d-1)/2$ bivariate margins.

The class of Archimedean copulas (see also the section Archimedean Copulas) is an important class for which the construction of multivariate copulas can be performed quite generally. A common example of a three-dimensional Archimedean copula is given by the following exchangeable Archimedean copula:

$$C(u_1, u_2, u_3) = \phi^{-1}(\phi(u_1) + \phi(u_2) + \phi(u_3)) \quad (1)$$

with appropriate generator ϕ . However, for appropriate ϕ_1, ϕ_2 ,

$$\phi_1^{-1}(\phi_2 \circ \phi_1^{-1}(\phi_1(u_1) + \phi_1(u_2)) + \phi_2(u_3)) \quad (2)$$

also gives a three-dimensional copula (see [14], Section 5.4.3). It is of course possible that ϕ_1 and ϕ_2 are generators of different types of Archimedean copulas.

The key to the separation of marginals and dependence structure is the *quantile transformation*. Let U be a standard uniform rv and $F^{-1}(y) := \inf\{x : F(x) \geq y\}$ be the generalized inverse of F . Then

$$P(F^{-1}(U) \leq x) = F(x) \quad (3)$$

This result is frequently used for simulation: the generation of uniform rv's is readily implemented in typical software packages and if we are able to compute F^{-1} , we can sample from F using equation (1).

On the contrary, the *probability transformation* is used to compute copulas implied from distributions, see the following section Copulas derived from Distributions. Consider X having a continuous distribution function F_2 , then $F(X)$ is standard uniform.^c

Sklar's Theorem

It is not surprising that every distribution function inherently embodies a copula function. On the other hand, any copula entangled with some marginal distributions in the right way leads to a proper multivariate distribution function. This is the important contribution of Sklar's theorem [22]. Ran F denotes the range of F .

Theorem Consider a d -dimensional cdf F with marginals $F_1, \dots, F_d$. There exists a copula C such that

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)) \quad (4)$$

for all x_i in $[-\infty, \infty]$, $i = 1, \dots, d$. If F_i is continuous for all $i = 1, \dots, d$ then C is unique; otherwise, C is uniquely determined only on $\text{Ran } F_1 \times \dots \times \text{Ran } F_d$. On the other hand, consider a copula C and univariate cdf's $F_1, \dots, F_d$. Then F as defined in equation (4) is a multivariate cdf with marginals $F_1, \dots, F_d$.

It is important to note that for discrete distributions, copulas are not as natural as they are for continuous distributions; compare [8].

In the following, we therefore concentrate on continuous F_i , $i = 1, \dots, d$. It is interesting to examine the consequences of representation (2) for the copula itself. Using that $F \circ F^{-1}(y) = y$ for any continuous CDF F , we obtain

$$C(\mathbf{u}) = F(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d)) \quad (5)$$

While relation (4) is usually the starting point for simulations that are based on a given copula and given marginals, relation (5) rather proves to be the theoretical tool to obtain the copula from any multivariate distribution function. This equation also allows to extract a copula directly from a multivariate distribution function.

Invariance Under Transformations

An important property of a copula is that it is invariant under strictly increasing transformations: for strictly increasing functions $T_i : \mathbb{R} \rightarrow \mathbb{R}$, $i = 1, \dots, d$ the rv's $X_1, \dots, X_d$ and $T_1(X_1), \dots, T_d(X_d)$ have the same copula.

Bounds of Copulas

Hoeffding and Fréchet independently derived that a copula always lies in between certain bounds; compare Figure 1. This is because of the existence of some extreme cases of dependency, co- and countermonotonicity. The so-called *Fréchet–Hoeffding bounds* are given by

$$\max \left\{ \sum_{i=1}^d u_i + 1 - d, 0 \right\} \leq C(\mathbf{u}) \leq \min \{u_1, \dots, u_d\} \quad (6)$$

which holds for any copula C . Although a comonotonic copula exists in any dimension d , there is no countermonotonicity copula in the case of dimensions greater than two.^d

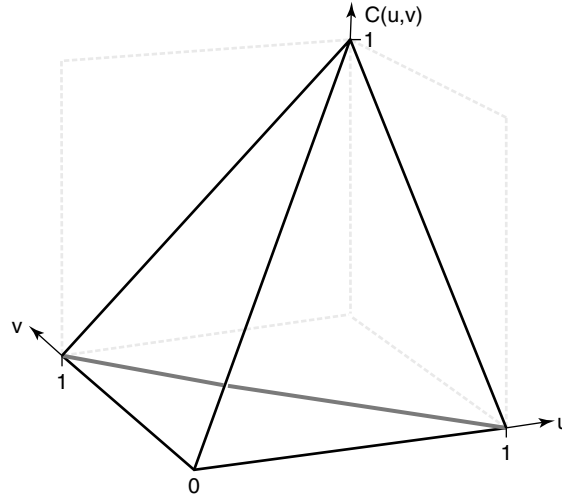


Figure 1 According to the Fréchet–Hoeffding bounds, every copula has to lie inside of the pyramid shown in the graph. The surface given by the bottom and back side of the pyramid (the lower bound) is the countermonotonicity copula $C(u, v) = \max \{u + v - 1, 0\}$, while the front side (the upper bound) is the comonotonicity copula, $C(u, v) = \min(u, v)$

Important Copulas

First, the *independence copula* is given by

$$\prod_{i=1}^d u_i \quad (7)$$

Random variables are independent if and only if their copula is the independence copula.

The *comonotonicity copula* or the *Fréchet–Hoeffding upper bound* is given by

$$\min\{u_1, \dots, u_d\} \quad (8)$$

Random variables $X_1, \dots, X_d$ are called *comonotonic*, if their copula is as in equation (8). This is equivalent to $(X_1, \dots, X_d)$ having the same distribution as $(T_1(Z), \dots, T_d(Z))$ with some rv Z and strictly increasing functions $T_1, \dots, T_d$. Therefore, comonotonicity refers to perfect dependence in the sense where all rv's are, in an increasing and deterministic way, depending on Z .

The other case of perfect dependence is given by countermonotonicity. The *countermonotonicity copula* reads

$$\max\{u_1 + u_2 - 1, 0\} \quad (9)$$

Two rv's with this copula are called *countermonotonic*. This is equivalent to (X_1, X_2) having the same distribution as $(T_1(Z), T_2(Z))$ for some rv Z and T_1 being increasing and T_2 being decreasing or vice versa. However, the *Fréchet–Hoeffding lower bound* as given in equation (6) is *not* a copula for $d > 2$; see [14], Example 5.21.

Copulas Derived from Distributions

The probability transformation^e allows to obtain the copula inherent in multivariate distributions: for a multivariate cdf F with continuous marginals F_i , the inherent copula is given by

$$C(\mathbf{u}) = F(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d)) \quad (10)$$

For example, for a multivariate normal distribution, the implied copula is called *Gaussian copula*. For a d -dimensional rv X , the correlation matrix^f Γ is obtained from the covariance matrix by scaling each component to variance 1. Therefore, Γ is given by the entries $\text{Corr}(X_i, X_j)$, $1 \leq i, j \leq d$ (see **Correlation**

Risk). For such a *correlation matrix* Γ the Gaussian copula is given by

$$\Phi_{\Gamma}(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)) \quad (11)$$

In a similar fashion, one obtains the *t-copula* or the *Student copula*

$$t_{v,\Gamma}(t_v^{-1}(u_1) \dots, t_v^{-1}(u_d)) \quad (12)$$

where Γ is the correlation matrix, t_v is the cdf of the one dimensional t_v distribution, and $t_{v,\Gamma}$ is the cdf of the multivariate $t_{v,\Gamma}$ distribution. The mixing nature of the t -distribution leads to a dramatically different behavior in the tails, which is an important property in applications. See the section *Tail Dependence*.

Archimedean Copulas

An important class of analytically tractable copulas are the *Archimedean copulas*. For the bivariate case, consider a continuous and strictly decreasing function $\phi : [0, 1] \rightarrow [0, \infty]$ with $\phi(1) = 0$, called the *generator*. Then $C(u_1, u_2)$ given by

$$\begin{cases} \phi^{-1}(\phi(u_1) + \phi(u_2)) & \text{if } \phi(u_1) + \phi(u_2) \leq \phi(0) \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

is a copula if and only if ϕ is convex; see [18], Theorem 4.1.4. If $\phi(0) = \infty$ the generator is said to be *strict* and $C(u_1, u_2) = \phi^{-1}(\phi(u_1) + \phi(u_2))$.

For the multivariate case, there are different possibilities of generalization. A relatively special case is when the copula is of the form

$$\phi^{-1}(\phi(u_1) + \dots + \phi(u_d)) \quad (14)$$

These are the so-called exchangeable Archimedean copulas and [15] give a complete characterization of such ϕ leading to a copula of the form (7). One may also consider asymmetric specifications of multivariate Archimedean copulas; see [14] Section 5.4.2 and 5.4.3. We present some examples of Archimedean copulas in bivariate case: from the generator $(-\ln u)^\theta$ one obtains the bivariate *Gumbel copula* or *Gumbel–Hougaard copula* as follows:

$$\exp\left(-\left[(-\ln u_1)^\theta + (-\ln u_2)^\theta\right]^{\frac{1}{\theta}}\right) \quad (15)$$

where $\theta \in [1, \infty)$. For $\theta = 1$ it coincides with the independence copula, and for $\theta \rightarrow \infty$, it converges

4 Copulas: Estimation

to the comonotonicity copula. The Gumbel copula has tail dependence in the upper right corner.

The *Clayton copula* is given by^g

$$(\max \{u_1^{-\theta} + u_2^{-\theta} - 1, 0\})^{-\frac{1}{\theta}} \quad (16)$$

where $\theta \in [-1, \infty) \setminus \{0\}$. For $\theta \rightarrow 0$ it converges to the independence copula, and for $\theta \rightarrow \infty$ to the comonotonicity copula. For $\theta = -1$ we obtain the Fréchet–Hoeffding lower bound. The generator $\theta^{-1}(u^{-\theta} - 1)$ of the Clayton copula is strict only if $\theta > 0$. In this case

$$C_{\theta}^{Cl}(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}} \quad (17)$$

The generator $\ln(e^{-\theta} - 1) - \ln(e^{-\theta u} - 1)$ leads to the *Frank copula* given by

$$-\frac{1}{\theta} \ln \left(1 + \frac{(e^{-\theta u_1} - 1) \cdot (e^{-\theta u_2} - 1)}{e^{-\theta} - 1} \right) \quad (18)$$

for $\theta \in \mathbb{R} \setminus \{0\}$.

The *generalized Clayton copula* is obtained from the generator $\theta^{-\delta}(u^{-\theta} - 1)^{\delta}$:

$$\left([(u_1^{-\theta} - 1)^{\delta} + (u_2^{-\theta} - 1)^{\delta}]^{\frac{1}{\delta}} + 1 \right)^{-\frac{1}{\theta}} \quad (19)$$

with $\theta > 0$ and $\delta \geq 1$. Note that for $\delta = 1$ the standard Clayton copula is attained; compare [18], Example 4.19.

Further examples of Archimedean copulas may be found in [18], and in particular one may consider Table 4.1 therein as well as Sections 4.5 and 4.6 (Table 1).

Table 1 List of some copulas. For the Gumbel, Clayton, Frank and the Marshall–Olkin copula, only the bivariate versions are stated. References to the multivariate versions are given in the text. Further copulas may be found in [18], Table 4.1 (p. 94) and Sections 4.5 as well as 4.6

Name	Copula	Parameter range
Independence or product copula	$\Pi(\mathbf{u}) = \prod_{i=1}^d u_i$	
Comonotonicity copula or Fréchet–Hoeffding upper bound	$M(\mathbf{u}) = \min \{u_1, \dots, u_d\}$	
Countermonotonicity copula or Fréchet–Hoeffding lower bound	$W(u_1, u_2) = \max \{u_1 + u_2 - 1, 0\}$	
Gaussian copula ^(a)	$C_{\Gamma}^{Ga}(\mathbf{u}) = \Phi_{\Gamma}(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d))$	
<i>t</i> - or Student copula ^(a)	$C_{v, \Gamma}^t(\mathbf{u}) = t_{v, \Gamma}(t_v^{-1}(u_1), \dots, t_v^{-1}(u_d))$	
Gumbel copula or Gumbel–Hougaard copula	$C_{\theta}^{Gu}(u_1, u_2) = \exp \left(- [(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}]^{\frac{1}{\theta}} \right)$	$\theta \in [1, \infty)$
Clayton copula ^(b)	$C_{\theta}^{Cl}(u_1, u_2) = (\max \{u_1^{-\theta} + u_2^{-\theta} - 1, 0\})^{-\frac{1}{\theta}}$	$\theta \in [-1, \infty)$
Generalized Clayton ^(b) copula	$C_{\theta, \delta}^{Cl}(u_1, u_2) = \left([(u_1^{-\theta} - 1)^{\delta} + (u_2^{-\theta} - 1)^{\delta}]^{\frac{1}{\delta}} + 1 \right)^{-\frac{1}{\theta}}$	$\theta \geq 0, \delta \geq 1$
Frank copula ^(b)	$C_{\theta}^{Fr}(u_1, u_2) = -\frac{1}{\theta} \ln \left(1 + \frac{(e^{-\theta u_1} - 1) \cdot (e^{-\theta u_2} - 1)}{e^{-\theta} - 1} \right)$	$\theta \in \mathbb{R}$
Marshall–Olkin copula or generalized Cuadras–Augé copula	$C_{\alpha_1, \alpha_2}(u_1, u_2) = \min \{u_2 \cdot u_1^{1-\alpha_1}, u_1 \cdot u_2^{1-\alpha_2}\}$	$\alpha_1, \alpha_2 \in [0, 1]$

^(a)Here Γ is a correlation matrix, that is, a covariance matrix where each variance is scaled to 1

^(b)For the (generalized) Clayton and the Frank copula, the case $\theta = 0$ is given as the limit for $\theta \rightarrow 0$, which leads to the independence copula in both cases

The Marshall–Olkin Copula

The *Marshall–Olkin copula* is a copula with singular component. For intuition, consider two components that are subject to certain shocks that lead to failure of either one or both the components. The shocks occur at times that are assumed to be independent and exponentially distributed. Denote the realized shock times by Z_1, Z_2 and Z_{12} . Then we obtain for the probability that the two components live longer than x_1 and x_2 , respectively,

$$P(Z_1 > x_1)P(Z_2 > x_2)P(Z_{12} > \max\{x_1, x_2\}) \quad (20)$$

This extends to the multivariate case in a straightforward way; compare [4] and [18]. The related copula equals

$$\min\left\{u_2 \cdot u_1^{1-\alpha_1}, u_1 \cdot u_2^{1-\alpha_2}\right\} \quad (21)$$

with $\alpha_i \in [0, 1]$. A similar family is given by the *Cuadras-Augé* copulas

$$\min\{u_1, u_2\} \cdot (\max\{u_1, u_2\})^\alpha \quad (22)$$

$\alpha \in [0, 1]$; see [3].

Measures of Dependence

Measures of dependence summarize the dependence structures of rv's. There are three important concepts: linear correlation, rank correlation, and tail dependence. A further concept of dependence is *association*; see [6, 17].

Linear Correlation

Linear correlation is a well-studied concept. It is a dependence measure, which is useful only for *elliptical distributions* (see **Multivariate Distributions**). This is because elliptical distributions are fully described by mean vector, covariance matrix, and a characteristic generator function. As mean and variances are determined by the marginal distributions, the copulas of elliptical distributions depend only on the covariance matrix and the generator function. Linear correlation, therefore, has a distinguished role in this class, which it does not have in other multivariate models.

Rank Correlation

Rank correlations describe the dependence structure of the ranks, that is, the dependence structure of the considered rv's when transformed to uniform marginals using the probability transformation. Most importantly, this implies a direct representation in terms of the underlying copula; compare equation (5). We consider *Kendall's tau* and *Spearman's rho*, which also play an important role in nonparametric statistics.

For rv's $\mathbf{X} = X_1, \dots, X_d$ with marginals $F_i, i = 1, \dots, d$, *Spearman's rho* is defined by

$$\rho_S(\mathbf{X}) := \text{Corr}(F_1(X_1), \dots, F_d(X_d)) \quad (23)$$

Corr is the correlation matrix whose entries are given by $\text{Corr}(F_i(X_i), F_j(X_j))$.

Consider an independent copy $\tilde{\mathbf{X}}$ of $\mathbf{X}$. Then *Kendall's tau* is defined by

$$\rho_\tau(\mathbf{X}) := \text{Cov}\left[\text{sign}(\mathbf{X} - \tilde{\mathbf{X}})\right] \quad (24)$$

For $d = 2$,

$$\begin{aligned} \rho_\tau(X_1, X_2) &= P\left((X_1 - \tilde{X}_1) \cdot (X_2 - \tilde{X}_2) > 0\right) \\ &\quad - P\left((X_1 - \tilde{X}_1) \cdot (X_2 - \tilde{X}_2) < 0\right) \end{aligned} \quad (25)$$

which explains this measure of dependency.

Both measures have values in $[-1, 1]$; they are 0 for independent variables (while there might also be nonindependent rv's with zero rank correlation) and they equal 1 (−1) for the comonotonic (countermonotonic) case. Moreover, they can directly be derived from the copula of $\mathbf{X}$; see [14], Proposition 5.29. For example,

$$\rho_S(X_1, X_2) = 12 \int_0^1 \int_0^1 (C(u_1, u_2) - u_1 u_2) du_1 du_2 \quad (26)$$

In the case of a bivariate Gaussian copula one obtains^h $\rho_S(X_1, X_2) = \frac{6}{\pi} \arcsin \frac{\rho}{2}$ and a similar expression for ρ_τ .

For other examples and certain bounds that interrelate those two measures, we refer the reader to [18],

Sections 5.1.1–5.1.3. For multivariate extensions see, for example, [21] and [23].

Tail Dependence

We distinguish between *upper* and *lower* tail dependence. Consider two rv's X_1 and X_2 with marginals F_1, F_2 and copula C . *Upper tail dependence* means intuitively that with large values of X_1 also large values of X_2 are expected. More precisely, the *coefficient of upper tail dependence* is defined by

$$\lambda_u := \lim_{q \nearrow 1} P(X_2 > F_2^{-1}(q) | X_1 > F_1^{-1}(q)) \quad (27)$$

provided the limit exists and $\lambda_u \in [0, 1]$. The *coefficient of lower tail dependence* is

$$\lambda_l := \lim_{q \searrow 0} P(X_2 \leq F_2^{-1}(q) | X_1 \leq F_1^{-1}(q)) \quad (28)$$

If $\lambda_u > 0$, X_1 and X_2 are called *upper tail dependent*, while for $\lambda_u = 0$ they are *asymptotically independent* in the upper tail; this applies analogously for λ_l . For continuous cdf's, Bayes' rule gives

$$\lambda_l = \lim_{q \searrow 0} \frac{C(q, q)}{q} \quad (29)$$

and

$$\lambda_u = 2 + \lim_{q \searrow 0} \frac{C(1-q, 1-q) - 1}{q} \quad (30)$$

A Gaussian copula has no tail dependence if the correlation is not equal to 1 or -1 . For the bivariate t -distribution

$$\lambda_l = \lambda_u = 2t_{v+1} \left(-\sqrt{\frac{(v+1)(1-\rho)}{1+\rho}} \right) \quad (31)$$

provided $\rho > -1$. Note that even for zero correlation this copula shows tail dependence.

Tail dependence is a key quantity for joint quantile exceedances; see Example 5.34 in [14]: a multivariate Gaussian distribution will give a much smaller probability to the event that all returns from a portfolio are below the 1% quantiles of their respective distributions than a multivariate t -distribution. This is because of the difference in the tail dependence.

Association

A relatively stronger concept than correlation is the so-called association introduced in [6]. If $\text{Cov}(X, Y) \geq 0$, then one would consider X and Y as somehow associated. If, moreover, $\text{Cov}(f(X), g(Y)) \geq 0$ for all pairs of nondecreasing functions f, g , they would be considered more strongly associated. If $\text{Cov}(f(X, Y), g(X, Y)) \geq 0$ for all pairs of functions f, g which are nondecreasing in each argument, an even stronger dependence holds. Random variables $X_1, \dots, X_d =: \mathbf{X}$ are called *associated* if $\text{Cov}(f(\mathbf{X}), g(\mathbf{X})) \geq 0$ for all f, g that are nondecreasing and the covariance exists. Examples of associated rv's include independent rv's, positively correlated normal variables, and also the generalized exponential distribution turning up in the Marshall–Olkin copula.

Sampling from Copulas

Consider given marginals $F_1, \dots, F_d$ and a given copula C . The first step is to simulate $(U_1, \dots, U_d)$ with uniform marginals and copula C . By equation (3), the vector $(F_1^{-1}(U_1), \dots, F_d^{-1}(U_d))$ has copula C and the desired marginals.

If the copula is inherited from a multivariate distribution, the task reduces to simulating this multivariate distribution, for example, Gaussian or t -distribution.

If the copula is Archimedean, this task is more demanding and we refer the reader to [14], Algorithm 5.48 for details.

Conclusion

On one hand, copulas are a very general tool to describe dependence structures and have been successfully applied in many cases. However, the immense generality is also a drawback in many applications and also the static characteristic of this measure of dependence has been criticized; see the referenced literature. Obviously, the application of copulas has been a great success to a number of fields, and they are a frequently used concept especially in finance. They serve as an excellent tool for calibrating dependence structures or stress testing portfolios or other products in finance and insurance as they allow to interpolate between extreme cases of dependence.

Literature

The literature on copulas is growing fast. The vital article on copulas **Copulas in Insurance** by P. Embrechts gives an excellent overview of the literature and applications. An introduction to copulas that extends this note in many ways may be found in [20]. For a detailed exposition of copulas with different applications in view, we refer the reader to [4, 14, 19]. Reference [2] gives additional examples and [13] analyzes extreme financial risks. Estimation of copulas is discussed in [14], Section 5.5, [1] and [10]. For an in-depth study of copulas consider [11, 18]. Interesting remarks of the history and the development of copulas may be found in [7]. For more details on Marshall–Olkin copulas, in particular the multivariate ones, see [4, 18]. In the modeling of Lévy processes (see **Lévy Processes**) one considers dependency of jumps where the measure is no longer a probability measure. This leads to the development of so-called Lévy copulas; compare [12] and **Lévy Copulas**. The pitfalls mentioned with linear correlation are discussed in detail in [5] or [14 Chapter 5.2.1]. For a discussion on the general difficulties in the application of copulas, we refer the reader to [9, 16].

End Notes

^aAlthough standard, it is not necessary to consider uniform marginals (see **Copulas in Insurance**).

^bOne example is the exponential distributions whose multivariate extension leads to the Marshall–Olkin copula, introduced in the following paragraph.

^cSee, for example [14], Proposition 5.2.

^dSee [14], Example 5.21, for a counterexample.

^eFor a rv X with continuous cdf F , the rv $F(X)$ is standard uniform, see Section 1.

^fSee **Correlation Risk**.

^gFor generating the Clayton copula, it would be sufficient to use $(u^{-\theta} - 1)$ instead of $\theta^{-1}(u^{-\theta} - 1)$ as generator. However, for $\theta < 0$, this function is increasing and the above result would not be applicable.

^hCompare [14], Theorem 5.36. ρ_S and ρ_τ for elliptic distributions are also covered.

Acknowledgments

The author thanks F. Durante and R. Frey for helpful comments.

References

- [1] Charpentier, A., Fermanian, J.-D. & Scaillet, O. (2007). The estimation of copulas: from theory to practice, in *Copulas: From Theory to Applications in Finance*, J. Rank, ed, Risk Books, pp. 35–60.
- [2] Cherubini, U., Luciano, E. & Vecchiato, W. (2004). *Copula Methods in Finance*, Wiley, Chichester.
- [3] Cuadras, C.M. & Augeé, J. (1981). A continuous general multivariate distribution and its properties, *Communications in Statistics: Theory and Methods* **10**, 339–353.
- [4] Embrechts, P., Lindskog, F. & McNeil, A.J. (2003). Modeling dependence with copulas and applications to risk management, in *Handbook of Heavy Tailed Distributions in Finance*, S.T. Rachev, ed, Elsevier, pp. 331–385.
- [5] Embrechts, P., McNeil, A.J. & Straumann, D. (2001). Correlation and dependency in risk management: properties and pitfalls, in *Risk Management: Value at Risk and Beyond*, M. Dempster & H.K. Moffatt, eds, Cambridge University Press, pp. 176–223.
- [6] Esary, J.D., Proschan, F. & Walkup, D.W. (1967). Association of random variables, with applications, *Annals of Mathematical Statistics* **28**, 1466–1474.
- [7] Fisher, N.I. (1995). Copulas, in *Encyclopedia of Statistical Sciences*, S. Kotz, C. Read, N. Balakrishnan, & B. Vidakovic, eds, Wiley, pp. 159–163.
- [8] Genest, C. & Nešlehová, J. (2007). A primer on copulas for count data, *Astin Bulletin* **37**, 475–515.
- [9] Genest, C. & Rémillard, B. (2006). Diskussion of “copulas: tales and facts”, by Thomas Mikosch, *Extremes* **9**, 27–36.
- [10] Genest, C., Rémillard, B. & Beaudoin, D. (2007). Goodness-of-fit tests for copulas: A review and a power study, *Insurance: Mathematics and Economics* in press, <http://dx.doi.org/10.1016/j.insmatheco.2007.10.005>.
- [11] Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.
- [12] Kallsen, J. & Tankov, P. (2006). Characterization of dependence of multidimensional Lévy processes using Lévy copulas, *Journal of Multivariate Analysis* **97**, 1551–1572.
- [13] Malevergne, Y. & Sornette, D. (2006). *Extreme Financial Risks*, Springer Verlag, Berlin.
- [14] McNeil, A.J., Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton University Press.
- [15] McNeil, A.J. & Nešlehová, J. (2008). Multivariate Archimedean copulas, d -monotone functions and l_1 -norm symmetric distributions, *Annals of Statistics*, forthcoming.
- [16] Mikosch, T. (2006). Copulas: tales and facts, *Extremes* **9**, 3–20.
- [17] Müller, A. & Stoyan, D. (2002). *Comparison Methods for Stochastic Models and Risk*, John Wiley & Sons, New York.

- [18] Nelsen, R.B. (1999). *An Introduction to Copulas, Lecture Notes in Statistics*, Springer Verlag, Berlin, Vol. 139.
- [19] Rank, J. (ed) (2007). *Copulas: from Theory to Applications in Finance*, Risk Books.
- [21] Schmidt, T. (2007). Coping with copulas, in *Copulas: From Theory to Applications in Finance*, J. Rank, ed, Risk Books, pp. 1–31.
- [20] Schmidt, F. & Schmidt, R. (2007). Multivariate conditional versions of Spearmans' rho, *Journal of Multivariate Analysis* **98**, 1123–1140.
- [22] Sklar, A. (1959). Fonctions de répartition à n dimensions e leurs marges, *Publications de l'Institut de Statistique de l'Université de Paris* **8**, 229–231.
- [23] Taylor, M.D. (2007). Multivariate measures of concordance, *Annals of the Institute of Statistical Mechanics* **59**, 789–806.

Related Articles

Copulas in Econometrics; Copulas in Insurance; Correlation Risk; Lévy Copulas; Operational Risk; Risk Exposures.

THORSTEN SCHMIDT

Correlation Risk

Correlation

Correlation is a well-known concept for measuring the linear relationship between two or more variables. It plays a major role in a number of classical approaches in finance: the capital asset pricing model (see **Capital Asset Pricing Model**) as well as arbitrage pricing theory (APT) rely on correlation as a measure for the dependence of financial assets (see [4]). In the multivariate Black–Scholes model correlation of the log-returns is used as a measure of the dependence between assets, [2, 5, 14]. The main reason for the importance of correlation in these frameworks is that the considered random variables (rvs) obey—under an appropriate transformation—a multivariate normal distribution. Correlation is moreover a key driver in portfolio credit models, and the term *default correlation* has been coined for this. Correlation as a measure of dependence fully determines the dependence structure for normal distributions and, more generally, elliptical distributions while it fails to do so outside this class. Even within this class, correlation has to be handled with care; while a correlation of zero for multivariate normally distributed rvs implies independence, a correlation of zero for, say, t -distributed rvs does *not* imply independence (compare the following paragraph on correlation pitfalls). More general measures of dependence help to avoid these pitfalls. Examples of more general measures for dependence are rank correlation, the coefficient of tail dependence, and association (see **Copulas: Estimation**).

Hence, approaches relying on multivariate Brownian motions and transformations thereof naturally determine the dependence structure *via* correlation. Extending this, there are a number of approaches generalizing the simple linear correlation to a time-varying and stochastic concept (see [3, 8] and references therein).

For two random variables X and Y with finite and positive variances, their *correlation* is defined as

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} \quad (1)$$

where

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] \quad (2)$$

is the *covariance* of X and Y . We state some properties of correlation: $\text{Corr}(X, Y)$ is a number in $[-1, 1]$ and it is equal to 1 or -1 if and only if X and Y are linearly related, that is, $Y = a + bX$ for constants a, b with $b \neq 0$. The correlation is -1 if $b < 0$ and 1 if $b > 0$. For constants a, b

$$\text{Corr}(X + a, Y + b) = \text{Corr}(X, Y) \quad (3)$$

If X and Y are independent, then $\text{Corr}(X, Y) = 0$. On the other hand, if $\text{Corr}(X, Y) = 0$, X and Y are called *uncorrelated*. In the case when (X, Y) has a bivariate normal distribution, this implies independence of X and Y . Otherwise, this implication is typically wrong: even when X and Y are normally distributed (but (X, Y) does not have a bivariate normal distribution—compare copulas and dependence measurement for an exposition how this may be achieved using copulas), $\text{Corr}(X, Y) = 0$ does not imply independence.

For two random variables belonging to a given class of elliptical distributions,^a which includes the normal distribution and the Student t -distribution, correlation fully determines the dependence structure. However, note that uncorrelated t -distributed random variables are *not* independent.

If X is m -dimensional and Y n -dimensional, then $\text{Cov}(X, Y)$ is given by the $m \times n$ -matrix with entries $\text{Cov}(X_i, Y_j)$. $\Sigma = \text{Cov}(X, X)$ is called *covariance matrix*. Σ is symmetric and positive semidefinite, that is, $x^\top \Sigma x \geq 0$ for all $x \in \mathbb{R}^m$. Moreover, one has

$$\text{Cov}(\mathbf{a} + BX, \mathbf{c} + DY) = B \text{Cov}(X, Y) D^\top \quad (4)$$

for $\mathbf{a} \in \mathbb{R}^o$, $\mathbf{c} \in \mathbb{R}^p$, $B \in \mathbb{R}^{o \times m}$, $D \in \mathbb{R}^{p \times n}$. Similarly, $\text{Corr}(X, Y)$ has the entries $\text{Corr}(X_i, Y_j)$, $1 \leq i \leq m, 1 \leq j \leq n$. The *correlation matrix* of X is $\text{Corr}(X, X)$. It is again symmetric and positive semidefinite. Correlation is invariant under linear increasing transformations such that

$$\text{Corr}(a + bX, c + dY) = \text{Corr}(X, Y) \quad (5)$$

if $bc > 0$. If $bc < 0$ only the sign of the correlation changes.

Correlation Pitfalls

When correlation is used as a measure of dependence a number of pitfalls arise; compare [7] or [13], Chapter 5.2.1 for a detailed exposition. In the following,

we list the typical pitfalls and give a hint why difficulties may arise when linear correlation is used.

1. *A correlation of 0 is not equivalent to independence.* For (X, Y) being jointly normal, $\text{Corr}(X, Y) = 0$ implies independency of X and Y . In general this is not true; even perfectly related rvs can have zero correlation: consider $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$. Then $\text{Corr}(X, Y) = 0$ and X and Y are clearly not independent.
2. *Correlation is invariant under linear transformations, but not under general transformations.* For example, two lognormal rvs have a different correlation than the underlying normal rvs; compare Example 5.26 in [13].
3. *For given distributions of X and Y and some given correlation in $[-1, 1]$ it is, in general, not possible to construct a joint distribution.* For example, if X and Y are lognormally distributed, the interval of attainable correlations is a strict subset of $[-1, 1]$ and becomes smaller with increasing volatility; compare again Example 5.26 in [13].
4. *A small correlation does not imply a small degree of dependency.* This is, in particular, implied by observation 3, and so it is, in general, wrong to deduce a small degree of dependency from a small correlation.

Stylized Facts

Asset correlation shows two typical stylized features:^b

1. *Correlation clustering:* periods of high (low) correlation are likely to be followed by periods of high (low) correlation.
2. *Asymmetry and comovement with volatility:* high volatility in falling markets goes hand in hand with a strong increase in correlation, but this is not the case for rising markets; see [12] or [1].

In [15] the 1987 crash is analyzed and correlation risk is identified as a reason of simultaneous stock-market declines and volatility increases. Notably, this reduces opportunities for diversification in stock-market declines.

Estimating Correlation

The estimation of correlation in financial data is a delicate task as the underlying distribution typically has heavy tails. If this is the case, it is preferable to use robust methods in comparison to nonrobust methods like the sample correlation. Evidence of the efficiency of robust methods like Kendall's rank correlation coefficient is provided in [13], Section 3.3.4. Acknowledging that correlation changes over time, a number of approaches for dynamic correlation have been developed; see, for example, [3, 8] and references therein.

Correlation Risk

Correlation risk refers to the risk of a financial loss when correlation in the market changes. It plays a central role in risk management and the pricing of basket derivatives.

Risk Management

In risk management, correlation risk refers to the risk of a loss in a financial position occurring due to a difference between anticipated correlation and realized correlation. In particular, this occurs when the estimate of correlation was wrong or the correlation in the market changed.

The risk management of a portfolio as well as portfolio optimization heavily depends on the used correlation. This is illustrated by the following simple example: assume that a financial position is given by portfolio weights $w_1, \dots, w_d$ and the distribution of the assets X is multivariate normal. Then the profit and loss (P&L) of the position is $\sum_{i=1}^d w_i X_i$, and is hence normally distributed with mean $\sum_{i=1}^d w_i E(X_i)$ and variance

$$\sum_{i,j=1}^d w_i w_j \text{Cov}(X_i, X_j) \quad (6)$$

which equals $\sum_{i=1}^d w_i^2 \text{Var}(X_i)$ if the positions are uncorrelated. Otherwise, the Value at Risk depends on the correlations of all assets and therefore a change in the correlation may significantly alter the risk of the position.

In the elliptical world, the use of coherent risk measures is related to the Markowitz approach where

the variance is used as a risk measure; see [13], Section 6.1.5. The effects of stochastic correlation on hedging strategies have also been considered in [9]. Section 5 therein also gives some examples on stochastic correlation.

Basket Derivatives

If the pricing of basket derivatives (*see Foreign Exchange Basket Options; Basket Options*) is considered, the value of the derivative itself depends on the (unknown) correlation. In this case, correlation risk refers to the change in the value of the derivative with changing correlation. Note that in contrast to the first example, already the *value* of the derivative itself depends on the correlation. Options of this kind typically are traded in interest markets, foreign exchange markets, or credit markets, examples being quanto (*see Quanto Options*) options, rainbow options, spread options, or collateralized debt obligations (*see Collateralized Debt Obligations (CDO)*). In particular, in credit markets, the term *default correlation* has been coined and contagion and dependencies play a central role. For more references on the role of default correlation risk in credit risky markets, see [10, 11].

By analyzing prices of options on single names and on market indices, Driessen *et al.* [6] show that correlation risk is priced in the options markets. Practitioners trade priced correlation risk by using short positions in index options and long positions in individual options, which is called *dispersion trading* (*see Dispersion Trading*). A similar position can be taken with a correlation swap (*see Correlation Swap*).

In particular, in credit markets, contagion and dependencies play a central role. For more references on the role of default correlation risk in credit risky markets, see [10, 11].

End Notes

^aCompare [13], Theorem 3.25. Section 3.3 therein gives a short introduction to elliptical distributions.

^bSee, for example, [16] and references therein.

References

- [1] Andersen, T.G., Bollerslev, T., Diebold, F.X. & Ebense, H. (2001). The distribution of realized stock return volatility, *Journal of Financial Economics* **61**, 43–76.
- [2] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.
- [3] Campbell, R.A.J., Forbes, C.S., Koedijk, K.G. & Kofman, P. (2008). Increasing correlations or just fat tails, *Journal of Empirical Finance* **15**, 287–309.
- [4] Campbell, J.W., Lo, A.W. & MacKinlay, A.C. (1997). *The Econometrics of Financial Markets*, University Presses of CA.
- [5] Carmona, R. & Durlleman, V. (2006). Generalizing the Black–Scholes formula to multivariate contingent claims, *Journal of Computational Finance* **9**, 43–67.
- [6] Driessen, J., Maenhout, P. & Vilkov, G. The Price of Correlation Risk: Evidence from Equity Options, *Journal of Finance*, forthcoming (June 2009).
- [7] Embrechts, P., McNeil, A.J. & Straumann, D. (2001). Correlation and dependency in risk management: properties and pitfalls, in *Risk Management: Value at Risk and Beyond*, M. Dempster & H.K. Moffatt, eds, Cambridge University Press, pp. 176–223.
- [8] Engle, R.F. (2002). Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroskedasticity models, *Journal of Business and Economic Statistics* **20**, 339–350.
- [9] Gozzi, F. & Vargiolu, T. (2002). Superreplication of European multiasset derivatives with bounded stochastic volatility, *Mathematical Methods of Operations Research* **55**, 69–91.
- [10] Gregory, J. & Laurent, J.P. (2004). In the core of correlation, *Risk* **17**(10), 87–91.
- [11] Lipton, A. & Rennie, A. (eds) (2007). *Credit Correlation: Life After Copulas*, World Scientific.
- [12] Longin, F. & Solnik, B. (2001). Extreme correlation of international equity markets, *Journal of Finance* **56**, 649–676.
- [13] McNeil, A.J., Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton University Press.
- [14] Merton, R. (1974). On the pricing of corporate debt: the risk structure of interest rates, *Journal of Finance* **29**, 449–470.
- [15] Rubinstein, R. (2001). Comments on the 1987 stock-market crash: 11 years later, *Investment Accumulation Products of Financial Institutions*, The Society of Actuaries.
- [16] Skintzi, V.D. & Refenes, A.-P.N. (2005). Implied correlation index: a new measure of diversification, *The Journal of Future Markets* **25**, 171–1197.
- Merton, R. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.

Further Reading

Related Articles

Copulas: Estimation; Correlation Swap; Dispersion Trading; Lévy Copulas; Multivariate Distributions.

THORSTEN SCHMIDT

Multivariate Distributions

Financial risk models, whether for market or credit risks, are inherently multivariate. For instance, the value change of a portfolio of traded instruments over a fixed time horizon depends on a random vector of risk-factor changes or returns; similarly, the loss incurred by a credit portfolio depends on a random vector of losses for the individual counterparties in the portfolio. In this article, we review some multivariate distributions (models for the distribution of random vectors) that are particularly useful for financial data. Our discussion is based mainly on Chapter 3 of [5]; the closely related issue of copulas and dependence measurement is discussed in **Copulas: Estimation**.

Random Vectors and Their Distributions

In this section, we briefly review some key notions from multivariate statistics.

Joint and Marginal Distributions

Consider a general d -dimensional random vector (think of risk-factor changes or log-returns) $\mathbf{X} = (X_1, \dots, X_d)'$. The dependence between the components of $\mathbf{X}$ is completely described by the *joint distribution function* (df)

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &= F_{\mathbf{X}}(x_1, \dots, x_d) = P(\mathbf{X} \leq \mathbf{x}) \\ &= P(X_1 \leq x_1, \dots, X_d \leq x_d) \end{aligned} \quad (1)$$

The *marginal* distribution function of X_i , written as F_{X_i} or often simply as F_i , is the df of that risk factor considered individually and is easily calculated from the joint df via

$$\begin{aligned} F_i(x_i) &= P(X_i \leq x_i) \\ &= F(\infty, \dots, \infty, x_i, \infty, \dots, \infty) \end{aligned} \quad (2)$$

where the last expression is understood as the limit, as the respective arguments tend to the upper boundaries of the support of the distribution. If the marginal df $F_i(x)$ is absolutely continuous, then we refer to its derivative $f_i(x)$ as the *marginal density* of X_i . The

df of a random vector $\mathbf{X}$ is said to be *absolutely continuous* if

$$\begin{aligned} F(x_1, \dots, x_d) \\ = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(u_1, \dots, u_d) du_1 \dots du_d \end{aligned} \quad (3)$$

for some nonnegative function f integrating to one, known as the *joint density* of $\mathbf{X}$.

Conditional Distributions and Independence

If we have a multivariate model for risks in the form of a joint df or density, we can make conditional probability statements about the probability that certain components take certain values given that other components take other values. More precisely, partition $\mathbf{X}$ into $(\mathbf{X}'_1, \mathbf{X}'_2)'$, where $\mathbf{X}_1 = (X_1, \dots, X_k)'$ and $\mathbf{X}_2 = (X_{k+1}, \dots, X_d)'$. Assume that $\mathbf{X}$ is absolutely continuous with density f and denote by $f_{\mathbf{X}_1}(\mathbf{x}_1) = \int f(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2$ the density of $\mathbf{x}_1$. Then the conditional distribution of $\mathbf{X}_2$ given $\mathbf{X}_1 = \mathbf{x}_1$ has density

$$f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{x}_1) = \frac{f(\mathbf{x}_1, \mathbf{x}_2)}{f_{\mathbf{X}_1}(\mathbf{x}_1)} \quad (4)$$

The components of $\mathbf{X}$ are called *mutually independent*, if and only if $F(\mathbf{x}) = \prod_{i=1}^d F_i(x_i)$ for all $\mathbf{x} \in \mathbb{R}^d$ or, in the case where $\mathbf{X}$ possesses a density, $f(\mathbf{x}) = \prod_{i=1}^d f_i(x_i)$.

Moments and Characteristic Function

The *mean vector* of $\mathbf{X}$, when it exists, is given by $E(\mathbf{X}) := (E(X_1), \dots, E(X_d))'$; the *covariance matrix*, when it exists, is the matrix $\Sigma = \text{cov}(\mathbf{X})$ with (i, j) th element given by $\sigma_{ij} = \text{cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$, the ordinary pairwise covariance between X_i and X_j . The diagonal elements $\sigma_{11}, \dots, \sigma_{dd}$ are the variances of the components of $\mathbf{X}$. Mean vectors and covariance matrices are extremely, easily manipulated under linear operations on the vector $\mathbf{X}$. For any matrix $B \in \mathbb{R}^{k \times d}$ and vector $\mathbf{b} \in \mathbb{R}^k$, we have

$$E(B\mathbf{X} + \mathbf{b}) = BE(\mathbf{X}) + \mathbf{b} \quad (5)$$

$$\text{cov}(B\mathbf{X} + \mathbf{b}) = B \text{cov}(\mathbf{X}) B' \quad (6)$$

Covariance matrices are symmetric and positive semidefinite. In case that Σ is positive definite ($\mathbf{a}'\Sigma\mathbf{a} > 0$ for any $\mathbf{a} \in \mathbb{R}^d \setminus \{0\}$), we can use the well-known *Cholesky factorization* $\Sigma = AA'$ for a lower triangular matrix A with positive diagonal elements known as the *Cholesky factor*; this decomposition is very useful for simulation purposes.

An important tool for studying multivariate distributions is the *characteristic function* of a random vector (or its distribution), given by

$$\phi_{\mathbf{X}}(\mathbf{t}) = E\left(e^{i\mathbf{t}'\mathbf{X}}\right), \quad \mathbf{t} \in \mathbb{R}^d \quad (7)$$

The Multivariate Normal Distribution

The multivariate normal distribution is central to much of classical multivariate analysis and provided the starting point for attempts to model market risk via the variance–covariance method (see **Market Risk** or Chapter 2 of [5]); moreover, it is an important building block for constructing more refined distributions.

Definition and Basic Properties

A random vector $\mathbf{X} = (X_1, \dots, X_d)'$ has a *multivariate normal* or *Gaussian* distribution if

$$\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + A\mathbf{Z} \quad (8)$$

where $\mathbf{Z} = (Z_1, \dots, Z_k)'$ is a vector of independent and identically distributed (iid) univariate *standard* normal random variables (rvs) (mean zero and variance one), and $A \in \mathbb{R}^{d \times k}$ and $\boldsymbol{\mu} \in \mathbb{R}^d$ are the matrix and vector of constants, respectively. It is easy to verify, using equations (5) and (6), that the mean vector of this distribution is $E(\mathbf{X}) = \boldsymbol{\mu}$ and the covariance matrix is $\text{cov}(\mathbf{X}) = \Sigma$, where $\Sigma = AA'$ is a positive semidefinite matrix. Moreover, using the fact that the characteristic function of a standard univariate normal variate Z is $\phi_Z(t) = \exp(-t^2/2)$, the characteristic function of $\mathbf{X}$ may be calculated to be

$$\phi_{\mathbf{X}}(\mathbf{t}) = E\left(e^{i\mathbf{t}'\mathbf{X}}\right) = \exp\left(i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}\right), \quad \mathbf{t} \in \mathbb{R}^d \quad (9)$$

Clearly, the distribution is characterized by its mean vector and covariance matrix, and hence a standard

notation is $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$. Note that the components of $\mathbf{X}$ are mutually independent if and only if Σ is diagonal. For example, $\mathbf{X} \sim N_d(\mathbf{0}, I_d)$ if and only if $X_1, \dots, X_d$ are iid $N(0, 1)$, the standard univariate normal distribution.

We concentrate on the *nonsingular* case of the multivariate normal when $\text{rank}(A) = d \leq k$. In this case, the covariance matrix Σ has full rank d and is therefore invertible (nonsingular) and positive definite. Moreover, $\mathbf{X}$ has an absolutely continuous distribution function with joint density given by

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \times \exp\left\{-\frac{(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})}{2}\right\}, \quad \mathbf{x} \in \mathbb{R}^d \quad (10)$$

where $|\Sigma|$ denotes the determinant of Σ . In the singular case, the support of the distribution of $\mathbf{X}$ is a strict subspace of $\mathbb{R}^d$ and a joint density does not exist. The form of the density clearly shows that points with equal density lie on *ellipsoids* determined by equations of the form $(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c$, for constants $c > 0$.

In order to generate a realization $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ with Σ positive definite, one would proceed as follows:

1. Perform a Cholesky decomposition $\Sigma = AA'$ of the covariance matrix Σ (see, e.g., [6]).
2. Generate a vector $\mathbf{Z} = (Z_1, \dots, Z_d)'$ of independent standard normal variates and set $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Z}$.

We now summarize further useful properties of the multivariate normal that underline the attractiveness of this distribution for computational work in risk management. *Linear combinations* of multivariate normal random vectors are multivariate normal; if $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$, we have for any $B \in \mathbb{R}^{k \times d}$ and $\mathbf{b} \in \mathbb{R}^k$ that

$$B\mathbf{X} + \mathbf{b} \sim N_k(B\boldsymbol{\mu} + \mathbf{b}, B\Sigma B') \quad (11)$$

As a special case, if $\mathbf{a} \in \mathbb{R}^d$, then $\mathbf{a}'\mathbf{X} \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\Sigma\mathbf{a})$, and this fact is used routinely in the variance–covariance approach to risk management. Similarly, marginal distributions of a multivariate normal random vector are univariate normal.

Assuming Σ is positive definite, the *conditional distributions* of $\mathbf{X}_2$ given $\mathbf{X}_1$ and of $\mathbf{X}_1$ given $\mathbf{X}_2$ may also be shown to be multivariate normal. For example, $\mathbf{X}_2 | \mathbf{X}_1 = \mathbf{x}_1 \sim N_{d-k}(\boldsymbol{\mu}_{2.1}, \Sigma_{22.1})$ where

$$\begin{aligned}\boldsymbol{\mu}_{2.1} &= \boldsymbol{\mu}_2 + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \quad \text{and} \\ \Sigma_{22.1} &= \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}\end{aligned}\quad (12)$$

are the conditional mean vector and covariance matrix.

Multivariate Normal Distributions as Model for Asset Returns

There are numerous empirical tests of the suitability of the multivariate normal distribution as a model for log-returns (risk-factor changes) for financial data. These tests involve both testing the hypothesis that marginal log-returns are normal (univariate tests) and tests for multivariate normality; details and further references can, for instance, be found in Section 3.1.4 of [3]. Broadly speaking, the outcome of these tests points to three main defects of the multivariate normal distribution (at least for data over a shorter time horizon such as daily or even monthly data):

1. The tails of its univariate marginal distributions are too thin; they do not assign enough weight to *extreme* events.
2. The joint tails of the distribution do not assign enough weight to *joint extreme* outcomes.
3. The distribution has a strong form of symmetry known as *elliptical symmetry*.

In the next section, we look at models that address some of these defects.

Variance Mixtures and Elliptical Distributions

Normal Mixture Distributions

The random vector $\mathbf{X}$ is said to have a *multivariate normal variance mixture distribution* if

$$\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + \sqrt{W} \mathbf{A} \mathbf{Z} \quad \text{where} \quad (13)$$

$\mathbf{Z} \sim N_k(\mathbf{0}, \mathbf{I}_k)$; $\mathbf{A} \in \mathbb{R}^{d \times k}$ and $\boldsymbol{\mu} \in \mathbb{R}^d$ are a matrix, respectively, a vector of constants; and where $W \geq 0$

is a nonnegative, scalar-valued rv, which is independent of $\mathbf{Z}$. Such distributions are known as *variance mixtures*, since $\mathbf{X} | W = w \sim N_d(\boldsymbol{\mu}, w \Sigma)$ where $\Sigma = \mathbf{A} \mathbf{A}'$. The distribution of $\mathbf{X}$ can be considered as a composite distribution, constructed by taking a set of multivariate normal distributions with the same mean vector and with the same covariance matrix up to a multiplicative constant w ; the mixture distribution—which is generally not a normal distribution—is then constructed by drawing randomly from this set of component multivariate normals according to a set of “weights” determined by the distribution of W .

Provided W has a finite expectation, we may easily calculate that $E(\mathbf{X}) = \boldsymbol{\mu}$ and that $\text{cov}(\mathbf{X}) = E(W) \Sigma$. Note that normal variance mixtures provide nice examples of models where a lack of correlation does not necessarily imply independence of the components of $\mathbf{X}$; indeed, it can be shown that two uncorrelated rvs (X_1, X_2) having a normal mixture distribution with $E(W) < \infty$ are independent, if and only if W is almost surely constant, that is, if (X_1, X_2) are bivariate normally distributed.

The *characteristic function* of a multivariate normal variance mixture is given by

$$\phi_{\mathbf{X}}(\mathbf{t}) = \exp(i \mathbf{t}' \boldsymbol{\mu}) \widehat{H}(\mathbf{t}' \Sigma \mathbf{t} / 2) \quad (14)$$

where $\widehat{H}(\theta) = \int_0^\infty e^{-\theta v} dH(v)$ is the Laplace–Stieltjes transform of the df H of W . We use the notation $\mathbf{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \widehat{H})$ for normal variance mixtures. Assuming that Σ is positive definite and that the distribution of W places no point mass at zero, the density of $\mathbf{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \widehat{H})$ becomes

$$\begin{aligned}f(\mathbf{x}) &= \int \frac{w^{-\frac{d}{2}}}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \\ &\times \exp \left\{ -\frac{(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2w} \right\} dH(w)\end{aligned}\quad (15)$$

Note that all such densities depend on $\mathbf{x}$ only through the quadratic form $(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$.

The most important example is provided by the *multivariate t -distribution*. Here, we take W in equation (13) to be a rv with an inverse gamma distribution $W \sim \text{Ig}(\nu/2, \nu/2)$ (which is equivalent to

saying that $\nu/W \sim \chi_\nu^2$). The notation for this distribution is $\mathbf{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$. Since $E(W) = \nu/(\nu - 2)$, we have $\text{cov}(\mathbf{X}) = \frac{\nu}{\nu - 2} \Sigma$ and the covariance matrix of this distribution is only defined if $\nu > 2$. Using equation (15), the density can be calculated to be

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{\nu + d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) (\pi\nu)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \times \left(1 + \frac{(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+d}{2}} \quad (16)$$

The multivariate t has heavier marginal tails than the normal distribution and a more pronounced tendency to generate simultaneous extreme outcomes (see also **Copulas: Estimation**).

Normal variance mixture distributions are easy to work with under linear operations; if $\mathbf{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{H})$ and $\mathbf{Y} = B\mathbf{X} + \mathbf{b}$ where $B \in \mathbb{R}^{k \times d}$ and $\mathbf{b} \in \mathbb{R}^k$, then $\mathbf{Y} \sim M_k(B\boldsymbol{\mu} + \mathbf{b}, B\Sigma B', \hat{H})$. Thus, linear transformations of $\mathbf{X}$ remain in the subclass of mixture distributions specified by $\hat{H}$ or, equivalently, by the mixing variable W of which $\hat{H}$ is the Laplace–Stieltjes transform. For example, if $\mathbf{X}$ has a multivariate t distribution with ν degrees of freedom, then so does any linear transformation of $\mathbf{X}$.

Normal Mean–variance Mixture Distributions

The definition (13) can be generalized to allow for skewed distributions: $\mathbf{X}$ is said to have a multivariate normal mean–variance mixture distribution if

$$\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + W\boldsymbol{\gamma} + \sqrt{W}A\mathbf{Z} \quad (17)$$

For instance, if we assume $W \sim N^-(\lambda, \chi, \psi)$ (a generalized inverse Gaussian distribution), we obtain the class of (multivariate) *generalized hyperbolic distributions* (see **Generalized Hyperbolic Models**). These distributions have important applications in finance and risk management; in particular, they are infinitely divisible and therefore directly connected to *Lévy processes* (see **Lévy Processes**). Note that mean–variance mixture distributions are, in general, skewed; in particular, the density is no longer constant on ellipsoids. We refer to Section 3.2.4 of [5] for further information and further references about this class.

Spherical and Elliptical Distributions

Many important distribution have densities that are constant on ellipsoids and thus belong to the class of *elliptical distributions*; this class has some theoretical properties that makes it attractive as a model for the joint distribution of risk factors. Elliptical distributions are defined as affine transformations of *spherical* random vectors. A random vector $\mathbf{X} = (X_1, \dots, X_d)'$ has a spherical distribution, if, for every orthogonal map $U \in \mathbb{R}^{d \times d}$ (i.e., maps satisfying $UU' = U'U = I$), one has $U\mathbf{X} \stackrel{d}{=} \mathbf{X}$. Spherical random vectors are thus distributionally invariant under rotations. It can be shown that $\mathbf{X}$ is spherical if and only if one of the following properties hold:

1. There exists a function ψ of a scalar variable, called *characteristic generator*, such that, for all $\mathbf{t} \in \mathbb{R}^d$,

$$\phi_{\mathbf{X}}(\mathbf{t}) = E\left(e^{i\mathbf{t}'\mathbf{X}}\right) = \psi(\mathbf{t}'\mathbf{t}) = \psi(t_1^2 + \dots + t_d^2) \quad (18)$$

2. For every $\mathbf{a} \in \mathbb{R}^d$, one has $\mathbf{a}'\mathbf{X} \stackrel{d}{=} \|\mathbf{a}\|X_1$, where $\|\mathbf{a}\|^2 = \mathbf{a}'\mathbf{a} = a_1^2 + \dots + a_d^2$.
 $\mathbf{X}$ has an *elliptical distribution* if

$$\mathbf{X} \stackrel{d}{=} \boldsymbol{\mu} + A\mathbf{Y} \quad (19)$$

where $\mathbf{Y}$ is spherical and $A \in \mathbb{R}^{d \times k}$ and $\boldsymbol{\mu} \in \mathbb{R}^d$ are a matrix and vector of constants, respectively. The characteristic function can be written as $\phi_{\mathbf{X}}(\mathbf{t}) = e^{i\mathbf{t}'\boldsymbol{\mu}} \psi(\mathbf{t}'A\mathbf{t})$, where $\psi(\cdot)$ is the characteristic generator of $\mathbf{Y}$ introduced in equation (18). Examples include the multivariate normal distribution and, more generally, the multivariate variance mixture distributions (13). Elliptical distributions share many properties with the multivariate normal distribution: affine transformations of elliptical random vectors and, in particular, marginal distributions remain elliptical with the same generator ψ ; if we condition on one or several components of an elliptical random vector, the ensuing conditional distribution is elliptical, albeit, in general, with a different generator.

An important property of elliptical distributions for risk management purposes is the fact that Value at Risk (VaR) (see **Value-at-Risk**) is a *coherent risk measure* (see **Convex Risk Measures**) if restricted to the set of all linear combinations of the components of some elliptical random vector.

Notes and Further References

Much of the material from the sections “Random Vectors and Their Distributions” and “The Multivariate Normal Distribution” can be found in greater detail in standard texts on multivariate statistical analysis such as [4, 7]. There are countless possible tests of univariate normality and a good starting point is the entry on *Departures from Normality, Tests for* in volume 2 of the Encyclopedia of Statistics [3]; the latter source also contains tests for multivariate normality.

A comprehensive reference for the spherical and elliptical distributions (including a special treatment of symmetric variance mixture models) is given by Fang *et al.* [2]. A useful reference on the multivariate generalized hyperbolic distribution is Blæsild [1].

References

- [1] Blæsild, P. (1981). The two-dimensional hyperbolic distribution and related distributions, with an application to Johanssen’s bean data, *Biometrika* **68**(1), 251–263.

- [2] Fang, K.-T., Kotz, S. & Ng, K.-W. (1987). *Symmetric Multivariate and Related Distributions*, Chapman & Hall, London.
- [3] Kotz, S., Johnson, N. & Read, C. (eds) (1985). *Encyclopedia of Statistical Sciences*, Wiley, New York.
- [4] Mardia, K., Kent, J. & Bibby, J. (1979). *Multivariate Analysis*, Academic Press, London.
- [5] McNeil, A., Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton University Press, Princeton.
- [6] Press, W., Teukolsky, S., Vetterling, W. & Flannery, B. (1992). *Numerical Recipes in C*, Cambridge University Press, Cambridge.
- [7] Seber, G. (1984). *Multivariate Observations*, Wiley, New York.

Related Articles

Copulas: Estimation; Copulas in Econometrics; Correlation Risk.

RÜDIGER FREY

Electricity Markets

Since the early 1990s, an increasing number of countries worldwide have liberalized their electricity power sectors. Contrary to the situation that prevailed earlier, when power sectors were not open to competition and prices were set by regulators according to the cost of generation, transmission, and distribution, electricity prices are now determined by an equilibrium of supply and demand, which introduces a substantial price risk with volatilities much higher than those of equity prices. A big share of the total electricity in liberalized power markets is traded over the counter through bilateral agreements. There exists a rich variety of exotic options traded in this market. On the other hand, similar to how the end of the Bretton Woods system in 1973 caused the appearance of currency exchanges, the deregulation of electricity markets has led to the creation of organized electricity exchanges, where electricity is quoted almost as any other commodity (see Table 1 for a list of the major European electricity exchanges).

Although commonly referred to as *commodity*, electricity is in many ways different from the more classical commodities like oil, coal, metals, and agriculture. One substantial difference is that electricity has very limited storage possibilities. To a certain degree, producers may store electricity indirectly in water reservoirs (for hydro-based electricity production) or *via* gas, oil, or coal (for thermal electricity production). However, the consumer of electricity cannot buy for storage. This implies that electricity is not a tradable asset (in the sense that one can buy the asset and sell it later on), and usual hedging arguments to price futures/forwards and other derivatives cannot be applied. Other effects of the lack of storability are strong seasonal and very volatile price behavior. Also, because electricity is only useful when sourced continuously in time, all power contracts concern the delivery of electricity over a *period* of time and not at a fixed point in time. In this sense, electricity markets share more similarities with temperature and also natural gas markets than with the more classical commodity markets. Like electricity, both temperature and natural gas exhibit restricted storage possibilities (storage is impossible for temperature and rather costly for gas), and the corresponding traded contracts are of *flow-over-a-period* type.

Contracts traded on electricity exchanges can typically be divided into two categories: contracts with physical delivery and financially settled contracts. In the following, we shortly sketch the organization of the Scandinavian exchange Nord Pool, but the situation is similar at most other electricity exchanges.

Contracts with physical delivery include actual consumption or production of electricity as part of contract fulfillment. The market for physical delivery is supervised by a so-called transmission system operator (TSO) who balances supply and demand, and market participants are those with proper facilities for production or consumption. Further, contracts with physical delivery are organized in two different submarkets: the *real-time* and the *day-ahead* markets, known as the *two-settlement system*. On the day-ahead market, hourly power contracts for the next day's 24 h (midnight to midnight) are traded. Each day at noon, the day-ahead market is closed for bids and prices for each hour the next day are derived. Electricity prices on the day-ahead market are referred to as *spot prices* as they are reference prices for the financially settled futures/forward contracts. The real-time market, on the other hand, is organized for short-term upward or downward regulation and bids may be posted or changed close to operational time. In both the day-ahead and the real-time market, prices are derived in auctions. The TSO lists bids for each hour according to the price, the so-called *merit order*, and prices are derived by balancing supply and demand.

Financial power contracts are settled financially against a reference price. The market for financial electricity contracts does not require central coordination but can be considered as side bets on the physical system. In contrast to the physical market, a big share of the market players are speculators. The predominantly traded financial contracts are futures/forward type contracts written on the (weighted) average of the hourly spot price (day-ahead market) over a specific *delivery period*. At Nord Pool, there exist daily and weekly contracts of futures type with margin accounts, and monthly, quarterly, and yearly contracts of forward type. Besides futures/forwards contracts, Nord Pool's financial market also organizes trade in European call and put options written on futures/forwards, however, in much smaller volumes.

The substantial electricity price risk introduced by the liberalization of electricity markets requires a precise statistical modeling for risk management,

Table 1 Major European electricity exchanges

Country	Starting date	Name
England and Wales	1990	Electricity pool
Scandinavia	2001	UK Power Exchange (UKPX)
	1993	Nord Pool (Norway only)
	From 1996	Sweden, Denmark, Finland consecutively joined Nord Pool
Spain	1998	OMEL
Netherlands	1999	Amsterdam Power Exchange (APX)
Germany	2000	Leipzig Power Exchange (LPX)
	2001	European Power Exchange (EEX)
Poland	2000	Polish Power Exchange
France	2001	Powernext
Italy	2004	Gestore Mercato Elettrico (GME)

pricing, and asset evaluation purposes. Over recent years, a number of models have been proposed, which basically can be separated in two categories. Models in the first category aim at directly modeling the dynamics of the complete electricity futures curve and, to this end, techniques known from interest rate modeling have been transferred to electricity futures curve modeling (see e.g., [4, 5, 16, 17], and the book [6] with references therein). One of the main problems with this approach is that electricity futures curves seem to have far more complex dynamics than interest rate curves. It seems hardly possible to model the complete futures curve including the spot price in the short end with a reasonable number of factors.

Models in the second category aim at modeling the dynamics of the spot price of electricity. In this framework, the main difficulty when it comes to futures/forwards and other derivatives pricing is the nonstorability of electricity, which makes the market highly incomplete. The cost-of-carry relationship between spot and forward prices breaks down, and a pricing measure (equivalently, market price of risk) has to be identified. Also, since forward-looking information on a nonstorable asset is not reflected by actual price behavior of this asset, one has to be cautious about the market information modeling [7].

In the following text, we consider reduced-form electricity spot price models, which are members of the second category. In particular, in the section Stylized Feature of Electricity Spot Prices we present the stylized features of electricity spot prices to take into account when specifying a model, before we give a review of reduced-form spot models existing in

the literature in the section Reduced-form Electricity Spot Price Models.

Stylized Feature of Electricity Spot Prices

In Figure 1, a section of the daily Nord Pool spot price is shown, which behaves very nicely in the sense that it illustrates well the qualitative characteristics to take into account when specifying a spot price model. In [23], a systematic statistical analysis is performed on spot data from seven electricity exchanges (two American and five European) and the following list of five stylized features common to all data sets has been identified.

- **Seasonality**
Electricity spot prices reveal seasonal behavior in yearly, weekly, and daily cycles. However, the seasonality has little effect on the overall variability of the price data.
- **Stationarity**
Similar to other commodities, electricity prices tend to exhibit stationary behavior. They are mean reverting to a trend which, however, may exhibit slow stochastic variations (see Figure 2).
- **Multiscale autocorrelation**
The observed autocorrelation structure of most European price series is described quite precisely with a weighted sum of exponentials:

$$\sum_{i=1}^n w_i e^{-h\lambda_i} \quad (1)$$

where the number n of factors needed for a good description is 2 or 3, and the weights w_i add

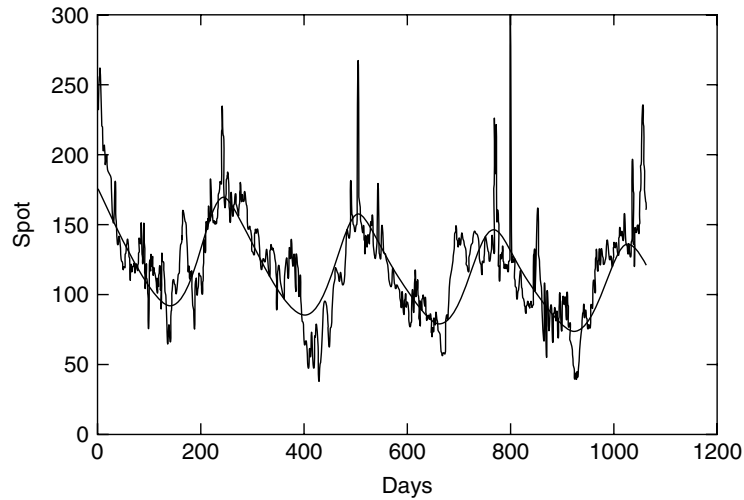


Figure 1 Daily spot price at Nord Pool spanning the period April 1, 1997–July 14, 2000

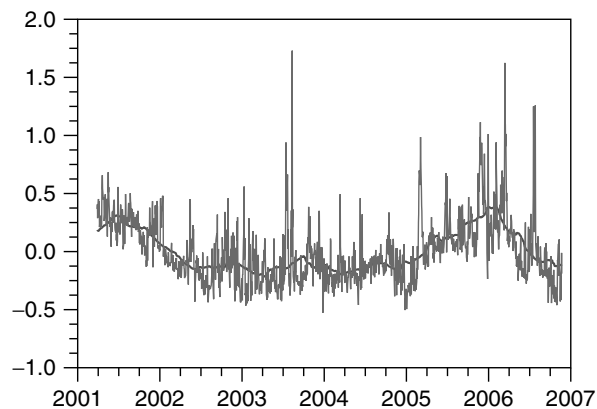


Figure 2 UKPX time series (after the seasonality has been removed) and its 6-month moving average

up to 1 (see Figure 3). We mention that the two American and also the Nord Pool price series exhibit a quite different, almost nonstationary, autocorrelation structure, which might be due to different market organization.

- **Spikes**

All data sets of electricity spot prices show impressive spikes, that is, violent upward jumps followed by rapid return to about the same level. The intensity of spike occurrence can vary over time. This fundamental property of electricity prices is due to the nonstorability

of this commodity, and any relevant spot price model must take this feature into account. In [19, 23], it is ascertained that appropriate modeling of spike risk requires a Pareto-like distribution with polynomial tail.

- **Non-Gaussianity**

The examination of daily spot prices reveals a highly non-Gaussian distribution, which tends to be slightly positively skewed and strongly leptokurtic. This high excess kurtosis is explained by the presence of the low-probability large-amplitude spikes.

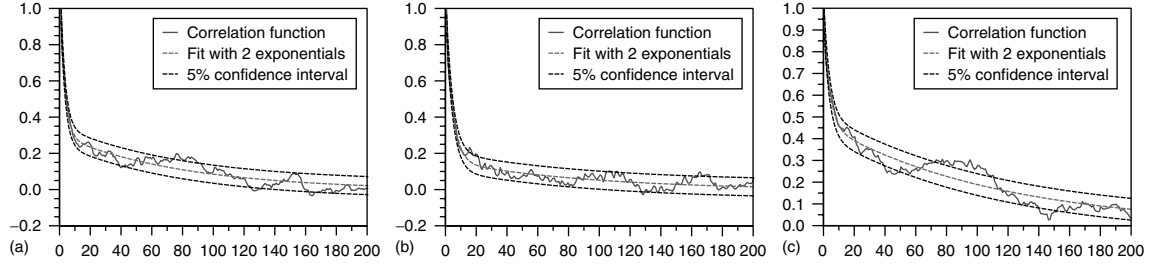


Figure 3 Fitting the autocorrelation function with a sum of two exponentials. (a–c) APX, EEX, UKPX

Reduced-form Electricity Spot Price Models

We conclude this article by presenting some spot price models existing in the literature. For this, we restrict to an overview as given in [23] of *reduced-form models* in continuous time, but we mention that there is a variety of different spot model types like *hybrid models* or *econometric time series models* (see, e.g., [20, 21, 24] or the text books [9, 12, 13, 26] with references therein).

Structural Models

Structural or equilibrium models as proposed in [1, 18] derive prices by balancing supply and demand. The (very inelastic) demand for electricity is described by a stochastic process

$$D_t = \overline{D}_t + X_t \quad (2)$$

$$dX_t = (\mu - \lambda X_t) dt + \sigma dW_t \quad (3)$$

where $\overline{D}_t$ describes the seasonal component and X_t corresponds to the stationary stochastic part. The price is obtained by matching the demand level with a deterministic supply function. In particular, the supply function must be nonlinear and strongly increasing in the right end to account for exploding cost of electricity generation (price spikes) in times of sudden rise in demand. Barlow [1] proposes

$$P_t = \left(\frac{a_0 - D_t}{b_0} \right)^{1/\alpha} \quad (4)$$

for some $\alpha > 0$, while Kanamura and Ohashi [18] suggest a “hockey stick” profile

$$P_t = (a_1 + b_1 D_t) 1_{D_t \leq D_0} + (a_1 + b_1 D_t) 1_{D_t > D_0} \quad (5)$$

Further, we mention in this category market equilibrium models as derived in [8, 15].

Markov Models

Geman and Roncoroni [14] model the electricity log-price as a one-factor Markov jump diffusion.

$$dP_t = \theta(\mu_t - P_t) dt + \sigma dW_t + h(t) dJ_t \quad (6)$$

The spikes are introduced by making the jump direction and intensity level-dependent: if the price is high, the jump intensity is high and downward jumps are more likely, whereas if the price is low, jumps are rare and upward-directed.

Regime-switching Models

In the one-factor Markov specification of Geman and Roncoroni [14], the “spike regime” is distinguished from the “base regime” by a deterministic threshold on the price process: if the price is higher than a given value, the process is in the “spike regime”; otherwise it is in the “base regime”. This threshold value may be difficult to calibrate and it is not very realistic to suppose that it is determined in advance. Regime-switching models as in [27] alleviate this problem by introducing a two-state unobservable Markov chain, which determines the transition from “base regime” to “spike regime” with greater volatility and faster mean reversion:

$$dP_t = \theta^1(\mu_t - P_t) + \sigma^1 dW_t \quad (\text{base regime}) \quad (7)$$

$$dP_t = \theta^2(\mu_t - P_t) + \sigma^2 dW_t \quad (\text{spike regime}) \quad (8)$$

Multifactor Ornstein–Uhlenbeck Models

Multifactor Ornstein–Uhlenbeck models describe the deseasonalized logarithmic spot price (geometric models), alternatively the deseasonalized spot

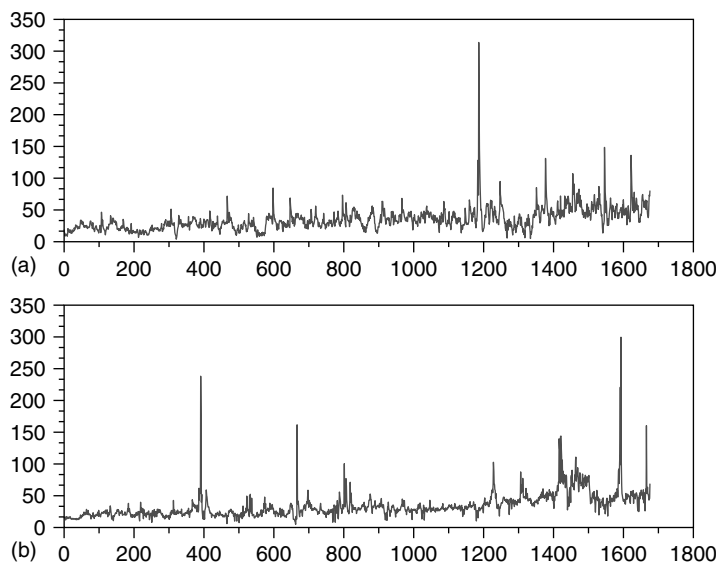


Figure 4 Comparison of the real EEX series (b) and the simulated series with estimated parameters (a)

price (arithmetic models), as a sum of independent Lévy-driven Ornstein–Uhlenbeck components:

$$X(t) = \sum_{i=1}^n Y_i(t) \quad (9)$$

$$dY_i(t) = -\lambda_i Y_i(t) dt + dL_i(t), \quad Y_i(0) = y_i \quad (10)$$

where processes $L_i(t)$ are independent, possibly time inhomogeneous Lévy processes. For example, in the case of a two-factor model, the first factor corresponds to the stochastic base signal with a slow rate of mean reversion λ_1 , and the second factor represents the spikes and has a high rate of mean reversion λ_2 . The earliest model in this family proposed in the literature is a Gaussian one-factor model in [25]. Subsequently, several authors have proposed various improvements [2, 3, 11, 22]. We also mention here the model in [10], which is a mixture of a structural and a multifactor model.

Multifactor Ornstein–Uhlenbeck models are capable of capturing all the stylized features presented in the previous section and to precisely describe daily spot price dynamics. In particular, the arithmetic model proposed in [3] can reproduce the multiscale autocorrelation structure of European spot prices, and, at the same time, it is mathematically very

tractable. However, due to their non-Markovianity, in case of several factors, estimation of multifactor models is not obvious. In [19, 23], we develop statistical estimation procedures based on separation of the data into a spike and a base component. In [23], this procedure is based on methods from nonparametric statistics, while we use tools from extreme value theory in [19]. Figure 4 shows a simulation of a two-factor model as a result of an estimation on EEX data in [19].

References

- [1] Barlow, M.T. (2002). A diffusion model for electricity prices, *Mathematical Finance* **12**, 287–298.
- [2] Barlow, M.T., Gusev, Y. & Manpo, L. (2004). Calibration of multifactor models in electricity markets, *International Journal of Theoretical and Applied Finance* **7**, 101–120.
- [3] Benth, F.E., Kallsen, J. & Meyer-Brandis, T. (2007). A non-Gaussian Ornstein-Uhlenbeck process for electricity spot price modeling and derivatives pricing, *Applied Mathematical Finance* **14**(2), 153–169.
- [4] Benth, F. & Koekebakker, S. (2008). Stochastic modeling of financial electricity contracts, *Energy Economics* **30**(3), 1116–1157.
- [5] Benth, F.E., Koekebakker, S. & Ollmar, F. (2008). Extracting and applying smooth forward curves from average-based commodity contracts with seasonal variation, *Journal of Derivatives* Fall, 52–66.

- [6] Benth, F., Koekebakker, S. & Saltyte-Benth, J. (2008). Stochastic modeling of electricity and related markets, in *Advanced Series on Statistical Science and Applied Probability*, O.E. Barndorff-Nielsen, ed., World Scientific, Vol. 11.
- [7] Benth, F. & Meyer-Brandis, T. (2009). The information premium in electricity markets, *Journal of Energy Markets* 2(3), accepted.
- [8] Bessembinder, H. & Lemmon, M. (2002). Equilibrium pricing and optimal hedging in electricity forward markets, *Journal of Finance* 57, 1347–1382.
- [9] Burger, M., Graeber, B. & Schindlmayr, G. (2007). *Managing Energy Risk*, John Wiley & Sons, Chichester.
- [10] Burger, M., Klar, B., Muller, A. & Schindlmayer, G. (2004). A spot market model for pricing derivatives in electricity markets, *Quantitative Finance* 4, 109–122.
- [11] Deng, S.-J. & Jiang, W. (2005). Lévy process-driven mean-reverting electricity price model: the marginal distribution analysis, *Decision Support Systems* 40, 483–494.
- [12] Eydeland, A. & Wolyniec, K. (2003). *Energy and Power Risk Management*, John Wiley & Sons, Chichester.
- [13] Geman, H. (2005). *Commodities and Commodity Derivatives*, Wiley-Finance, John Wiley & Sons, Chichester.
- [14] Geman, H. & Roncoroni, A. (2006). Understanding the fine structure of electricity prices, *Journal of Business* 79, 1225–1261.
- [15] Hinz, J. (2003). Modeling day ahead prices, *Applied Mathematical Finance* 10(2), 149–161.
- [16] Hinz, J., von Grafenstein, L., Verschuere, M. & Wilhelm, M. (2005). Pricing electricity risk by interest rate methods, *Quantitative Finance* 5(1), 49–60.
- [17] Hinz, J. & Wilhelm, M. (2006). Pricing flowcommodity derivatives using fixed income market techniques, *International Journal of Theoretical and Applied Finance* 9(8), 1299–1321.
- [18] Kanamura, T. & Ohashi, K. (2007). A structural model for electricity prices with spikes: measurement of spike risk and optimal policies for hydropower plant operation, *Energy Economics* 29(5), 1010–1032.
- [19] Klüppelberg, C., Meyer-Brandis, T. & Schmidt, A. (2008). Electricity spot price modelling with a view towards extreme spike risk, *Quantitative Finance* accepted.
- [20] Knittel, C. & Roberts, M. (2001). *An Empirical Examination of Deregulated Electricity Prices*. POWER WP-087, University of California Energy Institute.
- [21] Leon, A. & Rubia, A. (2004). Testing for weekly seasonal unit roots in the Spanish power pool, in *Modelling Prices in Competitive Electricity Markets*, D.W. Bunn, ed., Wiley Series in Financial Economics, Wiley, pp. 177–189.
- [22] Lucia, J. & Schwartz, E.S. (2002). Electricity prices and power derivatives: evidence from the Nordic Power Exchange, *Review of Derivatives Research* 5(1), 5–50.
- [23] Meyer-Brandis, T. & Tankov, P. (2008). Multi-factor jump-diffusion models of electricity prices, *IJTAF* 11(5), 503–528.
- [24] Misiorek, A., Trueck, S. & Weron, R. (2006). Point and interval forecasting of spot electricity prices: Linear vs. non-linear time series models, *Studies in Nonlinear Dynamics and Econometrics* 10(3), Article 2.
- [25] Schwartz, E.S. (1997). The stochastic behavior of commodity prices: implications for valuation and hedging, *Journal of Finance* 52(3), 923–973.
- [26] Weron, R. (2006). *Modeling and Forecasting Electricity Loads and Prices – A Statistical Approach*, John Wiley & Sons, Chichester.
- [27] Weron, R. (2009). Heavy-tails and regime-switching in electricity prices, *Mathematical Methods of Operations Research* 69(3), 457–473.

Related Articles

Commodity Forward Curve Modeling; Commodity Price Models; Electricity Forward Contracts; Stylized Properties of Asset Returns; Swing Options.

THILO MEYER-BRANDIS

Electricity Forward Contracts

In quantitative modeling of the electricity market, a fundamental question is to establish the link between spot and forward prices. The classical theory of arbitrage, including storage and convenience yield, does not easily carry over to electricity because of a single complicating fact: the nonstorability of the commodity. In this article, we discuss the spot-forward relation for electricity in view of the classical theory.

Forward Pricing by No Arbitrage

Let us recall how to derive forward prices using no-arbitrage arguments. Denote by $S(t)$ the spot price of an asset at time $t \geq 0$, and consider a forward contract with delivery at time $T \geq t$. To prevent arbitrage opportunities, the forward price $f(t, T)$ at time t must be

$$f(t, T) = S(t) \exp(r(T - t)) \quad (1)$$

with $r > 0$ being the risk-free interest rate. If this is not the case, we can exploit a buy-and-hold strategy to obtain arbitrage. For example, if $f(t, T) > S(t) \exp(r(T - t))$, finance a purchase of the spot contract with borrowing money and go short the forward. At delivery, hand over the spot, pay the debt with continuously compounding interest r , and receive the forward price. This entails in a sure win $f(t, T) - S(t) \exp(r(T - t))$. Obviously, if the forward price is less than $S(t) \exp(r(T - t))$, the strategy above is reversed. A fundamental assumption in this trading scheme is the storability of the spot.

If $S(t)$ is a *semimartingale process* on a *filtered probability space* $(\Omega, (\mathcal{F}_t), \mathcal{F}, P)$, there exists *risk neutral probabilities* Q such that the discounted spot price is a Q -martingale. The forward price can be represented alternatively as a conditional expectation with respect to Q , that is,

$$f(t, T) = \mathbb{E}_Q[S(T) | \mathcal{F}_t] \quad (2)$$

where $\mathcal{F}_t$ is the given information filtration in the market. We remark in passing that, in general, we have many pricing measures Q for a semimartingale

process, but because of the buy-and-hold strategy we get a unique forward price.

The *risk premium* is defined as the difference between the forward price and the predicted spot price at delivery, that is,

$$P(t, T) \triangleq f(t, T) - \mathbb{E}[S(T) | \mathcal{F}_t] \quad (3)$$

Letting $S(t)$ be a geometric Brownian motion,

$$dS(t) = \mu S(t) dt + \sigma S(t) dB(t) \quad (4)$$

we easily find that

$$P(t, T) = S(t) (e^{r(T-t)} - e^{\mu(T-t)}) \quad (5)$$

which is negative in “normal” markets since the expected return μ is naturally bigger than the risk-free interest rate. Hence, the forward price is below the expected spot, which is referred to as the forward market being in *normal backwardation*. In economical terms, this is explained as the producers wanting to secure their future income and hence creating a hedging pressure in the market. The opposite situation, when the risk premium is positive, is referred to as the forward market being *contango*.

One can easily imagine frictions in the market preventing the feasibility of the buy-and-hold strategy. Typically, in commodity markets, costs related to storage and transportation are incurred. The price for storage leads to a modification of the forward price. One also talks of a *convenience yield* in commodity markets, understanding a discount in the forward price due to the *convenience* to own the underlying spot. In effect, these modifications lead to a forward price

$$f(t, T) = S(t) \exp((r - \delta)(T - t)) \quad (6)$$

where δ is a factor accounting for storage, convenience yield, transportation, and so on.

The forward price can still be expressed in terms of a conditional expectation as in equation (2), but now we use pricing measures Q , which do not necessarily turn the discounted spot dynamics into a Q -martingale. Hence, by a specific choice of Q one may represent equation (6) as in equation (2). Considering the specific case of a geometric Brownian motion again, equation (6) is obtained by using the *Girsanov transformation* to introduce a measure Q under which

$$dW(t) = \frac{\mu - (r - \delta)}{\sigma} dt + dB(t) \quad (7)$$

is a Brownian motion. Note that this change of measure does not turn the discounted spot price into a Q -martingale. However, more importantly, normal backwardation is no longer the immediate implication, since the sign of the risk premium

$$P(t, T) = S(t) (e^{(r-\delta)(T-t)} - e^{\mu(T-t)}) \quad (8)$$

is determined by the relation between r , δ , and μ which is not *a priori* clear. Thus, we may have a market either in contango or in backwardation; however, we cannot have both in the sense that the sign of the risk premium changes with time to maturity of the contracts.

The Case of Electricity

In real electricity markets, the forward and futures contracts^a deliver the power over a period rather than at a specific future time point. Thus, a long position entered at time t in an electricity forward delivering over the period $[\tau_1, \tau_2]$, $t \leq \tau_1$ gives a profit/loss

$$\int_{\tau_1}^{\tau_2} S(T) dT - (\tau_2 - \tau_1) F(t, \tau_1, \tau_2) \quad (9)$$

with $F(t, \tau_1, \tau_2)$ being the forward price denoted in a currency per megawatt hour.^b The question is what $F(t, \tau_1, \tau_2)$ should be.

Electricity is a nonstorable commodity by physical nature. Once purchased, we must use it, and any buy-and-hold strategy is not feasible. One cannot talk of any storage costs for electricity and nor of any convenience yield. This means that there is no natural connection as in equation (6). However, based on equation (2) we can *define* the electricity forward price as

$$F(t, \tau_1, \tau_2) = \mathbb{E}_Q \left[\frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} S(T) dT \mid \mathcal{F}_t \right] \quad (10)$$

Interchanging expectation and dT -integration, we get

$$F(t, \tau_1, \tau_2) = \frac{1}{\tau_2 - \tau_1} \int_{\tau_1}^{\tau_2} f(t, T) dT \quad (11)$$

indicating that an electricity forward corresponds to the average of forward contracts delivering at each time point t in the settlement period.

In order to make equation (10) or (11) operational, one must specify a spot price dynamics and a pricing

measure Q . Since $S(t)$ is nonstorable, it cannot be used for any hedging purposes, and in this respect functions only like an index for the forward price. We can therefore choose any stochastic dynamics we like for the spot, even going beyond the semimartingale framework as is usually required in mathematical finance. The pricing measure Q is *a priori* any equivalent probability measure. In practice, however, we want to choose a Q such that we know the dynamics of S under this measure to be able to calculate the conditional expectation. A valid choice is $Q = P$; however, this would imply a zero-risk premium, which is hardly a realistic situation.

Empirical investigations in electricity markets show that one may have a change of sign in the risk premium. Thus, it may not be the case that the market is either in normal backwardation or in contango. In fact, one has observed a positive premium in the short end of the forward market, whereas contracts with maturity in the longer end have a negative premium. The latter is consistent with producers creating a hedging pressure since they want to lock in their production. However, the positive premium in the short end has been explained by the consumers' wish to hedge the spike risk inherent in spot prices. Thus, in order to hedge the risk of paying a high spot price, the consumers are willing to pay a premium. See [7] for more on this. Bessembinder and Lemmon [5] develop an equilibrium model for the electricity market with consumers and producers. On the basis of a discrete-time equilibrium model for the spot prices, the authors show that the risk premium in the forward market depends on both price variation and right skewedness of prices. Longstaff and Wang [8] found empirical evidence for these propositions in the Pennsylvania, New Jersey, and Maryland (PJM) market.

We observe that, by definition, $F(t, \tau_1, \tau_2)$ is a Q -martingale. This is in accordance with the no-arbitrage theory that states that all forward contracts must be martingales under the pricing measure. Observe that when t approaches start of delivery τ_1 , we do not, in general, have that $F(t, \tau_1, \tau_2)$ converges to the spot price. This is rather natural since the forward contract depends on the spot over a period of delivery, and not only at a fixed time of delivery. On the other hand, this is a distinct feature for electricity markets, which is not observed in most other commodity markets. An immediate effect of this is that one does not expect the same degree

of sensitivity to spot price variations in electricity forwards as for other commodities.

The range of spot models that allow for explicit electricity forward prices is rather limited. Much of the literature on commodities focuses on mean-reverting spot price models, and the exponential Ornstein–Uhlenbeck process (see **Ornstein–Uhlenbeck Processes**)

$$S(t) = S(0) \exp(X(t)) \quad (12)$$

where

$$dX(t) = \alpha(\theta - X(t)) dt + \sigma dB(t) \quad (13)$$

is the classical example; see [9]. This model has been generalized to allow for jumps modeling spikes (see [6] for the threshold model by Geman and Roncoroni). However, even though one can derive explicit expressions for the forward prices $f(t, T)$, it seems to be very hard to obtain the same for electricity forwards (see [4] for examples). On the other hand, arithmetic models as suggested in [2] seem to be appropriate for this.

The problem of choosing the right pricing measure Q can be approached from several different angles. The most common is by means of introducing a parameterized class of pricing measures Q , that is, pricing measures that depend on one or more parameters. These parameters will appear in the forward prices, and thus one may estimate them by calibration to observed forward prices. The relation between P and the estimated pricing measure $\hat{Q}$ would give us a statistical estimate on the risk premium in the market.

Benth *et al.* [1] propose to model the formation of forward prices from spot by using the certainty equivalence principle. Two representative agents, a consumer and a producer, decide whether to trade in the spot or forward market, based on where prices are more preferable. The certainty equivalence principle leads to forward pricing bounds within which the settled forward prices must lie. Introducing the market power as an explanatory variable, forward prices can be expressed in terms of the spot. Moreover, the risk premium is stochastic, and a positive premium in the short end of the forward curve is explicitly linked to the spikes in the spot, a feature which is in line with [7]. One may derive the risk neutral pricing measure for option pricing purposes. Empirical

support for this approach has been found in an analysis of spot and forward price data observed at the German electricity exchange EEX.

An alternative approach promoted in [3] is to use the *Heath–Jarrow–Morton* (HJM, see **Heath–Jarrow–Morton Approach**) idea from *interest-rate modeling* to electricity forwards. This entails modeling the price dynamics $t \mapsto F(t, \tau_1, \tau_2)$ for arbitrary delivery periods $[\tau_1, \tau_2]$. In fact, the application of HJM in electricity is not straightforward. At Nord Pool, say, the market trades in contracts with overlapping delivery periods, leading to certain price relations that must be met to avoid arbitrage possibilities. Benth and Koekebakker [3] show that this leads to rather strong restrictions on the feasible models; for instance, geometric Brownian motion with time-dependent volatility is ruled out. Furthermore, the connection to the underlying spot price is lost since there is no convergence of electricity forwards to the spot when time to delivery goes to zero. We refer to [3] and the monograph [4] for more discussion on these issues and empirical studies of data from Nord Pool where a market model (*LIBOR model*, see **LIBOR Market Model**) approach is taken.

End Notes

^aWe assume constant interest rate r and make no distinction between a forward and a futures contract.

^bAt Nord Pool, the settlement is with respect to the hourly spot price over the delivery period. This would entail a summation over the spot prices at the discrete hours. However, we choose to work with the slightly more notational convenient integration.

References

- [1] Benth, F.E., Carlea, A. & Kiesel, R. (2008). Pricing forward contracts in power markets by the certainty equivalence principle: explaining the sign of the market risk premium. *Journal of Banking Finance* **32**(10), 2006–2021.
- [2] Benth, F.E., Kallsen, J. & Meyer-Brandis, T. (2007). A non-Gaussian Ornstein–Uhlenbeck process for electricity spot price modeling and derivatives pricing, *Applied Mathematical Finance* **14**(2), 153–169.
- [3] Benth, F.E. & Koekebakker, S. (2008). Stochastic modeling of financial electricity contracts, *Energy Economics* **30**(3), 1116–1157.
- [4] Benth, F.E., Šaltytė Benth, J. & Koekebakker, S. (2008). *Stochastic Modelling of Electricity and Related Markets*, World Scientific.

- [5] Bessembinder, H. & Lemmon, M. (2002). Equilibrium pricing and optimal hedging in electricity forward markets, *Journal of Finance* **57**, 1347–1382.
- [6] Geman, H. & Roncoroni, A. (2006). Understanding the fine structure of electricity prices, *Journal of Business* **79**(3), 1225–1261.
- [7] Geman, H. & Vasicek, O. (2001). Forwards and futures on non-storable commodities: the case of electricity, *Risk* **August**, 12–27.
- [8] Longstaff, F.A. & Wang, A.W. (2004). Electricity forward prices: a high-frequency empirical analysis, *Journal of Finance* **59**(4), 1877–1900.
- [9] Lucia, J. & Schwartz, E.S. (2002). Electricity prices and power derivatives: evidence from the Nordic Power Exchange, *Review of Derivatives Research* **5**(1), 5–50.

Related Articles

Commodity Forward Curve Modeling; Commodity Price Models; Forwards and Futures.

FRED E. BENTH

Commodity Risk

Before the emergence of B2B exchanges and the contracting innovations of interest here, the focus in procurement was on bilateral negotiation, and the literature provides analysis of many aspects of supply contracting in this setting (see [3] for a review). Bilateral contracting and traditional supplier management practices will remain essential elements of procurement management in the future. However, the growth of commodity exchanges, and derivative instruments defined on these, has introduced a fine-tuning mechanism that improves operational performance and simultaneously helps value longer term contractual, capacity, and technology decisions. The centerpiece of this new perspective is the integration of traditional forms of contracting with shorter term market-driven physical and financial transactions. This integration is the focus of this article.

An example will help clarify the scope and importance of this subject. Consider the beverage industry. Aluminum is an important element of the cost structure. For major buyers like Anheuser Busch, a restricted set of sellers is used to source aluminum, even though the aluminum spot market price is a key benchmark for sourcing and hedging, and is determined by the actions of scores of global players. Here, sourcing arrangements with main sellers are typically set according to the spot price plus processing costs, and contracts are marked to market on a daily basis. Second, there may be value-added services undertaken by these contractors to take aluminum ingots and prepare them in a more suitable fashion for can production, and again here this would be done only with specifications for these services worked out with a few sellers. However, the typical setup in metals is for a restricted set of contract sellers, with spot purchases and hedging used to benchmark prices and to “top up” contract purchases and hedge overall supply cash flows. The same general market and contracting structure operates in many other commodity markets, including energy and agricultural markets [5].

The first question to address is the consequence of competition among multiple sellers with heterogeneous costs/technologies. This question can be addressed in two ways: one is to assume a closed spot market in which all participants in the contract market also participate in the spot market, and *vice*

versa. Alternatively, one can assume an open structure in which the spot market price is determined by a larger group of competitive sellers than the smaller, restricted group operating in the contract market for a particular buyer. In this article, we follow the latter assumption; for a discussion of closed spot markets and strategic behavior, see [10].

A Primer on Previous Literature

The literature on contracting in economics has been driven by the transactions cost framework developed by Williamson [11], and subsequently formalized in the principal–agent literature [8].

A key question addressed in the economics literature has been the efficiency of various contracting structures. A well-known result in the economics contracting literature is due to Allaz and Vila [1], which examines the efficiency of pure forward contracts in an oligopolistic setting but with a deterministic spot market (which is fixed and common knowledge). Assuming homogeneous sellers and instantaneous scalability (with no capacity limitations), they show that forward markets can yield inefficient outcomes because of strategic use of these markets by sellers with market power. Wu and Kleindorfer [12] show that the Allaz–Vila-forward-market inefficiency disappears when options contracts are introduced into the mix of traded instruments, underscoring the importance of derivative instruments for supply management.

The financial economics literature provides basic pricing results for derivative instruments traded on standard contracts in liquid markets. In the context of supply contracting, an additional distinction is required in the types of options involved, between purely financial options and those connected to physical delivery of a particular good at a particular time and place. The early discussion and literature was focused on physical delivery options to fulfill a buyer’s sourcing needs. These options would entail delivery at a particular time and place (e.g., FOB Pittsburgh). In these markets, options backed by physical delivery are central to the buyer’s problem of arranging for sufficient supply to meet the buyer’s demand. However, once a functioning spot market exists, this market can be used to define financial options on the basis of the spot price. Major buyers then find it in their interest not only to arrange optimal sourcing from the contract and spot markets for

physical delivery of goods but also to use the financial options defined in the market for price discovery and as risk hedge instruments. As Birge [2] notes, if either revenue or costs associated with a given product line are correlated with some well-defined price (or cost) index arising from a competitive market, then this price index can be used to define an appropriate derivative or options instrument that provides not only risk hedging benefits but also valuation information for the underlying operations decisions whose outputs are correlated with the index.

Standard Commodity Hedging Framework

The standard approach to supply management commodity risk [5, 12] is based on the following three-period timeline for trading in the B2B exchange, either using contracts or using the spot market (Figure 1).

- Period 0: At period 0, capacity and technology choices are made by buyers and sellers. As shown in [5], these choices will be different when rational managers know that they can fine-tune demand and supply through the spot market than when such a possibility does not exist.
- Period 1: At period 1, with updated information on the distribution of spot prices, sellers and buyers contract with one another, using options and forwards, for delivery of some good (either storable or nonstorable) at period 2.
- Period 2: Finally, at period 2, after possibly updating additional information, options are called, forwards are executed, deliveries are made, and additional sales and purchases are made in the short-term spot (or cash) market.

Between period 1 and period 2, there may be additional trading of options and additional, possibly continuous, updating of information on spot prices. The analysis here will assume a discrete-time framework with no secondary trading. Thus, we are only concerned with the indicated decision instants at periods 0, 1, and 2.

The objective of the buyer is to maximize expected profit by choosing among the available forward and spot contracts. The objective of sellers is to maximize expected profit, jointly obtained from sales in both the contract market and the spot market and subject to the seller's capacity constraint. Either buyer or sellers may have additional risk-based constraints on their transactions, such as those derived from a Value-at-Risk (VaR) framework, as explained further subsequently.

A key factor influencing the incentives of sellers and the buyer to sign contracts is the existence of imperfect market access on the day, capturing possible access inefficiencies of the spot market, which includes cost and quality differences between contract markets and the spot market. In addition, in supply management for commodities, different grades and specifications for commodities often require prior contracting and procurement relations. These alternative situations give rise to various forms of commodity risk management, as shown in Table 1.

Access conditions are modeled in [12] *via* a function $m(P_s)$, which is defined as the probability that the seller can find a last-minute buyer on the spot market when the realized spot price is P_s . This market access probability can also be thought of as the proportion of the seller's capacity that can be expected to be sold at the last minute. This function may be thought of as a measure of the liquidity of the market and the degree to which the commodity in question is standardized. When quality and yield issues are present, then $m(P_s)$ may also be

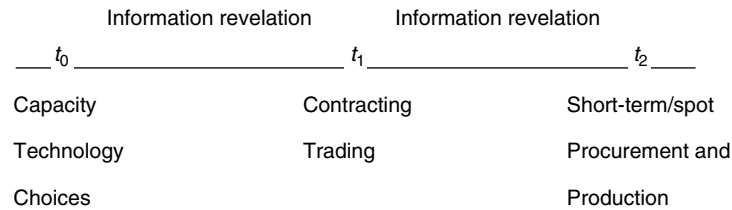


Figure 1 Market interactions in the standard commodity hedging framework

Table 1 Alternative contexts for commodity risk management of supply

Description of context	Instruments used in optimal portfolio	Examples
Cost and access differences are small and only standard commodities are sourced	Bilateral contracting and financial hedge instruments are defined on a common market and optimized jointly	energy commodity metals
Cost and access differences are large and only standard commodities are sourced	Bilateral contracting is used for most physical procurement, with spot market used for topping up supply, and for financial hedge instruments	Logistics services (standard air and maritime cargo) Fed cattle (beef), hogs, and lamb markets
Nonstandard commodities are sourced, but their prices are highly correlated with those of standard commodities	Bilateral contracting is used for all physical procurement, with financial hedge instruments, defined on correlated standard products, used as an overlay	Plastic resins and commodity chemicals

thought of as the quality risk associated with spot purchases.

The standard problem of commodity sourcing and hedging for a buyer can be stated as follows (this is the problem facing the buyer at period 1 in Figure 1):

$$\text{Maximize } E \left\{ \Pi(Q, \tilde{D}, \tilde{P}_s) \right\} \quad (1)$$

s.t.

$G(Q, F_D, F_{P_s}) \geq 0$ physical delivery constraints

$H(Q, F_D, F_{P_s}) \geq 0$ VaR constraints

$Q \in X$ other constraints on instruments

where the maximization in equation (1) is over the vector of available financial and physical instruments Q at the time of contracting, where “demand” uncertainties are represented by $\tilde{D}$, and where spot price uncertainty is represented by the random variable $\tilde{P}_s$, with respective cumulative density functions of F_D and F_{P_s} . At period 2, of course, once D and P_s are observed, instruments are executed to fulfill the physical delivery constraints and to optimize profits on the day by executing all options that are “in the money”, or needed for physical fulfillment. This problem “on the day” can sometimes be interesting, but it in theory is straightforward and we do not consider its solution here. Various forms of the problem in equation (1) have been developed for various types of markets, and the details for these differ considerably across these markets. To illustrate, let us consider the structure of this problem for the case of a distribution

company or supplier in the case of electric power operating in the presence of a liquid spot market.

Electric Power: An Illustrative Example of Commodity Hedging

We imagine an integrated utility, called the *Company*, which may own or lease generation, and which has a trading division that can sign contracts for power purchase agreements (PPAs), as well as puts, calls, and forwards, based on an underlying wholesale spot market. We abstract here from transmission constraints or markets.^a For simplicity, we also imagine that the Company provides energy to retail customers at prices that are regulated. We will take the simplest possible approach to the regulated sector, assuming a fixed, exogenously determined regulated price per Kilowatt hour, independent of time. More complicated regulatory scenarios are easily incorporated into the framework developed. This feature of customer demand at regulated prices, together with the weather-driven level of spot prices and the nonstorability of electric power, makes electricity supply a risky business.

We are interested in formulating the Company’s optimal portfolio problem for procuring and hedging its purchases of energy. The portfolio will be characterized by different levels of time-indexed instruments (puts, calls, forwards, etc.) that might be called upon either to fulfill retail demand or simply as part of profit-oriented trading/hedging activities by the Company’s trading division. We refer

to all potential assets for the portfolio, including owned/leased generation and PPAs, as *instruments*.

Following [6] and [7], we think of each instrument as having a capacity (measured in megawatts) that can be called or sold in a specific period (consisting of specified hours during a given week or month, typically the 5×16 h of “peak” or the 7×8 h of “off-peak”). Each instrument entails a reservation price, possibly zero, per megawatt to reserve, and an execution price, per megawatt hour, if used. In this framework, instruments such as own generation and certain PPAs that have been precommitted have a fixed execution price (e.g., the marginal running cost of own generation), but may be thought of as available at a reservation price of zero. Purchased forwards, which are prepaid, fixed obligations to deliver power, may be viewed as call options having a zero execution price that therefore will be executed by the Company on the day. Forwards sold by the Company have the same characteristic, that is, they may be viewed as options contracts with a zero execution price (that therefore will be executed on the day).

To set up the model, we will assume that the planning period for the instruments in question is a month, with hours in the month being denoted by the set $T = \{1, \dots, 720\}$. We will consider additional subsets of T below, such that certain instruments may be valid only for certain subsets of time (e.g., peak hours). We need the following additional notation:

Q_i = the amount (in megawatts) of instrument i that is purchased/sold

r_i = reservation price per megawatt hour for call asset i^b

c_i = execution price per megawatt hour for call asset i

s_i = reservation price per megawatt hour for put asset i

p_i = execution price per megawatt hour for put asset i
 $T_j \subseteq T$ = index of those hours of type “ j ”, where $j = 1, \dots, J$ (types of time periods might be, for example, peak periods, off-peak periods, week-ends, etc.)

I_{cj} = the set of indices for all call instruments (used in the following summation) that can be executed during hours of type j

I_{pj} = the set of indices for all put instruments (used in the following summation) that can be executed during hours of type j

n = the number of all instruments (so $n = \sum_{j=1}^J |I_{cj} \cup I_{pj}|$)

m_i = lower bound or minimum amount allowed for instrument Q_i

M_i = upper bound or maximum amount allowed for instrument Q_i

The random variable of “monthly cash flows” depends on the level of each instrument Q_i in the portfolio, where some instruments, within each month, may apply to different periods of time (e.g., just the peak periods or just the off-peak periods, or to some other selection of periods). We can write cash flows (for some period, say month t) in the following general form:

$$\begin{aligned} \Pi = & \sum_{\tau \in T} (P_{c\tau} - P_{s\tau}) D_{c\tau} \\ & + \sum_{j=1}^J \sum_{\tau \in T_j} \left[\sum_{i \in I_{cj}} [(P_{s\tau} - c_{i\tau})^+ - r_i] Q_i \right. \\ & \left. + \sum_{i \in I_{pj}} [(p_{i\tau} - P_{s\tau})^+ - s_i] Q_i \right] \end{aligned} \quad (2)$$

where $\tau \in T$ are the subperiods within the month in question, each assumed to be of length 1 h, and where we have indexed spot price P_s and the execution prices c_i and p_i for the options by time periods, in case these execution prices are time sensitive (e.g., different prices for peak or off-peak periods for the same instrument). Note that once an instrument is purchased, it is available for all hours covered by the instrument (e.g., a peak-period call option of Q_i MW is available for any peak hour in the month, subject to requirements—usually 24 h in advance—to execute the option).

Using equation (2), it is easily seen how this problem fits the standard form of equation (1). The profit function is clear (though additional payments to capital providers would be accounted for as in equation (2) in the risk constraint). As for the physical fulfillment constraints, these are also straightforward as long as the spot market is completely reliable as a source of power on the day (and very high reliability is typical of such markets). The VaR constraint in

equation (1) is represented as

$$\Pr \left\{ \Pi(Q, \tilde{D}, \tilde{P}_s) - F \geq -VaR \right\} \geq \gamma \quad (3)$$

where $\Pi(Q, \tilde{D}, \tilde{P}_s)$ are the cash flows in equation (2) resulting from the vector of contracts Q , where F represents fixed capital payment obligations, and VaR is the maximal Value-at-Risk allowed for the period in question, with confidence level γ (see **Value-at-Risk** and also [4, 9] for a discussion of VaR). In the case where there are sufficient subperiods in T to make the normality assumption reasonable for $\Pi(Q, \tilde{D}, \tilde{P}_s)$, this standard VaR constraint translates into a simple function of the mean and standard deviation of $\Pi(Q, \tilde{D}, \tilde{P}_s)$. Since the profit function is linear in Q_i for every realization of the random variables involved, finding the efficient frontier (in $E\{\Pi\}$, VaR space) is then easily solved in the standard fashion *via* quadratic programming. Multiperiod VaR constraints are also discussed in [7]. For more complex problems that arise in practice, including sophisticated models of the evolution of the spot price and various exotic options, Monte Carlo simulation (see **Monte Carlo Simulation**) with an appropriate optimization engine can be used.

Implementation and Research Challenges

It is unfamiliar territory for most organizations to integrate finance and supply management, and this is precisely what is required in order to have the full benefit of the options approach described here. Companies wishing to do so must radically expand the traditional focus of procurement on cost, quality, and dependability to include tracking of spot market conditions, valuing options in operational and hedging terms, and linking these activities to an appropriate risk management structure. Companies that have done this well have recognized the need to develop capabilities in trading, data management, and financial reporting and management. These include new skills in pricing and valuation of contracts, new approaches to managing the portfolio of sourcing options for key manufacturing inputs, and a very different approach to customer and customer segment valuation and management.

The open questions associated with B2B exchanges and contracting can be described under two general headings. First are the model-based

developments needed to capture the essence of the supply–demand coordination problem and the necessary options-based instruments to achieve efficiency in particular market contexts. Second is the continuing development of theory to understand the necessary guidelines for the structure and governance of sustainable business models for the exchanges supporting these instruments. Open research questions in these areas are noted in all of the references listed, especially in [5].

End Notes

^a.To the extent that these are based on principles of locational marginal prices, transmission constraints and options could also be included as part of the portfolio optimization described below.

^b.Thus, to reserve Q_i MW of callable capacity during a specific period of length L , the price paid is $r_i Q_i L$, where the maximum reserved/callable capacity during any hour of the period is Q_i MW. The reader can think of the reservation price in dollars per megawatt hour as the allocated cost to each hour of the period in question, though instruments will be typically traded for groups of hours in a month, for example, 100 MW of capacity callable for any peak hour during the month.

References

- [1] Allaz, B. & Vila, J.-L. (1993). Cournot competition, forward markets and efficiency, *Journal of Economic Theory* **59**, 1–16.
- [2] Birge, J.R. (2000). Option methods for incorporating risk into linear capacity planning models, *Manufacturing and Service Operations Management* **2**(1), 19–31.
- [3] Cachon, G.P. (2003). Supply chain coordination with contracts, in *Handbooks in Operations Research and Management Science: Supply Chain Management*, S. Graves & T. deKok, eds, North-Holland, Amsterdam, pp. 229–340.
- [4] Crouhy, M., Mark, R. & Galai, D. (2000). *Risk Management*, McGraw-Hill Trade, New York.
- [5] Geman, H. (2005). *Commodities and Commodity Derivatives*, Wiley Finance, New York.
- [6] Kaminski, V. (2004). *Managing Energy Price Risk*, Risk Books, London.
- [7] Kleindorfer, P.R. & Lide, L. (2005). Multi-period, VaR-constrained portfolio optimization in electric power, *The Energy Journal* **26**(1), 1–26.
- [8] Laffont, J.-J. & Tirole, J. (1998). *A Theory of Incentives in Procurement and Regulation*, The MIT Press, Cambridge, MA.
- [9] Marshall, C. & Siegel, M. (1997). Value at risk: implementing a risk measurement standard, *The Journal of Derivatives* **4**(Spring), 91–110.

- [10] Mendelson, H. & Tunca, T. (2007). Strategic spot trading in supply chains, *Management Science* **53**, 742–759.
- [11] Williamson, O.E. (1985). *The Economic Institutions of Capitalism*, The Free Press, New York.
- [12] Wu, D.J. & Kleindorfer, P.R. (2005). Competitive options, supply contracting and electronic markets, *Management Science* **51**(3), 452–466.

Related Articles

Electricity Forward Contracts; Risk Exposures; Swing Options; Value-at-Risk.

PAUL R. KLEINDORFER

Commodity Trading

Considerable concern has grown around speculative opportunities in commodity markets, which attracted hundreds of billions of dollars in recent years. In February 2008, a Reuter news item announced that “London coffee, cocoa and sugar futures rallied (...) to multi-year peaks as investment funds and speculators shifted money into soft commodities in search of stellar returns.” In July 2008, Brent oil was quoted at its historical peak of 174.27 USD per barrel, before losing about 70% of that quote four months later (December 2008). Equally significant changes occurred in the prices of cotton, coffee, wheat, soybeans, corn as well as aluminum and copper.

Since unit production costs of agricultural and nonagricultural commodities did not vary so extensively in the same period—nor did consumption—we cannot just attribute the recent marked price changes to their fundamentals. Commodity index funds, hedge funds, and speculators are, therefore, the center of huge fluctuations, changing vastly the drivers of commodity markets.

Speculative opportunities are an obvious reason for this wide interest.

In oil futures markets, a significant role of speculation has been identified to explain the large daily upward and downward shifts observed in prices [3]. Thus, even these extremely liquid instruments suffer from liquidity shocks that induce periods of increased volatility and significant returns predictability [12]. These authors conclude that the oil futures market is not driven only by fundamentals.

Empirical papers have reported evidence of predictability in commodity markets (see among others [14, 17, 20]). Moreover, theoretical contributions [5, 6, 21] attribute a key role to inventories to limit extreme price fluctuations, recognizing at the same time that inventories introduce predictability on (storable) commodity prices.

Finally, the attraction to speculation on commodity markets can be self-induced.

“Herd behavior” on the side of speculators [7] or “excess consensus” [22] can exacerbate price movements in commodity futures markets and large storage owners can profit from large price deviations. Spurred by the growing attention of specialized press and by the academic evidence showing that commodity

futures portfolios can be managed similarly to equities portfolios [16], the investment industry started offering products, allowing individuals and institutions to “invest” in commodities through several derivatives and structured notes. The performance of such products can be compared to that of popular commodity indices (e.g., Sachs commodity index) and have allowed some authors, as Domanski and Heath [9], to introduce the term *financialization* of commodity markets.

In commodity markets, predictability is essentially equivalent to mean reversion of prices as several theoretical and empirical contributions have shown (e.g., [1, 4, 10, 18] for securities [24], for commodities, and [8, 19] for electricity).

In their theoretical analysis on the role of inventories, Deaton and Laroque [5, 6] observe that the price boundaries introduced by inventory owners turn out to generate mean reversion in the price process.

From an empirical point of view, Elam [11] observes that the price of December cotton futures (at planting time) helps forecasting the cotton price at harvest. In particular, when futures price is higher than its long-term average, spot price tends to be lower and *vice versa*. This is exactly a mean-reverting behavior.

In their empirical analysis of several commodity markets [2], the authors adapt a simple trading strategy to exploit significant mean reversion observed in the corresponding price processes.

Till [25] takes into consideration two historically profitable trading strategies based on the mean-reverting properties of commodity prices.

Predictability and the relevance of mean reversion on commodity markets have also been contrasted, as in [23]. Nevertheless, the authors cannot deny the significance of mean reversion over periods longer than six months.

This evidence justifies the interest of this work. We devise optimal boundaries for the entry and exit timing of a trading strategy in a mean-reverting market.

Since our strategy is based on the peculiar properties of a mean-reverting stationary price process, it is important to observe that as long as such a strategy remains profitable, the presence of a mean-reverting stationary behavior is confirmed.

The Trading Strategy

Consider a mean-reverting process of an asset price x on which a long and a short position can be taken (such as a spot or a forward contract). The process can be described by the following stochastic differential equation (SDE):

$$dx = -\lambda(x - \Phi)dt + \theta dW \quad (1)$$

where λ is the continuous speed of reversion, θ the diffusion coefficient, and Φ is the long run average. If we consider discrete time intervals of length Δt , the corresponding discrete autoregressive process is

$$x_{t+1} = x_t + k(\Phi - x_t) + \sigma \varepsilon_{t+1} \quad (2)$$

where $k = 1 - e^{-\lambda \Delta t}$, $\sigma^2 = \frac{1 - e^{-2\lambda \Delta t}}{2\lambda} \theta^2$ and $\{\varepsilon_n\}$ are independent and identically distributed $N(0, 1)$. As soon as the process x_t diverges from Φ , the drift adjusts instantaneously to revert back to Φ . The reversion speed depends on k ($0 \leq k \leq 1$). The case $k = 0$ corresponds to a random walk.

To keep things simple, we assume that the long run average is a constant. Still, we are aware that for long-term trading strategies, it could be worth considering the inflationary or other trend elements of this average. Our one-factor process limits the shape of the forward curve more severely than it actually happens. Further, the long run average could be stochastic itself. This can be a key factor that makes the process nonstationary [15]. On the other hand, in the empirical application that follows, we will see that a constant long run price average is statistically consistent at the usual levels for several markets. Besides, the assumption of a constant Φ can be reasonable for finite (especially short) horizon strategies and has the advantage of improving the mathematical tractability.

We consider a strategy that is very close to a simple buy-and-hold. For ease of illustration, we account only for profits and losses, ignore the margin requirements, and do not consider transaction costs. We assume a zero-risk-free rate.

In a trading horizon $[0, T]$ choose $a > 0$ and construct two barriers around Φ : $\Phi + a$ and $\Phi - a$, respectively, the sell and buy barriers. Let $t^* \in [0, T]$ be the first time when $x_t \geq \Phi + a$ or $x_t \leq \Phi - a$, that is, when the price first crosses one of the two boundaries. A short or long position is then started

depending on which one of the two conditions is verified first.

Suppose that x_t goes below $\Phi - a$ in t^* , so one can take a long position (buy the asset on the spot market or a future/forward contract) for the nominal value of $\Phi - a$. The position can be closed in two ways. In the first case (*full closing*), the position is closed if $x_t \geq \Phi + a$ for some $t < T$; the net profit is $2a$. There is, however, the possibility that such a condition is not met before T . In this case, the position is closed at price x_T , with a random profit/loss $x_T - (\Phi - a)$. Analogously, the strategy is started with an opposite sign (short position) if $x_t \geq \Phi + a$ in t^* . If $t^* \notin [0, T]$ (i.e., x_t remains within the two boundaries over all the trading period), the strategy is not started and the net profit is 0 (Table 1).

This trading strategy can be implemented on the spot (respectively, futures) market, assuming that x_t is the spot (respectively, futures) price. In the case of commodity spot markets, some additional technicalities are required, such as borrowing/lending the commodity in a self-financing position. Such aspects do not alter our results. Further, by assuming a zero interest rate and equally balancing the frequencies of long and short positions, we do not need to account for gains and losses generated by borrowing/lending positions.^a

The expected profit can be optimized with respect to a under two different strategies:

1. no bounds on the full closing probabilities (maximum return strategy, MRS):

$$\max_a E[G(a; k, \sigma, T)] \quad (3)$$

2. a lower bound of 95% on the full closing probabilities (fixed risk strategy, FRS)

$$\begin{aligned} \max_a E[G(a; k, \sigma, T)] \\ \text{s.t. } p_{11} \geq 0.95 \end{aligned} \quad (4)$$

where p_{11} is the full closing probability.

Table 1 Plot of decision bounds over the price series. Outcomes of a trading strategy (time horizon T , boundaries $\Phi + a$ and $\Phi - a$)

		x_t goes above $\Phi + a$ before T	
		Yes	No
G	Yes	$2a$	$x_T - (\Phi - a)$
	No	$(\Phi + a) - x_T$	0

FRS is a VaR-like strategy as the “risk” involved in the strategy is constrained under the 5% probability of not having a full closing. We solved equations (3) and (4) numerically, searching for the optimal a on a large grid of k and T by a Monte Carlo simulation based on 2.727×10^7 trajectories. For more details, see [13].

The Risk–Return Trade-off

MRS. Simulations show that for every pair (k, T) there is a bound maximizing the expected profit. From Figure 1(a) we see that, in general, the maximum expected profit increases with T , but the effect is more pronounced for relatively low values of k , which for practical purposes is very interesting. From

Figure 1(b) (MRS depending on k for several values of T), we note that at short maturities MRS increases with k , while for longer maturities the expected profit is not monotonic: a positive reversion speed is required to generate positive expected profits; however, if it is too strong, it keeps the process close to the long run mean, reducing the profit opportunities generated by the diffusion.

FRS. Figure 2 shows the dependency of a on T and k . As a function of k (Figure 2b), short-term and long-term strategies differ: for short terms, faster reversion coefficients k imply higher closing probabilities (and so a larger a and a larger expected profit), but when the time horizon is high, an excessively strong reversion reduces the optimal bound and the

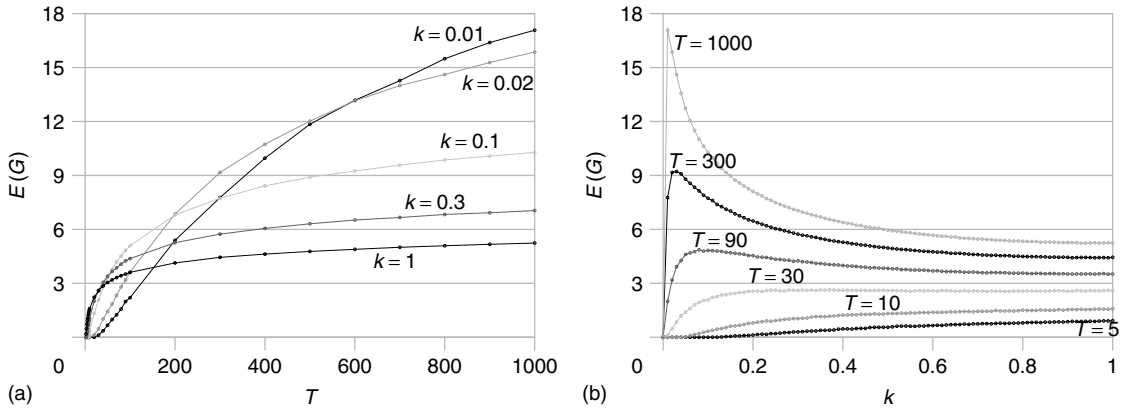


Figure 1 (a) MRS vs the maturity T for $k = 0.01, 0.02, 0.1, 0.3, 1$. (b) MRS vs the reversion k for $T = 5, 10, 30, 90, 300, 1000$

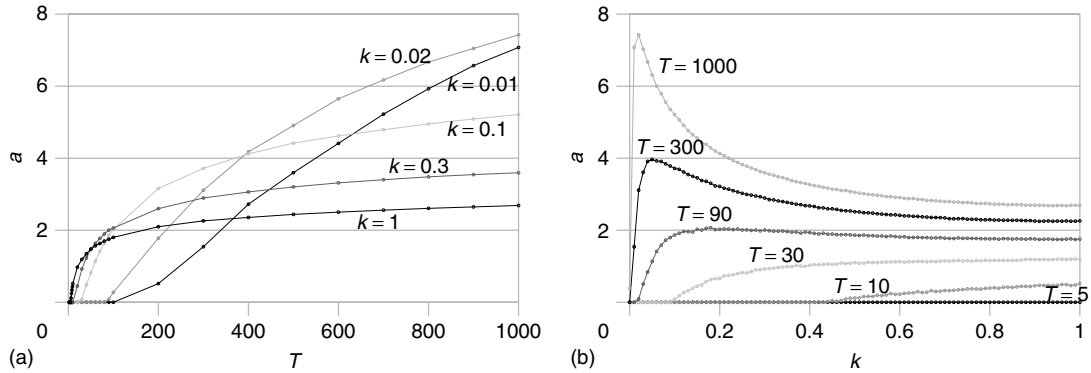


Figure 2 (a) The optimal bound a_{FRS} vs the maturity T for $k = 0.01, 0.02, 0.1, 0.3, 1$. (b) The optimal bound a_{FRS} vs the reversion k for $T = 5, 10, 30, 90, 300, 1000$

expected profit accordingly. For every large T , there is a k maximizing the full closing probability.

The Empirical Evidence

We collected the series of 6931 daily prices of nearby futures contracts of live cattle, cotton, gas oil, copper, wheat and sugar for the period January 1, 1981 to July 27, 2007. We simulated a trader taking a short/long position during a time interval j . The position is taken only once on the basis of the trading signals suggested by the optimal values of a calculated on data available up to the interval $j - 1$. The strategy starts only if a significant stationary mean reversion is observed up to $j - 1$. We evaluated the profit/losses generated by the two trading strategies, MRS and FRS, over several subperiods obtained by splitting the original series in intervals of different length (30, 50, 70, 100, 200, 300, 500 trading days).

As a refinement of those strategies, when selling or buying conditions are met (i.e., $x_t \geq \Phi + a$, $x_t \leq \Phi - a$), trades occur at price x_t instead of the boundary value. The strategies remain the same, but the order execution is similar to what is known as a *best price order* (the investor fixes a boundary price,

then the market trader executes the order at the best price available on the book satisfying the boundary).

Estimates and tests of stationarity and reversion speed are shown in Table 2. The stationarity test is based on a standard augmented Dickey–Fuller’s τ (Prob(τ) is the rejection probability under the null hypothesis that the series is stationary). The significance of the reversion speed parameter is given by a t -test on the ordinary least squares (OLS) regression 1 (Prob(k) is the rejection probability under the null hypothesis $k \neq 0$). Live cattle and cotton are significantly stationary and mean reverting. For gas oil and copper, stationarity is rejected, mainly due to the marked rise in prices that has occurred in the recent years; as a consequence, the reversion speed coefficients are not significantly different from zero. Coffee and wheat are not stationary, while their reversion coefficients are significant at 5% level.

Tables 3 and 4 report the results of MRS and FRS, respectively. To make results comparable among commodities, profits and losses are expressed as an average of all profit/losses under each maturity and are normalized as a percentage of the corresponding historical average price. Thus, figures may be taken as the average “rate of return” obtained by applying the strategy on one unit of the underlying commodity over a period of length T . A t -test assesses the statistical significance (1%, 5%).

FRS and MRS show a similar pattern, with MRS offering higher average profits as expected. Trading in live cattle gives on average sizeable MRS profits for $T = 100, 200, 300$, and 500. Cotton offers MRS positive returns (of a lesser size) for $T = 30, 100, 300$, and 500. Gas oil and copper show negligible or even significant negative returns (with an isolated exception for gas oil at $T = 500$): this is due to the nonstationarity of the series and the recent steady rise

Table 2 Stationarity test and significance reversion speed estimates and significance

Series	τ	Prob(τ)	k (%)	Prob(k)
Live cattle	-3.1927	0.0211	0.297	0.0026
Cotton	-3.8290	0.0028	0.360	0.0002
Gas oil	-1.1244	0.7086	0.084	0.3285
Copper	1.3040	0.9987	-0.058	0.3153
Coffee	-2.1375	0.2296	0.208	0.0390
Wheat	-2.3858	0.1459	0.218	0.0236

Table 3 MRS strategy—percentage return

Series	Maturity						
	$T = 30$	$T = 50$	$T = 70$	$T = 100$	$T = 200$	$T = 300$	$T = 500$
Live cattle	-0.375**	-0.333**	0.149	1.602**	5.680**	4.450**	8.793**
Cotton	0.013**	0.005*	-0.003	-0.022**	0.082	0.056**	0.211**
Gas oil	0.001	-0.004*	0.011**	-0.004	0.003	-0.037**	0.140**
Copper	-0.007**	-0.032**	-0.031**	-0.031**	-0.029**	-0.101**	0.019
Coffee	-0.013**	0.005	-0.024**	-0.064**	-0.068**	-0.100**	-0.157**
Wheat	0.903**	0.095	0.051	0.674*	-0.466	-5.864**	-7.769**

* significant at 5%, ** significant at 1%

Table 4 FRS strategy—percentage return

Series	Maturity						
	$T = 30$	$T = 50$	$T = 70$	$T = 100$	$T = 200$	$T = 300$	$T = 500$
Live cattle	−0.067	−0.127	0.029	0.797**	4.127	4.773**	7.584**
Cotton	0.007**	0.005*	0.006	−0.028**	0.061	0.033**	0.117**
Gas oil	0.001	−0.004**	0.003	−0.001	0.006	−0.032**	0.119**
Copper	−0.005**	−0.023**	−0.031**	−0.032**	−0.030**	−0.112**	−0.002
Coffee	−0.007**	0.002	−0.037**	−0.063**	−0.028*	−0.108**	−0.135**
Wheat	0.606**	0.357**	0.384*	0.830**	1.241	−3.318**	1.427

** significant at 1%, * significant at 5%

in prices. Under such conditions, prices do not reverse within the chosen boundaries. Thus, according to our strategy, the position is opened often, but is seldom “fully” closed.

Coffee for $T = 300$, 500 and Wheat for $T = 300$ (and also $T = 500$ for MRS) show FRS statistically significant negative expected returns (of a relevant size for wheat). Our trading strategies would have offered significant positive returns if applied with opposite trading signs.

The empirical outcomes confirm the theoretical expectations only for live cattle and cotton, where both stationarity and mean reversion are statistically acceptable. The other commodities, for which stationarity is weak or not significant, offer significant positive returns only on a strict range of the trading intervals or even show significant losses. The lack of stationarity of gas oil, copper, and coffee generates average losses almost over all maturities. Puzzling results as returns turning from positive (for shorter maturities) to negative (for longer ones) may be explained by markets mixing “momentum” and “reverting” regimes, where prices revert to their historical average in the short run but tend to systematically diverge on longer observation periods. This is the case of wheat from 1986 to 1995 or from 1998 to 2007.

Conclusions

We proposed a trading rule exploiting the peculiarities of a mean reversion process and investigated the related properties. The empirical results confirm the simulation analyses and show how our strategies perform nicely in markets that are stationary and mean reverting.

Such results lead to some relevant conclusions about the current hypothesis of mean reversion in commodity markets. Mean reversion can be a relevant component of the process governing many commodities. However, our results, in particular for wheat, are more consistent with markets showing regimes switching from reversion to momentum and back.

Moreover, commodity prices are not always stationary. Metals and energy prices were relatively stationary for a long time, but experienced a strong positive trend in the recent years. Many agricultural commodities mix trend with stationarity. The opportunity to fruitfully adopt our strategies is, therefore, linked to the ability of discriminating, for each commodity, stationarity from momentum market phases, which can be an easier task than forecasting size and sign of the future price changes. In such a case, switching from a momentum to a mean-reverting strategy can be a useful practice for many investors and traders.

End Notes

^aIn a risk-neutral setting, a zero interest rate implies the absence of trend for any traded asset.

References

- [1] Bessembinder, H., Coughenor, J.F., Seguin, P.J. & Smoller, M.M. (1995). Mean reversion in equilibrium asset prices: evidence from the futures term structure, *Journal of Finance* **50**(1), 361–375.
- [2] Carcano, G., Falbo, P. & Stefani, S. (2005). Speculative trading in mean reverting markets, *European Journal of Operations Research* **163**(1), 132–144.
- [3] Cifarelli, G. & Paladino, G. (2008). *Oil Price Dynamics and Speculation. A Multivariate Financial Approach*. Dipartimento di Scienze Economiche, Università

- degli Studi di Firenze, Working Paper N.15/2008, www.dse.unifi.it.
- [4] Cutler, D.M., Poterba, J.M. & Summers, L.H. (1991). Speculative dynamics, *Review of Economic Studies* **58**, 529–546.
- [5] Deaton, A.S. & Laroque, G. (1992). On the behavior of commodity prices, *Review of Economic Studies* **89**, 1–23.
- [6] Deaton, A.S. & Laroque, G. (1996). Competitive storage and commodity price dynamics, *Journal of Political Economy* **104**, 896–923.
- [7] De Long, J.B., Shleifer, A., Summers, L.H. & Waldmann, R.J. (1990). Positive feedback investment strategies and destabilizing rational speculation, *The Journal of Finance* **45**(2), 379–395.
- [8] Deng, S. (2000). *Stochastic Models of Energy Commodity Prices and Their Applications: Mean-reversion with Jumps and Spikes*. POWER, Working Paper PWP-073.
- [9] Domanski, D. & Heath, A. (2007). Financial investors and commodity markets, *Bank for International Settlements Quarterly Review* March, 53–67.
- [10] Dueker, M.J. (1997). Markov switching in GARCH processes and mean reverting stock market volatility, *Journal of Business and Economic Statistics* **15**(1), 26–34.
- [11] Elam, E. (2000). A marketing strategy for cotton producers based on mean reversion in cotton futures prices, *Beltwide Cotton Conferences Proceedings*, National Cotton Council, Memphis, TN, pp. 310–313.
- [12] Fagan, S. & Gencay, R. (2007). *Liquidity-induced Dynamics in Futures Markets*, Munich Personal RePEc Archive. MPRA Paper 6677, University Library of Munich.
- [13] Falbo, P., Felletti, D. & Stefani, S. (2007). *Optimal Trading in Mean Reverting Markets*. Dipartimento Metodi Quantitativi per le Scienze Economiche e Aziendali University Milano-Bicocca, Working Paper 125.
- [14] Finkenshtadt, B. & Kubbier, P. (1995). Forecasting nonlinear economic time series: a simple test to accompany the nearest neighbor approach, *Empirical Economics* **20**(2), 243–263.
- [15] Geman, H. (2005). Energy commodity prices: is mean reversion dead? *The Journal of Alternative Investments* Fall, 31–44.
- [16] Gorton, G.B., Hayashi, F. & Rouwenhorst, K.G. (2007). *The Fundamentals of Commodity Futures Returns*. National Bureau of Economic Research (NBER), Working Paper 13249.
- [17] Gregoriou, G.N. & Chen, Y. (2006). Evaluation of commodity trading advisors using fixed and variable and benchmark models, *Annals of Operations Research* **145**, 183–200.
- [18] Hsieh, D.A. (1995). Nonlinear dynamics in financial markets: evidence and implications, *Financial Analysts Journal* **51**(4), 55–62.
- [19] Lari-Lavassani, A., Sadeghi, A.A. & Ware, T. (2001). *Mean Reverting Models for Energy Option Pricing*, Electronic Publications of the International Energy Credit Association, www.ieca.net.
- [20] Moholwa, M. (2005). *Testing for Weak-form Efficiency in South African Futures Markets for White and Yellow Maize*, Michigan State University, Department of Agricultural Economics. Graduate Research Master's Degree Plan B Papers.
- [21] Ng, S. & Ruge-Murcia, F.J. (2000). Explaining the persistence of commodity prices, *Computational Economics* **16**, 149–71.
- [22] Roehner, B.M. & Sornette, D. (2000). Thermometers of speculative frenzy, *The European Physical Journal B - Condensed Matter and Complex Systems* **16**, 729–739.
- [23] Sam, Y.B. & Brorsen, B.W. (2000). Rollover hedging, *NCR-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management*, Chicago, <http://ageconsearch.umn.edu/handle/18938>.
- [24] Schwartz, E.S. (1997). The stochastic behavior of commodity prices: implications for valuation and hedging, *The Journal of Finance* **52**(3), 923–973.
- [25] Till, H. (2007). Trading strategies in the current commodity market environment, *Commodity Risk Magazine* May, www.energyrisk.com.

Related Articles

Commodity Forward Curve Modeling; Commodity Price Models; Commodity Risk.

PAOLO FALBO, DANIELE FELLETTI & SILVANA STEFANI

Commodity Forward Curve Modeling

Commodity Futures and Forward Prices

For several centuries, commodity forward markets have been the main place of commodity trading, as they allow producers and consumers to manage commodity price risk without going into actual delivery of a commodity. Forward contracts are traded over-the-counter between two parties and the grade of commodity, delivery point, exact amount, and exact delivery date are specified at the time of writing the contract. In this respect, they are different from *futures contracts*, which are standardized in terms of the grade, amount, delivery date, and location. Commodities are traded on exchanges such as the London Metal Exchange (LME), Chicago Board of Trade (CBOT), Chicago Mercantile Exchange (CME), InterContinental Exchange (ICE), and New York Mercantile Exchange (NYMEX). These futures markets provide the most reliable and liquid commodity futures prices. For example, most of the trading activity in oil markets takes place in futures and forward contracts, where the volume of trades is almost 10 times higher than in the spot market.

Throughout the article, we denote the commodity futures or forward price on day t for maturity date T by $F(t, T)$ and the spot price (i.e., the price for the immediate delivery) on day t by $S(t)$. The forward curve prevailing on day t is the collection of futures prices $\{F(t, T)\}_T$ for all traded maturities $T = 1, 2, \dots, N$. It has been known for some time that, in absence of credit risk, forward and futures prices are equal under nonstochastic interest rates [5]. In commodity markets, interest rate risk plays a secondary role compared to the commodity spot price risk, so we use the terms “futures price” and “forward price” interchangeably.

Forward curves are of paramount importance in commodity markets, for several reasons. They provide information about the views of market participants, anticipated price trends, and expectations about future supply and demand. Futures prices observed in liquid futures markets provide price discovery and are essential for daily marking to market existing portfolio positions as well as for risk management

activities such as Value-at-Risk calculations. Forward commodity prices influence storage, production, and other strategic decisions in related industries. Finally, futures contracts provide the way of calibrating derivatives pricing models under the risk-neutral probability measure.

The spot price of a commodity is often considered as the most important factor driving the whole forward curve [6, 13]. The influence of the spot price becomes greater as the futures contract approaches maturity, culminating in the following convergence property: on maturity date, the futures price must coincide with the spot price

$$F(T, T) = S(T) \quad (1)$$

This property follows from the absence of arbitrage and the possible occurrence of physical delivery of a commodity at the maturity of the futures contract. However, there are markets where the spot price is rather opaque or unreliable (or completely nonexistent, as it is the case, for instance, in electricity markets). In such cases, the convergence property does not necessarily hold or does not take place smoothly as the maturity approaches. These considerations, among others, may question the suitability of the spot price as the main driving factor of the forward curve.

Crude oil (and some metals) forward curves have traditionally been in one of the two states: backwardation, when the futures prices for short maturities are more expensive than those maturing later, or contango, which is the opposite situation. Figure 1 shows the NYMEX crude oil forward curve on February 21, 2008, when the market was in backwardation, and Figure 2 shows a contango forward curve observed on February 28, 2007.

Whether the market is in backwardation or in contango depends on the current price as well as inventory levels, transportation and storage costs, supply of storage, supply and demand equilibria, strategic and political issues, and many other factors. The shape of the commodity forward curve is closely related to the notion of the so-called *convenience yield*, that is, the benefit of holding a physical commodity over a futures contract. Backwardated forward curves arise when the benefit of holding a physical commodity (i.e., the convenience yield) is high and the interest rates and storage costs (representing together the “cost of carry”) are relatively low. This happens, for instance, when there is a general perception of

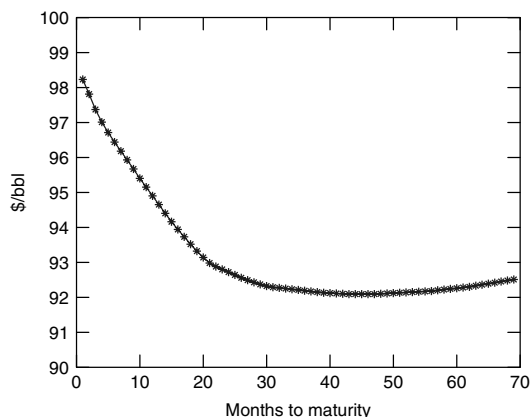


Figure 1 NYMEX crude oil forward curve, February 21, 2008

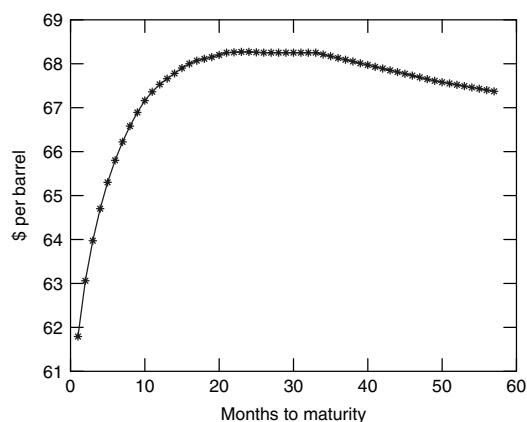


Figure 2 NYMEX crude oil forward curve, February 28, 2007

low availability of a commodity and instability of its supply. More generally, the shape of the forward curve summarizes perceptions of market participants about the current state and future developments of the global commodity market.

For seasonal commodities, such as natural gas, electricity, or agricultural commodities, futures prices are largely governed by seasonal demand (as it is the case for energy) or supply (for agricultural commodities). For those commodities, transportation or storage capabilities are limited, so the excess demand or supply shortage cannot be absorbed by transporting a commodity from a different part of the world with excess supply, or storing at times of plentiful supply and using it whenever necessary.

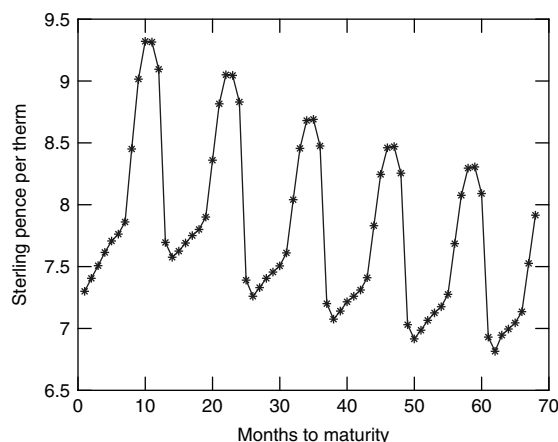


Figure 3 UK Natural Gas forward curve, March 7, 2007

This results in a high seasonal premium on futures contracts maturing during periods of high demand (such as winter, in the case of natural gas) or low supply (such as before harvest, for agricultural commodities). This is illustrated in Figure 3, where a forward curve for natural gas in the United Kingdom is shown. Futures maturing in the winter (the time of high demand for gas in the United Kingdom) are clearly at a premium compared to those maturing in summer.

The dynamic modeling of the forward curve (i.e., the description of its random movements over time) and the understanding of the main fundamental factors behind its evolution has been a subject of extensive research, both by practitioners and academics. A proper model of the dynamics of the forward curve is an essential input in most derivative pricing models, scenario simulation, and risk management applications. A successful forward curve model should be able to match current observed forward curves, generate realistic forward curves containing empirically observed features, and should ideally be able to extrapolate the forward curve beyond the longest observed maturity. Moreover, it should be easily and quickly calibrated to the market data. However, just as in case of the Treasuries yield curves, commodity forward curve modeling is a challenging task, owing to the inherent multidimensionality of the problem and a nontrivial dependence structure across maturities. In contrast to interest rates, features such as seasonality may further complicate the modeling process.

Forward Curve Models

Forward curve modeling approaches for commodities can be loosely classified into three main categories. The first one is a martingale-based approach, which models the dynamics of futures prices directly under the risk-adjusted probability measure Q , making use of the fact that futures prices are martingales under such a measure. One or possibly several risk factors driving the risk-neutral dynamics of the entire forward curve are usually assumed (resulting in the so-called *one-factor* or *multifactor models*).

Multifactor models such as those described in [4] are similar to the celebrated Heath–Jarrow–Morton (HJM) approach for modeling the yield curve [7], in that they directly model futures prices under the risk-neutral probability measure.^a So the following representation is often considered:

$$\frac{dF(t, T)}{F(t, T)} = \sum_{i=1}^n \sigma_i(t, T) dW_i(t) \quad (2)$$

where n is the number of risk factors, $W_i(t)$, $i = 1, \dots, n$ are Brownian motions (possibly correlated) representing sources of uncertainty and $\sigma_i(t, T)$ are volatilities associated with the risk factors (possibly varying deterministically in time and maturity). The statistical technique of *principal component analysis* is often used to derive the risk factors, in which case they are assumed to be uncorrelated. Such an approach focuses on the martingale property of $(F(t, T))_{t \leq T}$ under Q , which is useful for derivatives pricing. However, it is less suited for trading strategies involving spot and forward positions (the most common ones in a trading environment), as it says nothing about the actual evolution of the forward curves and hence cannot be calibrated to historical futures market data. The risk factors can be rather arbitrary and do not always have a clear meaning.^b Alternatively, more interpretable fundamental stochastic factors can be taken. For example, a popular choice is to take the short-term and long-term price components (e.g., [10, 14]), which together determine the commodity's spot price:

$$S(t) = s(t) + X(t) + L(t) \quad (3)$$

where $X(t)$ and $L(t)$ are respectively the short- and long-term price components and $s(t)$ is a possible (deterministic) seasonal component of the spot price.

Then the futures prices are obtained using the relationship

$$F(t, T) = E_Q(S(T) | \mathcal{F}_t) \quad (4)$$

where Q is the risk-adjusted probability measure and $\mathcal{F}_t$ is the information available up to date t . To evaluate the expression (4), the dynamics of the fundamental factors (such as $X(t)$ and $L(t)$) under the risk-adjusted probability measure Q should be specified. Usually, such dynamics include the market prices of risk (the spot price risk and convenience yield risk), which make these models rather hard to calibrate, as market prices of risk are not observable. Another problem with such models is that the forward curves derived from equation (4) usually do not match observed forward curves.

Another approach to the commodity forward curve modeling is a static arbitrage-based one, making use of the cash-and-carry arguments and describing the no-arbitrage relationship between the spot and futures prices. This approach is based on the no-arbitrage assumption and implies that the futures price of a commodity must be equal to the cost of acquiring the physical commodity and carrying it until the maturity of the futures. Physical commodities, unlike financial assets, cannot be stored costlessly; so for them, the cost of carry is largely determined by the accumulated storage costs (and the interest lost on the investment into a commodity). However, these arguments would imply that commodity forward curves should always be in contango, that is, futures prices for longer maturities should always be higher than those for shorter maturities and the spot price. In practice, this is rarely the case; for example, historically, the crude oil futures market has been in backwardation approximately 75% of the time. To cope with this observation, the notion of *convenience yield* was introduced by Kaldor [8] and Working [15]. It is defined as the premium received by the owner of the physical commodity and not accruing to the holder of a futures contract written on it. This concept leads to the famous cost-of-carry relationship:

$$F(t, T) = S(t)e^{[r(t)+c(t)-\tilde{y}(t)](T-t)} \quad (5)$$

where $r(t)$ is the riskless interest rate on the date t , $c(t)$ is the storage cost (per unit of time and per dollar worth of commodity), and $\tilde{y}(t)$ the convenience yield, all continuously compounded. Often, the convenience yield is defined net of storage costs: $y(t) = \tilde{y}(t) -$

$c(t)$, in which case the cost-of-carry relationship becomes

$$F(t, T) = S(t)e^{[r(t)-y(t)](T-t)} \quad (6)$$

In its early versions, the cost-of-carry relationship (6) included a constant or deterministic (but time-varying) convenience yield, such as in [3]. The next step was to define the convenience yield as a function of the spot price. More realistic versions of equation (6) (e.g., those in [6, 9]) included the convenience yield as a stochastic process. Some authors stressed a time-spread optionality embedded in the convenience yield (discussed in, e.g., [12]). To emphasize this option-like feature of the convenience yield, it should be considered as a function of maturity T (or time to maturity $T - t$), as well as time t : $y = y(t, T)$, as suggested in [2]. This representation is particularly useful in the case of seasonal commodities.

Obtaining a dynamic model for the forward prices from the cost-of-carry relationship (6) goes as follows. The spot price $S(t)$ and the convenience yield $y(t)$ are considered as the main fundamental stochastic factors driving the forward curve's evolution. A stochastic model is assumed for the joint dynamics of $S(t)$ and $y(t)$ (e.g., correlated geometric Brownian motions or mean-reverting diffusions) and then the corresponding dynamics for the futures prices is obtained by substituting the expressions for $S(t)$ and $y(t)$ into the cost-of-carry relationship (6). Sometimes, the interest rate can be added as another fundamental factor, as in [11]. The cost-of-carry relationship (6) holds under any probability measure, so for the purposes of derivatives pricing, the risk-neutral dynamics of futures prices can be obtained from the risk-neutral dynamics of the fundamental factors.

The Seasonal Forward Curve Models

Seasonal forward curves are often observed in energy and agricultural markets. One attempt to model these has been made by Sorensen [14] and Lucia and Schwartz [10] who consider a deterministic seasonal component of the spot price, which subsequently enters the expression for the forward prices. However, seasonality in forward prices obtained in this way does not match the empirically observed seasonal patterns present in a forward curve, especially the observation that the forward price's seasonality

is a function of the contract maturity T and not so much of the current date t .

Forward prices of seasonal energy commodities such as natural gas and electricity are influenced by seasonal demand due to heating or air-conditioning; agricultural commodities are influenced by seasonal supply (harvest). These constraints result in positive seasonal premia attached to futures maturing in calendar months of high demand or low supply. The seasonal cost-of-carry model by Borovkova and Geman [2] incorporates such a seasonal premium within the framework of statistical arbitrage-like relationship.

Assume that the forward curve contains liquid maturities up to one year (12 months) or an integer number of years. The first fundamental factor of the seasonal cost-of-carry model is defined as *the geometric average of the observed forward prices at date t* :

$$\bar{F}(t) = \sqrt[N]{\prod_{T=1}^N F(t, T)} \quad (7)$$

where N is the most distant maturity. The average level of the forward curve $\bar{F}(t)$ is a robust quantity and a good indicator of the overall state of the forward market (more so than the volatile spot price). Note that, if $N = 12 \times k$ ($k = 1, 2, \dots$), then $\bar{F}(t)$ is a nonseasonal quantity, unlike a possibly seasonal spot price. Furthermore, the commodity spot price can be unreliable or even unavailable, so it is often not a good candidate to be the main fundamental factor driving the forward curve.

The seasonal premia $s(M)$, ($M = 1, \dots, 12$), less associated with calendar months, are defined as *the average premia, expressed in per cent, in futures expiring in the calendar month M with respect to the day- t average forward price $\bar{F}(t)$* . It is assumed to be deterministic and $\sum_{M=1}^{12} s(M) = 0$.

The *seasonal cost-of-carry model* describes the relationship between the prevailing forward price $\bar{F}(t)$ and the futures price $F(t, T)$:

$$F(t, T) = \bar{F}(t)e^{[s(T)-\gamma(t, T-t)](T-t)} \quad (8)$$

where the quantity $\gamma(t, T - t)$, defined by the relationship above, is called the *stochastic premium* at date t (as opposed to the deterministic seasonal premium $s(T)$), associated with the time to maturity $T - t$. Denoting $\tau = T - t$, the stochastic premium becomes $\gamma(t, \tau) = \gamma(t, T - t)$, which emphasizes

that γ is associated with the time to maturity τ rather than the maturity date T .

Therefore, futures maturing in a particular calendar month can be either at a premium or a discount relative to the average level of the forward curve. This is a periodic, that is, seasonal effect, which is associated with periods of high demand or low supply. The seasonal premium is the nonstochastic quantity $s(M)$, which can be consistently estimated from the historical data. The quantity $\gamma(t, \tau)$, on the other hand, is a stochastic process in t , indexed by the time to maturity τ , and summarizes all deviations of the actual observed futures prices from the anticipated futures prices given by $\bar{F}(t)e^{s(T)}$. It is called the *stochastic cost of carry*, or *stochastic premium devoid of seasonal premium*, for time to maturity τ . Factors influencing the function γ are, for example, levels of stocks, political news and events, or unusual weather conditions.

The seasonal cost-of-carry model (8) is related to the traditional cost-of-carry models such as the one in [3], but with the traditional convenience yield replaced by the stochastic premium $\gamma(t, \tau)$. This stochastic premium can be seen as the “relative convenience yield” (with respect to its average across all maturities), net of the scaled seasonal premium:

$$y(t, T) - \frac{1}{N} \sum_{M=1}^N y(t, M) = \gamma(t, \tau) - \frac{s(T)}{\tau} \quad (9)$$

where $y(t, T)$ is the traditional convenience yield.

The convergence property of the futures price to the spot price (1) provides the following expression for the spot price within the seasonal cost-of-carry

model:

$$S(t) = \bar{F}(t)e^{s(t)} \quad (10)$$

hence, if the spot price is not available in a particular market, the above relationship can be used to define a valuable proxy for the spot price.

The stochastic premium can be represented by a mean-reverting process, where the parameters of the process (such as the mean-reversion speed, level, and volatility) may differ per time to maturity τ . The entire term structure of $\gamma(t, \tau)$ can be driven by the same Brownian motion $W(t)$; more sources of uncertainty can be easily incorporated. The average forward price $\bar{F}(t)$ can be modeled by any appropriate diffusion or jump-diffusion process, uncorrelated to the term structure of $\{\gamma(t, \tau)\}_\tau$.

The seasonal premium and other model parameters can be estimated from the set of historical futures prices by the ordinary least squares method or a Kalman filter methodology. Figure 4 shows the estimated seasonal premia for UK natural gas, electricity, and heating oil futures. Our estimates are based on historical data sets of ICE futures prices from 2001 to 2008.

As expected, futures expiring in winter are at a premium with respect to the average price level, and summer futures at a discount. December gas futures are on average at a 28% premium and December electricity futures are at 15% premium. The seasonal premium for heating oil is generally smaller, and is at most 3%, which reflects a wider availability of storage and transportation for heating oil.

For longer forward curves (such as the one depicted in Figure 3), a significant slope (defining an overall backwardation or contango state) can be

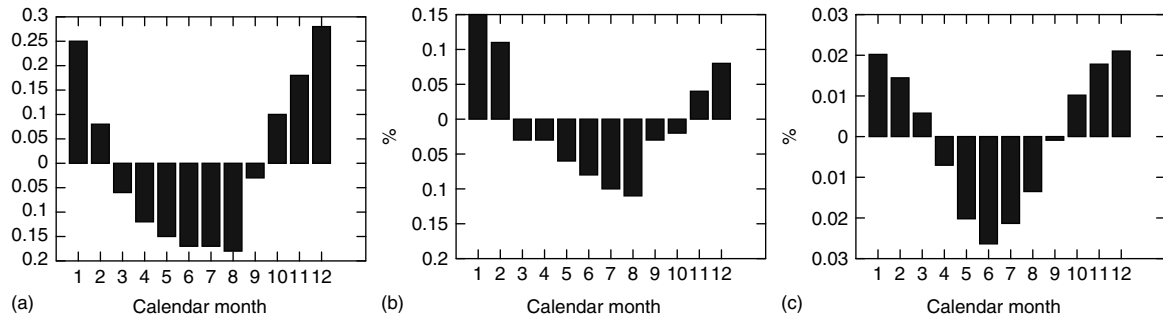


Figure 4 Natural gas, electricity, and heating oil seasonal premia. (a) Seasonal component, natural gas futures. (b) Seasonal forward premium, electricity futures. (c) Gasoil seasonal component

detected. In such cases, instead of defining the $\bar{F}(t)$ as the day- t level of the forward curve, we can approximate the overall shape of the log-forward curve on day t by a straight line fitted to all available forward log prices:

$$X_\tau(t) = X_0(t) + \alpha(t)\tau \quad (11)$$

where the parameters $\alpha(t)$ (slope) and $X_0(t)$ (intercept) are estimated on each day t by the ordinary least squares method. Note that equation (11) defines an affine function of the time to maturity. If the slope $\alpha(t)$ on day t is negative, then the forward curve is in overall backwardation, such as in Figure 3, where the seasonal premium is still clearly present. The positive $\alpha(t)$ defines the contango forward curve shape. If $\alpha(t)$ is not significantly different from zero, then we are in the situation of the initial seasonal cost-of-carry model, with the parameter $\bar{F}_0(t) = e^{X_0(t)}$ being exactly the geometric average of the forward curve $\bar{F}(t)$.

The seasonal premia $s(M)$ ($M = 1, \dots, 12$) are defined, as before, by the average premia in futures expiring in the calendar month M , but now with respect to the day- t overall shape of the forward curve, given by the exponent of equation (11). The *generalized seasonal cost-of-carry model* is given by

$$F(t, T) = \bar{F}_0(t)e^{\alpha(t)\tau + s(T) - \gamma(t, \tau)\tau}, \quad \tau = T - t \quad (12)$$

where $\gamma(t, \tau)$ is again the stochastic forward premium, defined by the relationship (12) above. This is a three-factor model. The slope $\alpha(t)$, the intercept $X_0(t)$, and the stochastic premium $\gamma(t, \tau)$ are the three fundamental factors. The factors $X_0(t)$ and $\alpha(t)$ have a well-known interpretation, closely related to the principal component analysis of the forward curve. They reflect the *level and slope* of the forward curve on date t and contain most of the forward curve's variability. The remaining stochastic variation of the forward curve, in particular its curvature, is summarized in the stochastic forward premia $\gamma(t, \tau)$. All three fundamental factors are pairwise uncorrelated.

The dynamics of the factor $X_0(t)$ can be a process mean reverting to a constant or an increasing level, or a geometric Brownian motion. Jumps can be incorporated into $X_0(t)$ as well. The slope $\alpha(t)$ and

the stochastic premium $\gamma(t, \tau)$ can be modeled as (uncorrelated) mean-reverting processes.

Specifying the factors' dynamics leads to the corresponding forward price dynamics *via* the relationship (12), which holds under any probability measure (statistical or risk neutral). Risk management applications, scenario simulations, and portfolio strategies involving spot and forward positions can be devised using the representation under the statistical probability measure. For purposes of derivatives pricing, the risk-neutral dynamics is of interest. Specifying such dynamics for the three fundamental factors implies the risk-neutral dynamics and distributions for the futures prices. With these in hand, derivatives such as futures options or calendar spread options can be easily valued by a Black–Scholes-like approach or by Monte Carlo simulations. Recall that, under the risk-neutral measure Q , futures prices are martingales, that is, their drift is identically zero. This results in a restriction on the model parameters under Q , similar to the HJM condition [7] on Treasury forward rates, or the conditions obtained by Amin *et al.* [1].

Conclusions

Building a well-functioning model for commodity forward curves is a challenging task owing to particular features of commodity markets and an inherent multidimensionality of the problem. Applications of commodity forward curve models include the generation of realistic forward curves for risk management and scenario simulations. Moreover, an appropriate model can be used for extending the observed forward curves beyond the most distant liquid maturity, a necessary task in the valuation of physical assets such as gas storage facilities or refineries. The formulation of the fundamental factors' risk-neutral dynamics provides a unifying framework for pricing derivatives, such as calendar spread options, that depend on the entire forward curve.

End Notes

^aAs a futures contract does not require an initial investment, under Q its price remains constant on average and there is no drift term representing the average growth on the investment.

^bIf these risk factors are obtained by the principal component analysis, they are usually interpreted as changes in the level, slope, and curvature of the forward curve.

References

- [1] Amin, K., Ng, V. and Pirrong, C. (2004). *Arbitrage-Free Valuation of Energy Derivatives In: Managing Energy Price Risk*, The New Challenges and Solution. Third Edition Ed. Vincent Kaminski, Risk Books.
- [2] Borovkova, S. & Geman, H. (2006). Seasonal and stochastic effects in commodity forward curves, *Review of Derivatives Research* **9**, 167–186.
- [3] Brennan, M.J. & Schwartz, E.S. (1985). Evaluating natural resource investments, *Journal of Business* **58**(2), 135–157.
- [4] Clewlow, L. & Strickland, C. (2000). *Energy Derivatives: Pricing and Risk Management*, Lacima Publications, London, UK.
- [5] Cox, J.C., Ingersoll, J.E. & Ross, S.A. (1985). A theory of the term structure of interest rates, *Econometrica* **53**, 385–407.
- [6] Gibson, R. & Schwartz, E.S. (1990). Stochastic convenience yield and the pricing of oil contingent claims, *Journal of Finance* **45**(3), 959–976.
- [7] Heath, D., Jarrow, R. & Morton, A. (1992). Bond pricing and the term structure of interest rates: a new methodology for contingent claims evaluation, *Econometrica* **60**(1), 77–105.
- [8] Kaldor, N. (1939). Speculation and economic stability, *Review of Economic Studies* **7**, 1–27.
- [9] Litzenberger, R.H. & Rabinowitz, N. (1995). Backwardation in oil futures markets: theory and empirical evidence, *Journal of Finance* **50**(5), 1517–1545.
- [10] Lucia, J. & Schwartz, E.S. (2002). Electricity prices and power derivatives: evidence from the nordic power exchange, *Review of Derivatives Research* **5**(1), 5–50.
- [11] Miltersen, K.R. & Schwartz, E.S. (1998). Pricing of options on commodity futures with stochastic term structures of convenience yield and interest rates, *Journal of Financial and Quantitative Analysis* **33**(3), 33–59.
- [12] Routledge, B.R., Seppe, D.J. & Spatt, C.S. (2000). Equilibrium forward curves for commodities, *Journal of Finance* **55**(3), 1297–1338.
- [13] Schwartz, E.S. (1997). The stochastic behaviour of commodity prices: implications for valuation and hedging, *Journal of Finance* **53**(3), 923–973.
- [14] Sorensen, C. (2002). Modeling seasonality in agricultural commodity futures, *Journal of Futures Markets* **22**(5), 393–426.
- [15] Working, H. (1948). The theory of price of storage, *Journal of Farm Economics* **30**, 1–28.

Further Reading

- Gabillon, J. (1994). Analyzing the forward curve, in *Managing Energy Price Risk*, Risk Publications.
- Geman, H. (2005). *Commodities and Commodity Derivatives*, Wiley Finance.

Related Articles

Commodity Price Models; Electricity Forward Contracts; Forwards and Futures.

SVETLANA BOROVKOVA

Commodity Price Models

Arbitrage models of commodity prices are relative pricing devices. They allow market operators to evaluate commodity-linked securities in terms of *quoted* commodity prices or indices. Spark and crack spread options, multicommodity basket options, swing contracts, and collateralized commodity obligations are just a few examples of financial securities that can be priced (and thus hedged) using commodity models [17]. Moreover, the valuation of certain real assets may benefit from the use of these models. That is the case of long-term investments in natural resources, which can be interpreted as real (vs. financial) options written on the value of the corresponding commodity price(s) [37].

Models differ in the *primitives* from which prices of more complex positions can be derived and in the way the stochastic dynamics of these building blocks is described. As a consequence, model selection mainly depends on the level of development characterizing the market under consideration. For instance, oil markets are rather mature; they provide traders with a large quantity of reliable quotes for futures as well as vanilla options; these prices may thus be considered as model primitives. On the contrary, several electricity markets quote spot prices only, which then constitute candidate primitives for valuation purposes.

We begin by deriving the most important commodity valuation formula based on arbitrage considerations; then, we consider four major classes of commodity price models, each one gathering models sharing a common set of primitives; finally, we address a number of modeling issues underlying the design of risk management systems for commodity portfolios.

Net convenience revenue is defined as the sum of all benefits (e.g., saved costs of shortage, consumption value, and other side cash flows) minus the sum of all costs (e.g., storage expenses and replacement costs for perishable goods) stemming from physically holding a storable commodity over a period of time [8]. For the purpose of pricing forward contracts, which represent the simplest class of commodity derivatives, net convenience revenue must be computed net of the risk-free rate of interest on the currency of denomination representing the purely financial cost of funding the operation. This

computation results in a quantity referred to as *cost of carry*. The underlying argument goes as follows.

Current time is denoted by t . We consider a forward contract delivering a single unit of a commodity at a future time $T > t$ for a price $x(t)$. Although this price is determined at current time t , it represents a cash flow occurring at time T . It is assumed that holding the asset over the time interval $[t, T]$ leads to an overall benefit B and bears a total cost C , both quantities expressed in units of currency available at time T . A short forward position at time t leads to a time T pay off equal to $x(t) - S(T)$. At a first glance, one may think of hedging this cash flow by borrowing the present value of delivery price $x(t)$ and then buying the underlying asset at current price $S(t)$. However, this strategy generates the required cash flow $S(T) - x(t)$ *plus* an additional amount given by the net convenience revenue resulting from the asset holding. We then have to modify our trading strategy in a way to offset the resulting net convenience revenue. This can be done by borrowing funds equal to the present value of $B - C$. Table 1 illustrates the resulting *replicating strategy*, which is commonly referred to as *cash-and-carry*. V_f denotes the forward contract value and $P_T(t)$ is the discount factor representing the time t value of 1 unit of currency of denomination available at time T .

The net cash flow $V_\pi(T)$ resulting from this trading strategy at delivery is zero. Barring arbitrage opportunities, the time t value $V_\pi(t)$ of the resulting position must vanish as well. This fact leads to a formula for the commodity *forward (fair) value* V_f :

$$\begin{aligned} V_\pi(t) &= 0 \Leftrightarrow V_f(t) \\ &= S(t) - [x(t) + B - C] \times P_T(t) \end{aligned} \quad (1)$$

The corresponding time t *forward price* $f_T(t)$ is defined as the unique delivery price $x(t)$ negotiated at time t and making the value V_f of the forward contract equal to zero at the same time:

$$f_T(t) = x(t) : V_f(t) = 0 \quad (2)$$

Solving this equation with respect to the unknown delivery price $x(t)$ leads to a formula for the arbitrage-free forward price:

$$f_T(t) = \frac{S(t)}{P_T(t)} - (B - C) \quad (3)$$

Table 1 Replicating strategy for a commodity forward contract

	Position value at t	Value at T
1 short forward	$-V_f(t)$	$-S(T) + x(t)$
Borrowing \$	$-[x(t) + B] P_T(t)$	$-x(t) - B$
Lending \$	$+C \times P_T(t)$	$+C$
1 long asset	$+S(t)$	$+S(T) + B - C$
Portfolio π	$V_\pi(t) = -V_f(t) - [x(t) + B - C] \times P_T(t) + S(t)$	$V_\pi(T) = 0$

If side benefits and costs quote in terms of continuously compounded rates of appreciation β and depreciation χ respectively, then we may write the *spot-forward parity* (3) in the continuously compounded multiplicative form:

$$f_T(t) = \frac{S(t)}{P_T(t)} e^{-(\beta - \chi)[T - t]} \quad (4)$$

The difference $c := \beta - \chi$ defines the notion of *average spot convenience yield* per time unit. If r denotes the continuously compounded risk-free interest yield prevailing on average over the forward contract lifetime, then $f_T(t) = S(t) e^{[r - (\beta - \chi)][T - t]}$ and the factor $r - (\beta - \chi)$ represents the yield quoted *average cost of carry* per time unit.

Spot-forward parities (3) and (4) result from arbitrage considerations under the assumption that:

1. a spot price is quoted;
2. forward delivery occurs at a single fixed point in time;
3. the underlying asset is storable.

In contrast to hypothesis 1, some markets (e.g., oil) do not quote prices for spot delivery. In this case, we may either identify spot with the shortest maturity forward, or simply avoid considering spot prices and focus on forward curve models, whereby forward prices are model primitives and the spot is a derived quantity. As opposed to assumption 2, a few markets (e.g., electricity and shipping freight) quote forwards for delivery over a whole time period. A possible way out consists in finding implied forward prices compatible to the existing market quotes [28]. As far as assumption 3 is concerned, some commodities (e.g., electricity) cannot be stored. This fact prevents traders from performing a cash-and-carry strategy in the way described earlier, which may lead to denying the validity of the spot-forward parity. However, asset storability is not actually required to derive this relationship. One simply need be able to enter

a commitment whereby the asset is delivered on the future date T at the standing spot price $S(T)$. This is the case, for instance, of a trader owning the right to handle a production unit with available capacity on the date of delivery.

If any process between interest rate and convenience yield is stochastic, one has to rely on the general arbitrage pricing formula. This formula states that the fair value of a financial security is the risk-neutral expected value of the sum of all discounted future cash flows going to the holder. In a diffusion model setting, risk neutrality operatively translates into a price drift $\mu_S^*(t)$ that can equivalently be written either as an *instantaneous cost of carry* $r(t) - c(t)$ or as the historical price drift $\mu_S(t)$ plus the product between price volatility $\sigma_S(t)$ and a *market price of risk* process $\lambda(t)$. This latter can be estimated using methods proposed by Benth *et al.* [4] and Kolos and Ronn [29], among others.

In the case of a forward contract issued at time t and maturing at a future time $T > t$, the fair value at time $s : t \leq s \leq T$ reads as follows:

$$V_f(s) = \mathbb{E}_s^{\mathbb{P}^*} \left(e^{-\int_s^T r(u) du} (S(T) - x(t)) \right) \quad (5)$$

where $\mathbb{P}^*$ denotes the risk-neutral probability, and $r = (r(u), t \leq u)$ is the interest rate process. Substituting this quantity into the zero-value condition (2) and solving for the delivery price leads to a *general spot-forward parity*:

$$f_T(t) = \frac{\mathbb{E}_t^{\mathbb{P}^*} \left(e^{-\int_t^T r(s) ds} S(T) \right)}{P_T(t)} \quad (6)$$

We note that the underlying convenience revenue implicitly appears via the risk-neutral drift of spot price dynamics, that is, $\mu_S^*(t) = r(t) - c(t)$. We refer to $c(t)$ as the *instantaneous spot convenience yield* associated to commodity S . A more explicit expression is derived in [25] under the assumption

of a deterministic covariance between spot price and the exponentially integrated convenience yield processes.

Gibson and Schwartz [21] propose a model where commodity prices evolve through a continuous diffusion process driven by two factors: one factor is idiosyncratic to spot price returns; the other affects price evolution through the instantaneous convenience yield process. The resulting $\mathbb{P}^*$ -dynamics read as follows:

$$dS(t) = [r - c(t)]S(t)dt + \sigma_S S(t)dW_S(t) \quad (7)$$

$$dc(t) = \kappa[\alpha^* - c(t)]dt + \sigma_c dW_c(t) \quad (8)$$

where σ_S (respectively, σ_c) represents the spot price return (respectively, instantaneous convenience yield) volatility coefficient κ denotes the mean reversion frequency of convenience yield; $\alpha^* = \alpha + \sigma_c \lambda \kappa^{-1}$ (respectively, α) indicates the risk-neutral (respectively, historical) long-term convenience yield level; and the two $\mathbb{P}^*$ -Brownian motions W_S and W_c are assumed to be correlated with a coefficient ρ . These dynamics are assumed to start at initial values $S(t_0)$ and $c(t_0)$. The model represents a prototype for the *spot-convenience yield model* class (SC models). Bjerk Sund (1991, unpublished manuscript) and Jamshidian and Fein [26] show that forward prices in this model are exponentially affine functions of the pair $(\ln S(t), c(t))$. This feature is shared by most SC models, a fact leading Björk and Landén [6] to pursue a systematic study of affine models for commodity prices in the general setting of jump-diffusions (see **Affine Models** for details on affine models in the context of interest rates). SC models have been extensively studied in a number of papers, such as [10, 16, 30, 33, 37, 38], among others. Extensions to jump-diffusions can be found in [12] and [22].

SC models are intuitive in terms of economic interpretation and deliver analytically tractable expressions. However, they experience all consequences stemming from the highly reduced observability of their primitives. First, commodity markets do not quote instantaneous convenience yields. Consequently, filtering techniques (e.g., Kalman) are required to estimate the corresponding process. However, these procedures may deliver an unstable assessment of model parameters, a fact that can undermine the significance of hedging prescriptions resulting

from a fitted model. Second, fitting quoted forward curves or term structures of volatility and correlation requires the use of calibration procedures whose effectiveness and numerical stability must be evaluated case by case. Once more, the attention of the modeler is drawn on the nature of the primitives involved in the modeling process. These considerations led a few authors to propose the *forward model* class (FD models) as a viable alternative to SC models.

Jamshidian [24] directly assigns *futures* price $F_T(t)$ dynamics under the risk-neutral probability $\mathbb{P}^*$ as

$$\frac{dF_T(t)}{F_T(t)} = \Sigma_T(t) d\mathbf{W}(t) \quad (9)$$

where the initial condition $(F_T(t_0), T \geq t_0)$ is represented by the currently quoted forward price curve and $\mathbf{W}$ is an n -dimensional standard Brownian motion. As theory requires, these processes are all $\mathbb{P}^*$ -martingales, that is, they exhibit no drift [14]. If interest rates are deterministic, model (9) can be used to describe forward price dynamics as well. This is particularly important in view of the observation that interest rate volatility is usually dominated by commodity price volatility and thus the time value of money can assumed to be nonrandom in this kind of model [37]. Any particular instance of this class results from assigning a value to each of the following three quantities:

1. a futures/forward curve quoted at time t_0 ;
2. a number n of statistically independent driving factors;
3. a volatility structure $(\Sigma_T(t), t_0 \leq t \leq T)$.

FD models have been studied by Cortazar and Schwartz [11] and Audet *et al.* [2], among others. In particular, Andersen [1] illustrates a number of alternative modeling frameworks and examines their performance in option calibration.

The main benefit from using FD models is that primitives are observable quantities, possibly up to a deterministic transformation of market price data. Moreover, FD models are very intuitive and flexible from a trader's perspective. For instance, they easily allow the user to identify noise terms with a number of forward curve deformations resulting from a statistical analysis of historical market moves. As a result, one obtains a picture of the number of significant noise driving terms as well as the way

they impact on the various sections of the forward curve (see [39], among others).

However, spot dynamics implied by FD models *via* the self-explaining assignment $S(t) := F_t(t)$ need not be Markovian processes. Jamshidian [24] first derives conditions on the volatility structure, ensuring that the resulting spot price dynamics is Markovian. Unfortunately, this and subsequent results pursued within one-factor Heath–Jarrow–Morton interest rate setting cannot be applied to SC models that involve two-factor Markovian processes. Using the Cheyette’s technique developed for multidimensional HJM interest rates models, Andersen [1] disentangles abstract Markovian factors underlying forward price dynamics. Alternatively, Roncoroni and Id Brik [36] pursue a systematic study on the direct correspondence between SC models and FD models.

A third class of models has been proposed by Miltersen and Schwartz [32]. These authors consider the notion of *instantaneous forward convenience yield* $c_T(t)$ as implicitly defined through:

$$F_T(t) = S(t) e^{\int_t^T c_s(t) ds} \quad (10)$$

for all $T \geq t$. It can be shown that the process $c_t(t)$ is actually the instantaneous convenience yield defined as the formal “instantaneous dividend rate” process $c(t)$ entering the risk-neutral drift in expression (7). More precisely, Bjork and Landen [6] show that

$$c_T(t) = \mathbb{E}_t^{\mathbb{P}^*}(c(T)) + \frac{\text{Cov}_t^{\mathbb{P}^*}\left(e^{-\int_t^T r(u) du} S(T), c(T)\right)}{\mathbb{E}_t^{\mathbb{P}^*}\left(e^{-\int_t^T r(u) du} S(T)\right)} \quad (11)$$

from which we deduce that $c_t(t) = c(t)$.

Although formula (10) defines $c_T(t)$ in terms of forward and spot prices, for modeling purposes it is actually used the other way around: processes S and c_T are defined as primitives for a set of maturities T ; then forward prices are derived using expression (10) and drift restrictions are calculated on the c_T dynamics. This methodology defines the *forward convenience yield* model class (FC models). An interesting application of this setting is represented by the calibration procedure developed in [31].

Compared to both SC and FD models, this setting has the advantage of clearly exhibiting the contribution of spot price and convenience revenue to the formation of a forward price. FC models suffer from similar drawbacks as those affecting SC models: spot prices may be unavailable in the market under consideration; also, (instantaneous forward) convenience yields are rather difficult to imply from market quotes. In particular, differentiating a term structure of forward prices resulting from fitting observed market quotes may generate a curve exhibiting pronounced spikes because of the instability of the differential operator.

A final modeling framework is perhaps the simplest and most popular one. The idea is to focus on the fine properties of a particular class of commodity spot prices without assuming any stochastic structure about the underlying convenience yield dynamics. This restrictive assumption is partly offset by the flexibility introduced upon selecting an appropriate spot price process for modeling and pricing purposes. Spot models for energy price dynamics include [1, 3, 7, 9, 13, 15, 18, 20, 23, 27, 34, 35], among others. Eydeland and Wolyniec [17] and Fusai and Roncoroni [19] report a large number of examples and references about one-factor commodity price models.

At the time of writing, there seems to be no general consensus on a benchmark setting for pricing commodity-linked derivatives. As a consequence of the arguments illustrated here, we may suggest for the modeler to focus on primitives, structural elements, and driving noise terms.

1. *Primitives.* They should be selected within the set of market quantities for which reliable data are available. In our view, these may involve an appropriate combination of spot, forward, and swap prices or indices according to considerations of representativeness and liquidity. For instance, a market model for electricity or shipping freight indices may build on average prices as introduced in [5], while the general Markovian FD models framework may provide a model generator more suitable for gas and oil markets.
2. *Structural elements.* A proper form for price drift, volatility, and jump components, if any, should be identified using statistical analysis of historical data and then fitted to observed prices. One may perform a nonparametric estimation of

these characteristics and then define an appropriate family of functions reproducing the features displayed by the estimated graphs. Besides representativeness, structural elements ought to be selected in such a way that the resulting model is stable under statistical estimation, a property that goes with the reliability of all valuation assessments and hedging prescriptions stemming from using the model in practice.

3. *Driving noise terms.* Number and nature of underlying noise terms should be assessed on the basis of historical price analysis (e.g., examination of the trajectorial properties of price paths, principal components analysis) and jump filtering. For instance, a strong deviation of market price returns measured over short-time-lagged data should be interpreted as a sign that prices experience shocks of jump nature. Also, slowly decreasing eigenvalues of the historical covariance matrix of forward price returns are an indication of the significance of asynchronous movements affecting forward curve dynamics.

References

- [1] Andersen, L. (2008). *Markov Models for Commodity Futures: Theory and Practice*, Working Paper, Banc of America Securities.
- [2] Audet, N., Heiskanen, P., Keppo, J. & Vehviläinen, I. (2002). Modelling of electricity forward curve dynamics, in *Modelling Prices in Competitive Electricity Markets*, D.W. Bunn, ed, John Wiley & Sons, pp. 251–265.
- [3] Barlow, M.T. (2002). A diffusion model for electricity prices, *Mathematical Finance* **12**, 287–298.
- [4] Benth, F.E., Cartea, A. & Kiesel, R. (2008). Pricing forward contracts in power markets by the certainty equivalence principle: explaining the sign of the market risk premium, *Journal of Banking and Finance* **32**(10), 2006–2021.
- [5] Benth, F.E. & Koekebakker, S. (2008). Stochastic modeling of financial electricity contracts, *Energy Economics* **30**(3), 1116–1157.
- [6] Björk, T. & Landén, C. (2001). On the term structure of futures and forward prices, in *Mathematical Finance—Bachelier Congress 2000*, H. Geman, D. Madan, S.R. Pliska & T. Vorst, eds, Springer Verlag, Berlin, Heidelberg, New York.
- [7] Black, F. (1976). The pricing of commodity contracts, *Journal of Financial Economics* **3**, 167–179.
- [8] Brennan, M. (1991). The price of convenience and the valuation of commodity contingent claims, in *Stochastic Models and Option Values*, D. Lund & B. Oksendal, eds, Elsevier Science.
- [9] Cartea, A. & Figueroa, M.G. (2005). Pricing in electricity markets: a mean reverting jump diffusion model with seasonality, *Applied Mathematical Finance* **12**(4), 313–335.
- [10] Casassus, J. & Collin-Dufresne, P. (2005). Stochastic convenience yield implied from commodity futures and interest rates, *Journal of Finance* **60**, 2283–2331.
- [11] Cortazar, G. & Schwartz, E.S. (1994). The valuation of commodity contingent claims, *Journal of Derivatives* **1**, 27–39.
- [12] Crosby, J. (2008). A multi-factor jump-diffusion model for commodities, *Quantitative Finance* **8**, 181–200.
- [13] De Jong, C. (2006). The nature of power spikes: A regime-switch approach, *Studies in Nonlinear Dynamics and Econometrics*, **10**(3) Article 3.
- [14] Duffie, D. (2001). *Dynamic Asset Pricing Theory*, Princeton University Press.
- [15] Escribano, A., Peña, J.I. & Villaplana, P. (2002). *Modeling Electricity Prices: International Evidence*, Working Paper 02–27(8), Universidad Carlos III, Madrid.
- [16] Eydeland, A. & Geman, H. (1998). Pricing power derivatives, *Risk* September.
- [17] Eydeland, A. & Wolyniec, K. (2003). *Energy and Power Risk Management: New Developments in Modeling, Pricing and Hedging*, John Wiley & Sons.
- [18] Fusai, G., Marena, M. & Roncoroni, A. (2008). Analytical pricing of discretely monitored Asian-style options: theory and application to commodity markets, *Journal of Banking and Finance* **32**(10), 2033–2045.
- [19] Fusai, G. & Roncoroni, A. (2008). *Implementing Models in Quantitative Finance: Methods and Cases*, Springer Finance.
- [20] Geman, H. & Roncoroni, A. (2006). Understanding the fine structure of electricity prices, *Journal of Business* **79**, 1225–1262.
- [21] Gibson, R. & Schwartz, E.S. (1990). Stochastic convenience yield and the pricing of oil contingent claims, *Journal of Finance* **45**(3), 959–976.
- [22] Hilliard, J.E. & Reiss, J. (1998). Valuation of commodity futures and options under stochastic convenience yields, interest rates, and jump diffusion in the spot, *Journal of Financial and Quantitative Analysis* **33**(1), 61–86.
- [23] Huisman, R. & Mahieu, R. (2003). Regime jumps in electricity prices, *Energy Economics* **25**, 425–434.
- [24] Jamshidian, F. (1992). Commodity option valuation in the Gaussian futures term structure model, *Review of Futures Markets* **10**, 61–86.
- [25] Jamshidian, F. (1993). Option and futures evaluation with deterministic volatilities, *Mathematical Finance* **3**(2), 149–159.
- [26] Jamshidian, F. & Fein, M. (1990). *Closed-Form Solutions for Oil Futures and European Options in the Gibson-Schwartz Model: A Comment*, Working Paper, Financial Strategies Group, Merrill Lynch Capital Markets, New York.
- [27] Jarrow, R.A. (1987). The pricing of commodity options with stochastic interest rates, *Advances in Futures and Options Research* **2**, 19–45.

- [28] Koekebakker, S. & Os Adland, R. (2004). Modelling forward freight rate dynamics—empirical evidence from time charter rates, *Maritime Policy and Management* **31**(4), 319–335.
- [29] Kolos, S.P. & Ronn, E.I. (2008). Estimating the commodity market price of risk for energy prices, *Energy Economics* **30**, 621–641.
- [30] Korn, O. (2005). Drift matters: an analysis of commodity derivatives, *Journal of Futures Markets* **25**, 211–241.
- [31] Miltersen, K. (2003). Commodity price modelling that matches current observables, *Quantitative Finance* **3**, 51–58.
- [32] Miltersen, K. & Schwartz, E.S. (1998). Pricing of options on commodity futures with stochastic term structures of convenience yield and interest rates, *Journal of Financial and Quantitative Analysis* **33**, 33–59.
- [33] Nielsen, M.J. & Schwartz, E.S. (2004). Theory of storage and the pricing of commodity claims, *Review of Derivatives Research* **7**, 5–24.
- [34] Ribeiro, D. & Hodges, S. (2005). A contango-constrained model for storable commodity prices, *Journal of Futures Markets* **25**(11), 1025–1044.
- [35] Roncoroni, A. (2002). *Essays in quantitative finance: modelling and calibration in interest rate and electricity markets*, Ph.D. Dissertation, Université Paris IX Dauphine, France.
- [36] Roncoroni, A. & Id Brik, R. (2008). *On The Relationship Between Commodity Forward Models and Spot-Convenience Yield Model*, Working Paper, ESSEC Business School.
- [37] Schwartz, E.S. (1997). The stochastic behavior of commodity prices: implications for valuation and hedging, *Journal of Finance* **52**(3), 923–973.
- [38] Schwartz, E.S. & Smith, J. (2000). Short-term variations and long-term dynamics in commodity prices, *Management Science* **46**, 893–911.
- [39] Tolmasky, C. & Hindanov, D. (2002). Principal components analysis for correlated curves and seasonal commodities: the case of the petroleum, *Journal of Futures Markets* **22**(11), 1019–1035.

Related Articles

Commodity Forward Curve Modeling; Electricity Forward Contracts; Forwards and Futures; Heath–Jarrow–Morton Approach.

ANDREA RONCORONI

Oil Market

The development of the internal combustion engine one century ago was the principal cause of the development of the oil market. In the 1950s, oil became more important than coal. Since then its importance has kept increasing, though it has been recently largely displaced by natural gas for heating and electricity generation. The rapid economic growth of China and of India has increased the world demand for oil. Currently, demand exceeds 90 billion barrels a day, stretching the global production capacity.

Bringing new oil fields to production requires typically several years and very large investments. The economic hurdles are compounded by environmental worries and by the unstable market structure of oil. Market instability is due to the low short-term elasticity of oil consumption to price changes. The large price fluctuations implied by the low elasticity of oil prices being related to trends in world economy carry a substantial risk premium that increases the cost of capital for oil projects. Over longer time periods, high prices discourage demand and increase production. Lower prices have the opposite effects. This leads to substantial mean reversion in oil prices over longer intervals. Over the long run, real oil prices have been mostly declining in real terms, with the exception of 1973–1981 and the current uptrend since 1998 (see Figure 1). Demand has been steadily increasing and it is projected that it will continue to grow (Figure 2). The concentration of oil reserves in the Middle East (Figure 3) suggests that members of the Organization of Petroleum Exporting Countries (OPEC) will cover a greater share of world production in the future (Figure 4). The very low marginal cost of production of most of OPEC oil acts as an additional deterrent to the investment necessary for the development of alternative resources.

Crude Oil

The price of crude oil varies with the ease of refining it. Lower density oil with lower sulfur content is more easily processed. The market for crude oil is based on the two benchmarks of Brent and West Texas Intermediate (WTI) oil. Changes in these benchmarks show little day-to-day correlation with changes in prices of Middle Eastern oil. It is likely that, owing

to the increasing oil flow to Asia, a new benchmark will soon emerge.

The spot market for oil is dwarfed by the futures market that currently has volumes about 15 times larger than the market for physical oil. Oil swaps bridge gaps between short-term maturities and allow also for trading over very long maturities. Historically, the WTI price was slightly higher than the Brent price, reflecting the fact that the difference in oil price across the Atlantic was centered around the marginal cost of shipping oil westward. The emergence of China as the largest oil importer has led recently to frequent spread reversals. Local imbalances are now more random because of the growing importance of the eastward flow of Middle Eastern oil.

Futures

Futures contracts require the delivery of oil at a later date. They are standardized for quantity and quality, they are traded in exchanges, and their delivery price is adjusted daily to reflect changes in value and limit credit risk (mark to market). Margin requirements ensure the funds necessary for marking to market [1].

Traders are commonly classified as hedgers, market makers, or speculators. Hedgers want to offset price risk. They regard these contracts as insurance contracts. They are willing to pay a premium to reduce their risk. Market makers provide liquidity and hope to gain from temporary imbalances of supply and demand. Speculators take a position based on their view of price trends, with no specific interest to trade physical oil.

Most futures trade for maturities up to five years, though contracts up to eight years exist. Over-the-counter trading is active up to 10 years. Mean reversion reduces the correlation of price changes at maturities far apart and limits the relevance of current shocks for far delivery. Because of low correlation, diversification across maturities may achieve significant risk reduction. Diversifying across qualities, or even across most commodities, leads to much smaller risk reduction because the price of most commodities is highly correlated with the price of oil for the same maturity.

Most futures contracts are settled by entering the opposite trade. This allows hedgers more flexibility on delivery time, place, and quality. The difference

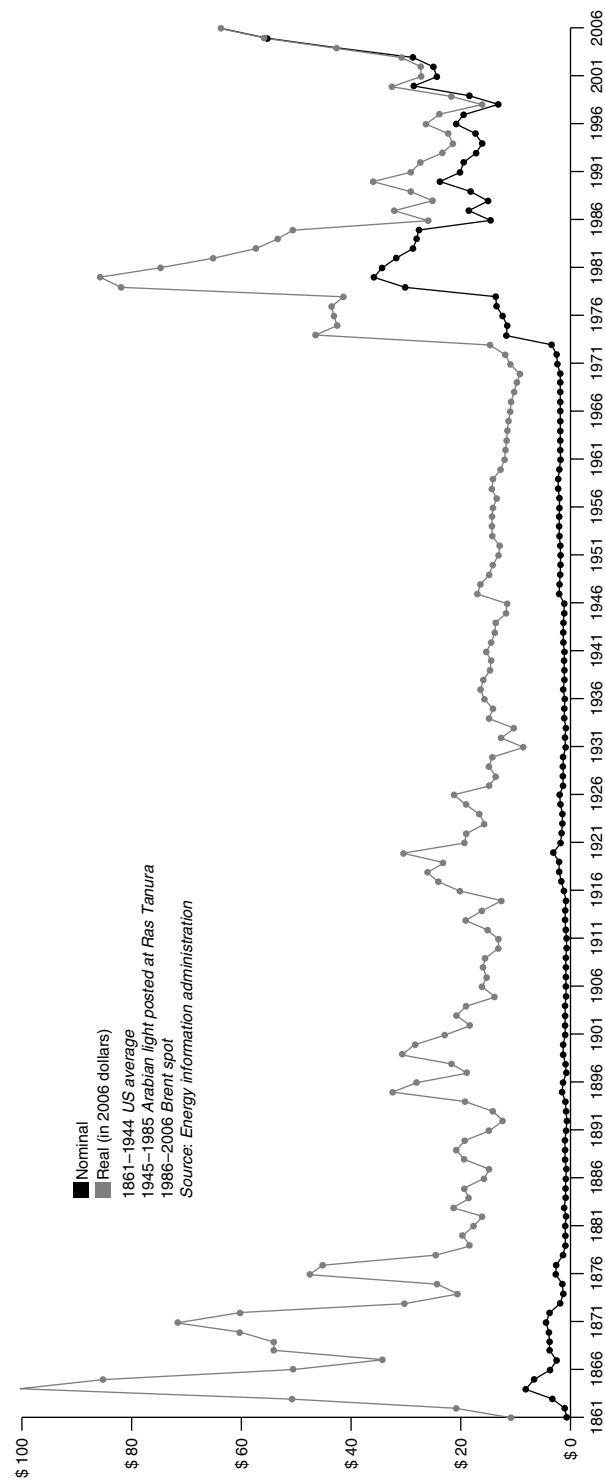


Figure 1 Real and nominal oil prices, 1861–2004

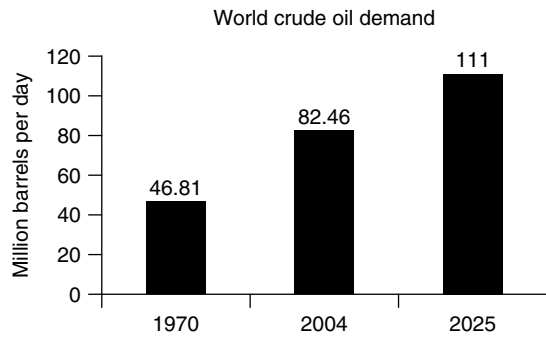


Figure 2 World oil demand

between the spot and the futures price is called *basis*. Because hedgers, after closing their future position, need to trade in the physical market, their final result will depend not only on the change in value of their futures, but also on the change of basis. That is known as *basis risk*.

Crude Oil Curve

The crude oil curve shows the price of oil for delivery at different dates. Its shape changes through time, from upward (contango) to downward (backwardation) sloping, occasionally with a hump. The shape of

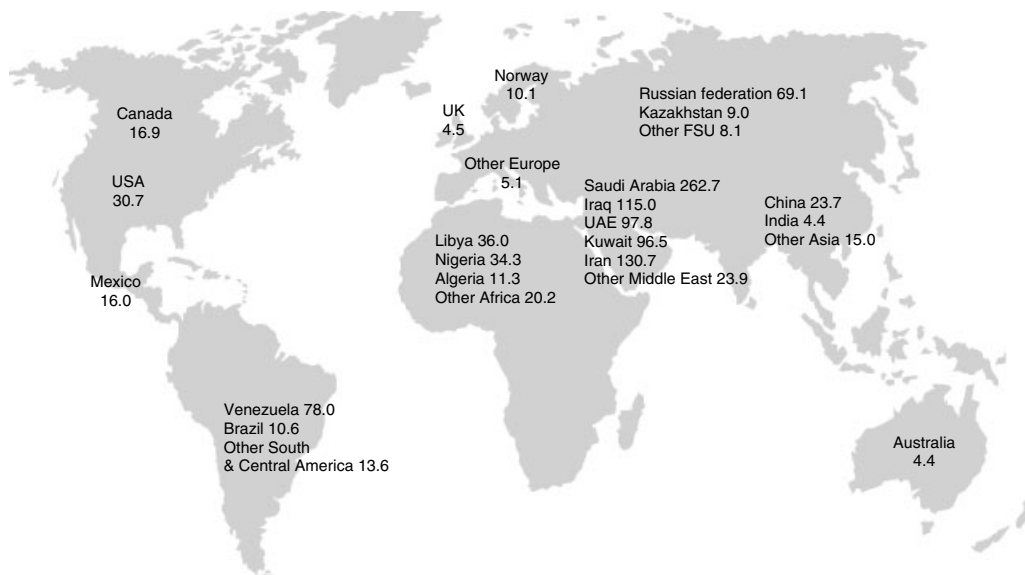
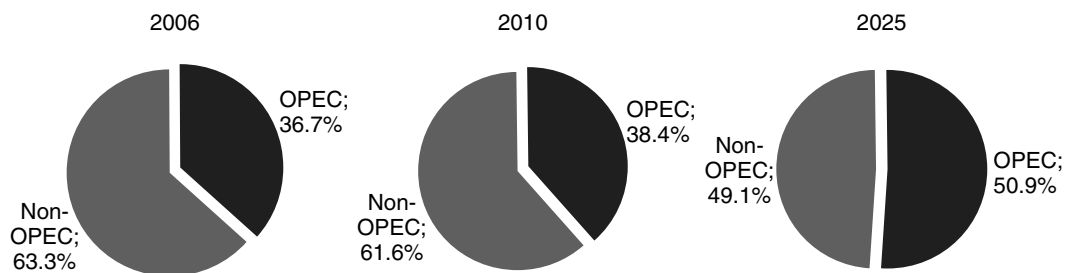


Figure 3 World oil reserves in billions of barrels, 2005



Source: BP Statistical Review

Figure 4 Market share of OPEC and other countries

the crude oil curve is influenced by the risk-free rate, the storage cost of oil, and inventory levels. When inventories are plentiful, investors are willing to hold them because the oil curve is typically upward sloping. The cost of financing and storing oil determines then the futures price in relation to the spot price. When the crude oil curve is downward sloping, the willingness of investors to hold oil is limited. Only investors who worry about the possibility that disruptions in supply may interfere with timely delivery are then willing to hold a physical inventory. The implicit cost of holding expensive oil that may only be delivered with certainty at a loss through the sale of futures contract is known as *convenience yield*. It represents the marginal benefit that investors assign to the flexible availability of oil relative to term delivery. It follows that convenience yield is negatively related to the slope of the crude oil curve. This traditional relationship has been challenged by the influx of financial institutions into the crude oil market since 2004. Moreover, oil tankers ensure that supply is very predictable for the next ninety days, while greater uncertainty dominates longer maturities.

Crack Spread

All the crude oil has to be processed in refineries. Refiners buy crude and sell refined products: gasoline,

diesel, heating oil. The difference between the price of a barrel of oil and the price of its refined components is known as *crack spread*. It represents the refiners' revenue for processing oil. It depends on the spare refining capacity available and the conditions prevailing in the markets for crude and refined products [2].

References

- [1] Geman, H. (2005). *Commodities and Commodity Derivatives. Modeling and Pricing for Agriculturals, Metals and Energy*, Wiley Finance, United Kingdom.
- [2] Kolb, R.W. (2000). *Financial Derivatives*, Blackwell Publishers Inc.

Further Reading

Hull, J. (2003). *Options, Futures and Other Derivatives*, Pearson Education Inc.

Related Articles

Commodity Forward Curve Modeling; Commodity Price Models; Commodities and Numéraire.

GIOVANNI BARONE-ADESI & CORINNE
CATTANI

Commodities and Numéraire

In the last decade, there has been considerable interest in the commodities world, in energy, in particular, as evidenced by the number of press editorials that are being published on the subject. Commodity prices have been experiencing unprecedented moves in the last few years and there is no sign of change at the date of writing (January 2009). Demand for metals, energy, and cereals from Brazil and Russia, two of the fastest growing economies, has undoubtedly increased the volatility across all commodity classes. As an example, between 2001 and 2005, China's demand for copper, aluminum, and iron increased by 78%, 85%, and 92%, respectively.

Energy prices, for example, those of crude oil and coal, have witnessed a very high volatility over the last two years, under the combined effect of a greater public awareness of the exhaustible nature of fossil energy, the massive arrival of new players into commodities, and the current looming depression worldwide. The West Texas Intermediate (WTI) crude oil, that had undergone a respite in price increase during the year 2006, resumed in 2007 its irresistible ascent to go over the symbolic threshold of \$100 per barrel in early 2008. It went above \$140 per barrel in June and July, resulting in prices multiplying by more than 500% in less than four years, with—among other reasons—supply disruptions in Nigeria, a structural decline in production in Mexico and in other countries like the United Kingdom and the United States. Subsequently, crude oil prices decreased by more than \$100 between July and November 2008, another move never observed previously. This “spike”, unique in the history of oil markets, took place a year after the start of the subprime crisis and an explanation is difficult, at least, for the remarkable upward move. Call options on crude oil with strikes of \$100 or more were the subject of great surprise when they appeared in New York in summer 2006. By May 2008, there were 21 000 outstanding contracts for the NYMEX December 2008 call options with a strike of \$200 per barrel. In parallel, put options for the same maturity of December 2008 and a strike of \$100 traded at \$1.84 in July 2008; when they settled on November 17, 2008, crude oil prices were at \$54/bbl

and the buyers of the puts realized a net return of 2200%!

Commodities and Numéraire: Some Qualitative Aspects

Since commodities are essentially denominated in dollars, the weakness of the “numéraire”^a currency, at least until September 2008, has often been cited as a major cause for the rise of commodity prices. Still, in constant dollars, average crude oil prices rose by 124% over the period 2002–2006; and the increase up to July 2008 was even steeper. The entire WTI crude forward curve was trading over the level of \$105 with a backwardated shape in July 2008. Since the oil price more than halved the level of the forward curve changed to lower values and its shape went to a steep contango, with medium-term forward prices being higher than the first nearby price by more than \$15. And in the oil market, the contango shape has essentially been associated with a situation of crisis, as during at the time of the First Gulf War or after September 11, 2001 (Figure 1).

The same volatile market conditions across the spectrum of the three commodity classes—energy, metals, and agriculturals—are unlikely to drastically improve in the near future, with land itself becoming rare and water insufficient, disruptions occurring (like in 2008 in South African mines that represent, in particular, 80% of the world platinum production) because of electricity shortages, geopolitical issues in a number of countries producing commodities, and the world population growth. We can observe that these elements illuminate an interesting property, namely, that the three commodity subasset classes are increasingly *correlated*, a property that we can certainly view as novel in the commodity markets. Hence, energy companies like agri-food firms, bankers, and private equity funds, have all moved forward in the acquisition of new physical assets such as power plants, gas storage facilities, aluminum smelters, or grain elevators.

Looking back into the past, one should keep in mind that the history of commodities has been filled with booms, busts, seasonal volatility, weather events, geopolitical tensions, and occasional attempts to “corner” the market, features that were reasonably acting as a deterrent for new entrants. Moreover, the physical constraints of delivery and storage make spot

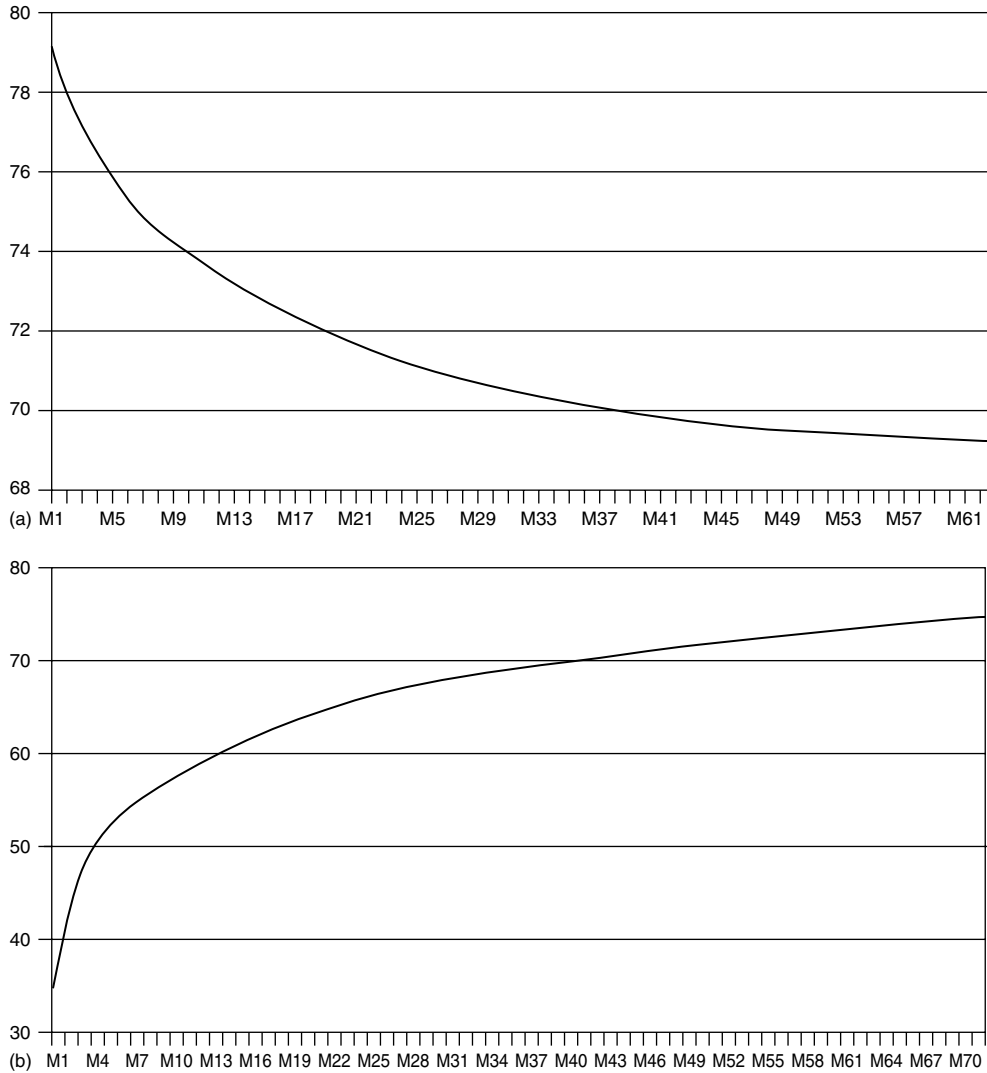


Figure 1 (a) Oil futures price curve (September 14, 2007); (b) Crude oil futures price curve (January 14, 2009)

commodity trading difficult or impossible; transactions on commodity futures exchanges can only be performed through a broker who is a member of the exchange (hence, the remarkable success of the recently introduced exchange-traded futures (ETFs)), that are accessible to individual investors. All market participants are increasingly aware of the fact that volatility swings reflect not only supply and demand concerns but also that these markets are sometimes dominated by large players, often opaque, and always directed by the long lead time between a production

decision and its actual viability. Trading volumes have also fluctuated widely over time: during the high inflation era of the 1970s, commodity futures trading exploded and the real estate sector boomed since bonds and stocks were generating a real return close to zero for over a decade. The commodity boom at the end of the 1970s can be identified with the crash in 1980 of precious metals prices, when the famous squeeze of the silver market by the Hunt Brothers (who were holding at some point an estimated 50% of the global deliverable supply of silver)

failed. Afterwards there was a long period, nearly 20 years, of stagnation in commodity prices that some experts attribute to supply fundamentals and low inflation, which lasted until the early 2000s. The present effects of the worldwide recession, together with the awareness that at any time the amount of every exhaustible commodity is *bounded*, tend to pull analysts' expectations in opposite directions.

Returning to the analysis of the numéraire effect, we plotted the trajectories over the period 2000–2008 of crude oil (Figure 2), copper (Figure 3), and gold prices (Figure 4) with respect to two “currency numéraires”, namely, the dollar and the euro; and the CRB (Commodity Research Bureau) index, which has been perceived for a long time in the United States as a very good proxy for inflation. For the

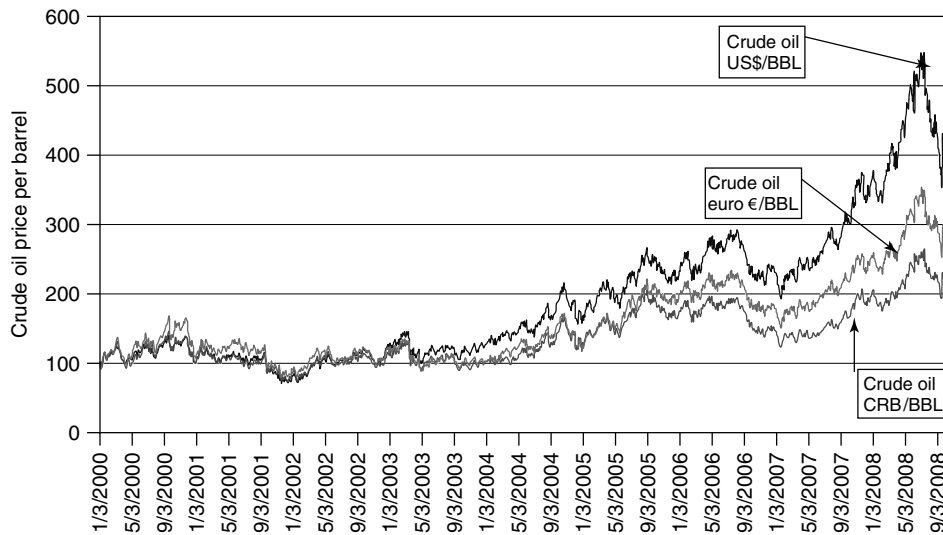


Figure 2 Crude oil relative prices

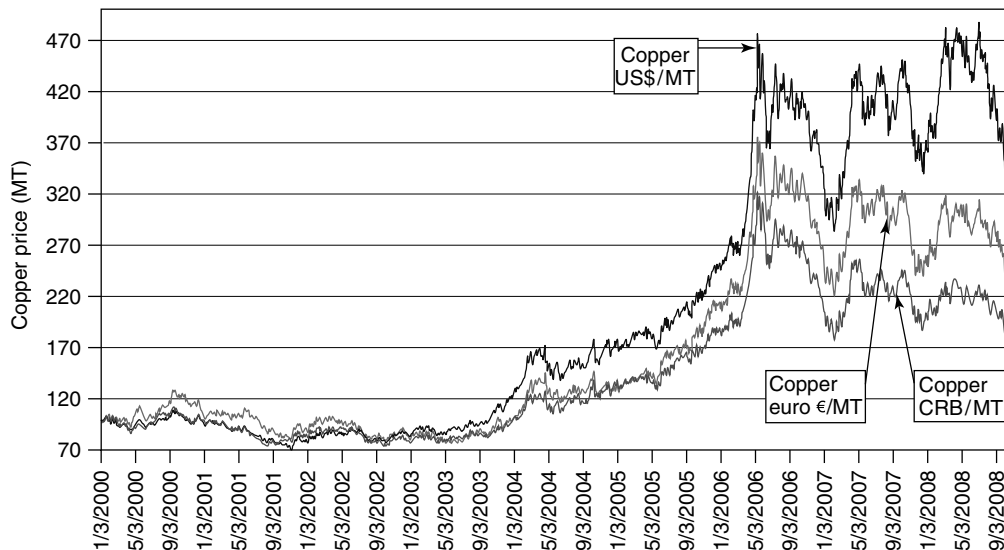


Figure 3 Copper relative prices

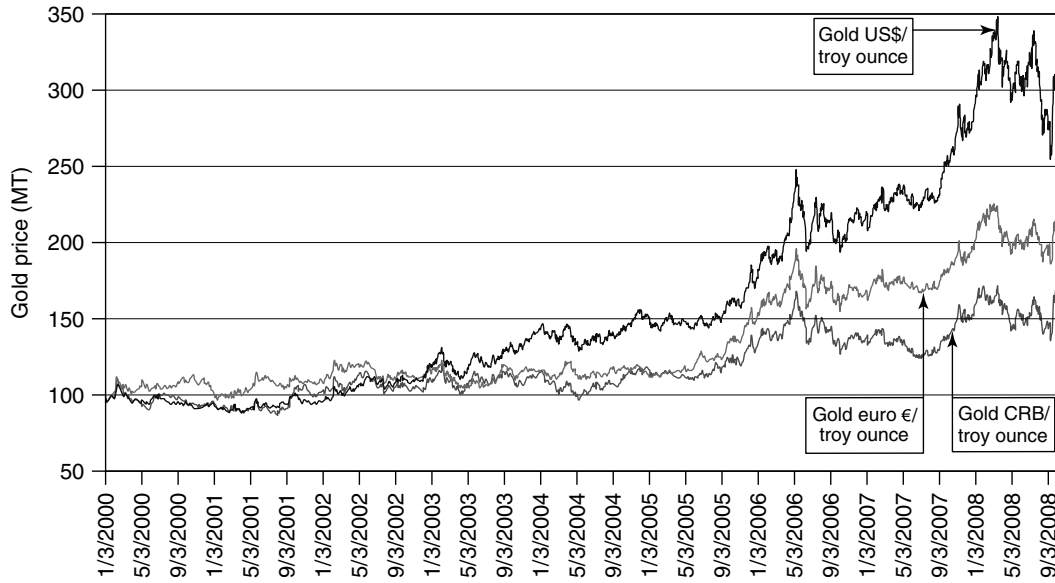


Figure 4 Gold relative prices

sake of comparison, we have normalized all relative prices to have a value of 100 at the starting date of analysis.

We observe that for both oil and copper, the change of numéraire has a remarkably little impact on the shape of the three trajectories, indicating that major effects such as supply and demand or the massive arrival of index funds and new market participants are the ones driving the commodity price changes.

Interestingly, the same property holds in the case of gold, namely, the absence of sensitivity of the trajectory shape to the numéraire currency or index. This is consistent with the fact that—for the current period at least—gold has lost its status of a special currency since its price has been falling sharply together with all commodity prices.

Commodities and Numéraire: Some Quantitative Aspects

Brief Reminders

A large and fascinating literature has established that the assumptions of no arbitrage, constant interest rates, and a general semimartingale for the stock price S lead to show the existence of a probability measure

Q such that the process $(S(t)e^{-rt})$ is a Q martingale. It is worth noting that this discounted price $S(t)e^{-rt}$ is nothing but the price of S with respect to the money market account numéraire M , where $M(0) = 1$ and $M(t) = e^{rt}$. As for the martingale representation, it brings a windfall of beautiful properties, including an alternative proof of the Black and Scholes [1] formula, as well as an essentially simple valuation of path-dependent options such as barrier, lookback, or Asian.

When moving to stochastic interest rates, M is not a valuable numéraire for pricing of contingent claims maturing at date T since it is not any more a “riskless asset” relative to date T . Instead, as shown in [4], the proper numéraire^b becomes the default-free zero-coupon bond B^T maturing at date T , that satisfies $B^T(T) = 1$. Moreover, the price of the risky stock S with respect to numéraire B^T is defined as $f^T(t) = S(t)/B^T(t)$ and has two fundamental properties:

- It represents the forward price of S , and in commodity markets, forward contracts are the central instruments.
- It is a martingale with respect to Q_T , the T -forward-adjusted probability measure, and hence offers an easy extension to stochastic interest rates of the Black and Scholes, Merton [1, 15], and other option pricing formulas.

From Stocks and Bonds to Commodities

Our view when turning to commodities is multifold and, in summary, is as follows:

1. Introducing stochastic *versus* constant interest rates when trading commodities (forwards/futures or options) is not a big gain in precision since interest rate moves are essentially quite small compared to commodity price moves and the recent period has been more extreme than ever in this direction.
2. This being stated, introducing stochastic interest rates is not a difficult matter since the most traded instruments in the world of commodities are the forward contracts whose prices, as mentioned before, are martingales under the “right” pricing measure, namely, the T -forward-adjusted probability measure. For futures contracts,^c the situation is different because of the margin calls (see [8] for a discussion in the case of non-storable commodities).
3. Turning now to a very different issue, which is the relevance of interest rates arbitrage models for commodities as discussed in the seminal paper [11] (see also **Heath–Jarrow–Morton Approach**), our view is that the benefit is limited. In fact, the large literature that transposed arbitrage models of interest rates may be misleading: default-free bonds (and even less so, interest rates) are very different from commodities. The world of commodities is characterized by possible squeezes, sudden lack of liquidity, and insufficient transparency, which make the assumption of no arbitrage (let alone market completeness and unicity of the risk-adjusted measure Q) essentially invalid [16].
4. The convenience yield y and the cost-of-carry relationship. The formula

$$f^T(t) = S(t)e^{(r-y)(T-t)} \quad (1)$$

introduced in the *Theory of Storage* by Keynes [13], did make sense in the minds of the famous economists Kaldor (1939) and Working (1949) because the convenience yield y —net of storage costs if defined as above—translates the preference for the ownership of the physical commodity over a forward contract as well as many other effects and market frictions, hence mitigating the importance of the no-arbitrage argument.

In the case of electricity, a nonstorable commodity, the existence of a commodity yield and the spot-forward relationship collapse together [3, 10]) as evidenced, for instance, by the abrupt convergence of the forward price $f^T(t)$ to the spot price at date T even in markets like the Nordpool, where a significant fraction of electricity is hydro [5]. The same holds for agriculturals with distant maturities because of their finite lifetime.

5. To avoid the doubtful arguments consisting in deriving the whole forward curve evolution from the spot price dynamics ($S(t)$), [2], also (see **Commodity Forward Curve Modeling**), when analyzing the risk management of a book of forward contracts (long or short, with different maturities), propose an original model consisting in replacing the spot price as the first state variable as in [9], by the average value $\bar{f}(t)$ of all liquid monthly forward contracts maturing before date H , and H a multiple of 12 to properly incorporate the seasonality in the case of natural gas and agricultural commodities ([7] for the case of soybeans).

Now $\bar{f}(t)$, the “backbone” of the forward curve—itsself the most fundamental tool when trading any commodity—becomes the numéraire with respect to which all forward prices are written as

$$f^T(t) = \bar{f}(t) \exp[s(T) - (T-t)\gamma(t, T)] \quad (2)$$

where

- a) $s(T)$ is the deterministic component of the relative cost-of-carry accounting for seasonality if it exists, expressed as an absolute quantity (for instance, the month of January effect).
- b) $\gamma(t, T)$ is the *rate of cost of carry* (including financing, convenience yield, and storage cost) relative to $\bar{f}$, assumed to be stochastic in t and deterministic in T (the latter assumption being possible relaxed).

Note that the relationship between $\bar{f}$ and f^T holds by definition of $\bar{f}$ and $\gamma(t, T)$ rather than by arbitrage arguments. Borovkova *et al.* (2006) show the validity of the model using data relative to seasonal (natural gas) and nonseasonal

(crude oil) commodities, and exhibit the property that the volatility $\gamma(t, T)$ decreases in T , an interesting extension of the Samuelson effect.

Lastly, the use of $\bar{f}$ as the relevant numéraire when trading a book of forward contracts can naturally be extended to electricity, a commodity for which we have already mentioned that the spot-forward relationship is not quite appropriate.

6. When turning to options on commodities, our main comments are as follows:
 - a) There should be a clear consistency between the two fundamental classes of derivatives, forwards, and options, and the measure $\mathcal{Q}$ (assuming its unicity) has to be derived from the forward curve where the liquidity is located.
 - b) Most financial options sold in the commodity world are backed by the physical optionalities attached to the ownership of physical assets such as a gas storage facility or a gas-fired power plant [5]. Hence, the valuation of these options is not tied to a particular model (a desirable feature after the catastrophic meltdown of the credit risk market), but rather to the “real” value of the externalized optionality.
 - c) Because of property b, many financial options on commodities are options to exchange a commodity S_1 for a commodity S_2 (for instance, crude oil for kerosene) or spread options. The natural numéraire is then S_1 and the relevant underlying the relative price S_2/S_1 , outside any assumption on interest rates. Note that Margrabe [14] (see **Margrabe Formula**) used the numéraire terminology in his remarkable paper, without introducing the associated change of measure; the latter would have made unnecessary the assumption of constant interest rates, as shown in [6].
7. Lastly, a worthwhile observation relative to the subject of commodities and numéraire is the following: let us consider the situation of a position held by an energy company based, for instance, in the United Kingdom, whose equity and profits are denominated in sterling pounds. Its portfolio typically contains crude oil, natural gas, electricity, coal, physical assets, and spot and derivatives contracts with counterparties based

in the United Kingdom, Norway, Germany, and Switzerland. In order to avoid, in a first stage, an array of currency risk, (Norwegian krone, Euro, *etc.*) and minimize the cost of risk management while exploiting the correlations between the various commodities when denominated in a single currency (US dollar for instance), one may decide to use the price X of a barrel of Brent crude oil, dollar-denominated, as the numéraire. Using affine relationships (e.g., indexation of natural gas to crude oil in Germany and other countries of continental Europe), correlations or co-integrations between the various energy commodity prices, the portfolio will be risk managed over time in terms of long and short positions with respect to crude oil. In the second stage, the “missing” components (orthogonal to crude oil) in the various other energy commodities will be accounted for; and in the third stage, the different currencies will be incorporated in the picture.

Conclusion

In a mature Walrasian economy [17], the general equilibrium output is enhanced by the optimal selection of the components of a multicommodity numéraire, according to the decisions of borrowers and lenders. The commodity evaluations that then emerge from transactions are accomplished by requiring that money in circulation be representative of this numéraire. This multicommodity numéraire is welfare optimal and should then constitute the real reserves of the monetary system.

End Notes

^aThe numéraire is the money unit of measure within an abstract economic model in which there is not necessarily actual money or currency. From a mathematical standpoint, it is a stochastic process, $X(t)$; where $X(t)$ is strictly positive almost surely at any date t and represents a self-financing portfolio.

^bNote that Jarrow [12] was the first one to introduce stochastic interest rates in option pricing and deliver the pricing formulas in the case of options on commodities and Gaussian interest rates. Jamshidian (1989) kept the

Gaussian interest rate setting to compute the prices of options on bonds. Geman [4] introduced the zero-coupon bond B^T as the new numéraire and showed that the introduction of the T -forward-adjusted probability measure Q_T does not require any assumption on the interest rate dynamics.

^cNote that if interest rates are nonstochastic, forward and futures prices are equal (in absence of credit risk), $Q = Q_T$, and the martingale property holds for both instruments under Q .

References

- [1] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**(3), 637–654.
- [2] Borovkova, S. & Geman, H. (2006). Seasonal and stochastic effects in commodity forward curves, *Review of Derivatives Research* **9**, 167–186.
- [3] Eydeland, A. & Geman, H. (1998). Pricing power derivatives, *Risk* **11**, 71–73.
- [4] Geman, H. (1989). *The Importance of the Forward Neutral Probability in a Stochastic Approach of Interest Rates*. Working Paper, ESSEC.
- [5] Geman, H. (2005). *Commodities and Commodity Derivatives: Modeling and Pricing for Agriculturals, Metals and Energy*, John Wiley & Sons, Ltd.
- [6] Geman, H., El-Karoui, N. & Rochet, J. (1995). Changes of numéraire, changes of probability measure and option pricing, *Journal of Applied Probability* **32**, 443–358.
- [7] Geman, H. & Nguyen, V. (2005). Soybeans inventory and forward curve dynamics, *Management Science* **51**(7), 1076–1091.
- [8] Geman, H. & Vasicek, O. (2001). Forwards and futures on non storable commodities: the case of electricity, *Risk* **14**, 12–27.
- [9] Gibson, R. & Schwartz, E.S. (1990). Stochastic convenience yield and the pricing of oil contingent claims, *Journal of Finance* **45**, 959–976.
- [10] Harris, C. (2006). *Electricity Markets: Pricing, Structures and Economics*, John Wiley & Sons, Ltd.
- [11] Heath, D.C., Jarrow, R.A. & Morton, A. (1992). Bond pricing and the twem structure of interest rates: a new methodology for contingent claim valuation, *Econometrica* **60**, 77–105.
- [12] Jarrow, R.A. (1987). The pricing of commodity options with stochastic interest rates, *Advances in Futures and Options Research* **2**, 19–45.
- [13] Jamshidian, F. (1989). An exact bond option formula, *Journal of Finance* **44**(1), 205–209.
- [14] Kador, N. (1939). Speculation and economic stability, *Review of Economic Studies* **7** 1–27.
- [15] Keynes, J.M. (1936). *The General Theory of Employment, Interest, and Money*, Macmillan.
- [16] Margrabe, W. (1978). The value of an option to exchange one asset for another, *Journal of Finance* **33**, 177–186.
- [17] Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.
- [18] Pirrong, C. (2008). Structural models of commodity prices, in *Risk Management in Commodity Markets: from Shipping to Agriculturals and Energy*, H. Geman, ed, John Wiley & Sons, Ltd.
- [19] Walras, L. (1874). Principe d'une théorie mathématique de l'échange, *Journal des économis 3e série* **34**(100), 5–21.
- [20] Working, H. (1949). The theory of the price of storage, *American Economic Review* **39**, 1254–1262.

Further Reading

- Cesarano, F. (2006). *Monetary Theory and Bretton Woods*, Cambridge University Press.
- Litzenberger, R & Rabinowitz, N. (1995). Backwardation in oil futures markets: theory and empirical evidence, *Journal of Finance* **50**, 1517–1545.
- Schwartz, E.S. (1997). The stochastic behavior of commodities prices: implication for valuation and hedging, *Journal of Finance* **52**, 923–937.

Related Articles

Change of Numeraire; Commodity Forward Curve Modeling; Commodity Price Models.

HELYETTE GEMAN & YIH-FONG SHIH

Swing Options

Unlike paper assets, trading in physical commodities often takes place over time and therefore involves volume as a second key state variable. Often, consumption rates of the commodity are unpredictable and make fixed delivery amounts uneconomic. To mitigate such *volume risk*, a swing contract gives the buyer the opportunity to manage fluctuating commodity demand levels in exchange for a fixed upfront fee. By exercising his or her swing up/down rights, the buyer can dynamically match supply and demand levels while hedging his or her costs. Swing options are widely offered by market makers and used extensively by major energy companies, especially in electricity and fossil fuel markets. Contracts with swing features are available as stand-alone financial tools; they can also be found embedded within structured physical transactions.

Definition and Use

A typical definition of a swing option is as follows. At times $n = 1, 2, \dots, N$ the buyer is to receive a *baseload* amount $\bar{K}$ of the commodity paying the strike price $\bar{P}$. In addition, over the life of the contract, the buyer has $N_c \leq N$ swing rights to temporarily vary these deliveries, instead requesting to receive an amount k_n . This request is usually made on a one-period ahead basis, so that the supplier has time to adjust delivery. The swing amount k_n is subject to the constraint $k_{\min} \leq k_n \leq k_{\max}$, where $k_{\min}$ (respectively, $k_{\max}$) is the minimal (respectively, maximal) allowable one-period delivery volume. The timing of these exercises is at the discretion of the buyer. Thus, depending on the needs of the buyer, the amount received can be swung up or down up to N_c times.

Let Y_n denote the discounted present value of the net cash flow beyond the income expected from the baseload contract, due to the exercise of a swing right on date n . Assuming that the cashflow amount is linear in the swing volume and no other constraints are present, it is optimal to always apply full (bang–bang) volumetric exercise, that is, $k_n \in \{k_{\min}, k_{\max}\}$. It follows that

$$Y_n = e^{-rn}(k_n - \bar{K})(P_n - \bar{P}) \quad (1)$$

where P_n is the spot price of the commodity in period n , and r is the discount rate for one period. Note that if the exercise decision was made *before* n , then Y_n in equation (1) may be negative.

Take-or-Pay Provisions

To bound possible swings from the baseload contract, global volume constraints are often imposed. A take-or-pay provision adds the constraint $S_{\min} \leq \sum_{n=1}^N k_n \leq S_{\max}$, where $S_{\min}$ (respectively, $S_{\max}$) is the minimal (respectively, maximal) total volume over the N periods. Thus, the buyer must place his or her total cumulative consumption within a predefined volume band. This rule is enforced *via* penalties on the buyer (hence take or pay) if the provision is violated.

Nomination Contracts

A nomination contract is a variant of a swing option, where the swing amounts are permanent. Thus, changing the nomination amount from $\bar{K}$ to k_n implies that the new baseload amount will be k_n for the remainder of the contract. Such ratcheting rules are common in pipeline contracts.

Refraction Periods

Some contracts impose a minimal refraction period δ between consecutive swings to prevent too rapid an exercise. Refraction periods also appear in research papers that consider swing options in continuous time; see the section Valuation and Theoretical Issues below.

Most existing contracts combine several of the above features and also include further custom rules, so that “vanilla” swing options are actually a rarity. Before proceeding with our discussion, we give a few examples of real-life swing contracts.

Swing Option in the Natural Gas Market

Consider an industrial natural gas consumer that wishes to hedge its future input costs. The consumer does not know consumption volume precisely (which depends on demand for output, operational

constraints, other suppliers, etc.), but has some volume range based on past usage. The consumer purchases a month-long swing option to offset this “volumetric risk” exposure. The swing option gives the consumer the right to buy up to 10 000 MMBtu at the price of $\bar{P} = \$6$ in each of the four weeks; the total volume must be between $S_{\min} = 10\,000$ and $S_{\max} = 30\,000$ MMBtu over the month. In this example, $N_c = N = 4$, $k_{\min} = \bar{K} = 0$, and $k_{\max} = 10\,000$.

Recall Option for an Oil Pipeline

An oil pipeline company delivers oil from hub A to an industrial customer at point B. The pipeline plans on delivering 1000 bbl/day for the next month at the forward price of $\bar{P} = \$100/\text{bbl}$. However, owing to pipeline congestion, supply disruption, or engineering problems, the pipeline may be unable to fulfill its obligation. To mitigate this risk, the pipeline purchases a recall option from its customer. Upon giving notice the day before, the recall option gives the pipeline the right to withhold delivery (or deliver less than the baseload amount) on up to three days during the month. In this example, $N_c = 3$, $N = 30$, $k_{\min} = 0$, $\bar{K} = k_{\max} = 1000$. Similar interruptible delivery contracts are common in electricity markets; see [10, 16, 22].

Load Serving Contract for an Electricity Utility

Electricity swing options have become popular with the deregulation of electricity markets, which led to volume risk management by the power serving entities (PSE). A PSE is obliged by law to provide power to its customers and must adjust its production to maintain equilibrium with the fluctuating demand. Such dynamic load management can be synthetically reproduced by a swing contract that the PSE in effect sells to its customers. By buying an offsetting swing option on its input fossil fuel (such as natural gas), the PSE can transfer the corresponding volume risk exposure to other market participants.

Valuation and Theoretical Issues

From a financial engineering perspective, a swing option is an exotic compound timing option on the underlying commodity asset. More precisely,

it closely resembles a series of American options. Indeed, if the number of swing rights is $N_c = 1$, then the swing collapses to the usual American option (see **American Options**); if the number of swing rights is equal to the number of periods ($N_c = N$), then the value of the swing contract is equal to a strip of European options for each period. Let us denote by $V(t, m, p)$ the value of owning a swing contract on date t with m remaining swing exercise rights and current price p of the commodity. Then the following *a priori* bound is always true irrespective of the underlying price process:

$$m \cdot EC(t, p) \leq V(t, m, p) \leq m \cdot AC(t, p) \quad (2)$$

where EC (respectively, AC) denotes the value of a European (respectively, American) option with expiration date N .

In the following discussion, we assume that $k_n \in \{0, 1\}$; modulo constant scaling this is equivalent to full volumetric exercise. To set up a mathematical model, it is necessary to construct a framework for the multiple exercise opportunities. Recalling the concept of stopping times, define the set of m -tuples of stopping times used to model the multiple exercises as

$$\begin{aligned} \mathcal{S}_{\alpha, \beta}^{(m)} &= \{\bar{\tau} = (\tau_1, \dots, \tau_m): \tau_i \text{ stopping time} \\ &\text{and } \alpha \leq \tau_1 < \tau_2 < \dots < \tau_m \leq \beta\} \end{aligned} \quad (3)$$

Then the optimal multiple-stopping problem corresponding to a basic swing option with maturity T reduces to computing

$$\begin{aligned} V(t, m, p) &= \sup_{\bar{\tau} \in \mathcal{S}_{t, T}^{(m)}} \mathbb{E}_t[Y_{\bar{\tau}}], \quad \text{where } Y_{\bar{\tau}} := \sum_{i=1}^m Y_{\tau_i} \\ \text{and } P_t &= p \end{aligned} \quad (4)$$

Beyond finding the value of the supremum in equation (4), it is of interest to know whether this supremum is actually attained (existence of optimal swing exercise strategy), and if yes, finding a computational algorithm giving a “minimal” set of optimal exercise times $\bar{\tau}$. For models of equation (4) in continuous time, the refraction time constraint $\tau_i \geq \tau_{i-1} + \delta$, $\delta > 0$ must be added to equation (3); otherwise $\tau_{i-1} < \tau_i$ is vacuous.

The problem (4) is best solved by an inductive procedure based on the solution of a sequence of

single-exercise American options. For example, Carmona and Touzi [9] have proved that

$$V(t, m, p) = \sup_{t \leq \tau \leq T} \mathbb{E}_t [Y_\tau + V(\tau, m - 1, P_\tau)] \quad (5)$$

where the maximization is over all stopping times τ , and where $\mathbb{E}_t$ denotes conditional expectation at time t . Equation (5) is of the same flavor as the American/Bermudan option problem, but with additional state variable m .

In a Markovian setting, an optimal swing exercise policy is characterized by an *exercise boundary* B , similar to American options, that is, the exercise boundary divides the commodity price space into a *continuation region*, where the holder of the swing does nothing, and an *exercise region*, where it is optimal to exercise a swing right. However, analysis of the swing exercise boundary is complicated by the fact that today's exercise reduces future delivery flexibility. Thus, the swing exercise boundary is a function both of time and number of remaining rights, $B = B(t, m)$. In certain settings [6, 9], it can be shown that $B(t, m)$ is connected, and, moreover, is monotone in number of rights m and, similarly, is monotone in t . However, general properties of $B(t, m)$ remain unresolved. An excellent summary of the structural features of swing options can be found in the book by Eydeland and Wolyniec [15, Chapter 8].

The major difficulties in valuing and analyzing swing options arise because of the high-dimensional setting and the contract's path-dependent features. Analysis of a basic swing option involves two state variables, namely, the current price P_n and the remaining number of exercise rights M_n . If a take-or-pay provision is present, then one must also consider the current cumulative volume $S_n := \sum_{j=1}^n k_j$. Both the M and S state variables depend on the particular swing exercise policy and are therefore path and control dependent. This precludes simple application of dynamic programming to equation (5). Moreover, models of underlying commodity price processes (see **Commodity Price Models**) are often complex and include multiple factors to capture the observed seasonality, mean reversion, and nonlinearity in prices. Recall that if the exercise decision is made on a forward basis, modeling of the spot-forward relationship (see **Commodity Forward Curve Modeling**) is also needed.

To illustrate the difficulties involved, consider a simple binomial tree model that assumes a one-dimensional random walk description of the underlying commodity price (P_n). To price a swing option, we then need to set up a *forest* of N_c such trees, with a tree for each possible number of remaining exercise rights. One can then proceed using the backward dynamic programming algorithm like for standard American options (see **American Options**), where a swing exercise triggers a jump from a tree with $M_n = m$ remaining rights to a tree with $M_n = m - 1$. Such an algorithm imposes memory and space constraints and raises efficiency issues. More sophisticated models might involve multiple factors or jumps (or take-or-pay provisions) and would require even more complex forests of trees.

Historical Perspective

Initial discussion of swing options appeared in the early 1990s in energy practitioner magazines; see, for example, [3, 11, 12]. At the same time, a limited literature studied the embedded multiple exercise opportunities [14, 25, 30]. Theoretical treatment of swing options was first carried out rigorously by Jaillet *et al.* [21]. The connection to multiple-stopping problems was established by Carmona and Touzi [9], who showed in a certain mathematical framework that a swing option is equivalent to N_c compound American (optimal stopping) options on the underlying; see also (5). This fact suggests pricing of swing options based on existing methods for American options. In a simplified model, with single underlying risky asset, swing options could be, for example, priced by solving partial differential equations like in the classical Black–Scholes theory (see **Finite Difference Methods for Early Exercise Options**). However, nonlinearities rule out solutions in the classical sense, and complex free boundaries make the design and control of numerical schemes quite a challenge.

To overcome these challenges, a variety of approaches have been suggested, including other partial differential equation methods ([13, 31] and see **Finite Element Methods**), stochastic dynamic programming [2, 17], the aforementioned binomial and trinomial trees ([21] and see **Tree Methods**), Monte Carlo simulation methods ([4, 16, 20, 27] and see

Bermudan Options), approximate techniques ([14, 23, 29] and *see Exercise Boundary Optimization Methods*), and optimal quantization (*see Quantization Methods* and [1]). Also see [16, 19, 24, 32] for analytic pricing of swing contracts under certain exotic models for the underlying commodity and [28] for game-theoretic aspects of swing options. Beyond the basic contracts, the modern practitioner/research trend is to value swing options using simulation tools that can achieve excellent computational complexity properties with respect to underlying models. However, as a trade-off, the error analysis of many simulation schemes is difficult, and usually only a lower bound on price can be obtained.

Hedging Issues

Hedging of swing options can be carried out in the same framework as American options. However, several additional issues make hedging difficult in practice. First, recall that many buyers of swing options are industrial customers. Empirical evidence [15] suggests that such buyers often exercise suboptimally to meet their other obligations (e.g., as suppliers of electricity). Suboptimal exercise may be justified if the spot market is not sufficiently liquid, as is the case in many commodities, especially those with strong locational characteristics. This leads to nontrivial financial implications to sellers of swing options who hedge optimally.

Secondly, swing exercise decisions are typically made one-day-ahead on the basis of forward prices. Since the cashflow is based on spot prices, basis risk between forward and spot price becomes a cause of concern (both to buyer and seller of swing contracts). This is particularly so in electricity markets, where the commodity is nonstorable and there is no definite convergence between day-ahead and real-time prices.

The underlying commodity market may be financially incomplete (*see Complete Markets*) for other reasons. As such, the seller of the swing option cannot fully hedge his or her exposure and therefore the no-arbitrage risk-neutral valuation method might not make sense. Finally, in some less liquid markets, the buyer may be exercising market power by manipulating demand through his or her swing policy and therefore affecting future prices.

Extensions and other Contracts with Multiple Exercise

With wide use of swing options, other modifications to the basic contract have become common.

Chooser Flexible Caps

A cap contract with strike $\bar{P}$ on a three-month interest rate is a portfolio of options on the quarterly interest payments that have to be made. Let us assume that the life of the contracts is N quarters, and the nominal is X . Each individual option is called a *caplet*. It covers one of the N quarters, and its payoff is $X \cdot \Delta t \cdot (L_3 - \bar{P})^+$, where Δt is the day-count fraction, and L_3 denotes the quarter-end three-month LIBOR rate. A *chooser flexible cap* is a contract such that, at each reset date, the holder of the contract has the right to decide whether to exercise that particular caplet and count it as part of the N_c allowable ones, or spare it for later use, each decision being final. The remaining caplet rights expire worthless if not used before the end of the N periods. A chooser cap reduces to a regular cap (*see Caps and Floors*) when $N_c = N$, while it can be viewed as a standard American/Bermudan option on the interest rate L_3 when $N_c = 1$. Valuation of such contracts has been studied in [27].

Two-sided Swing Options

Some physical contracts combine the swing and recall features described above. Thus, the baseload contract is augmented with N_r recall rights for the party delivering the commodity, and N_c swing rights for the buyer. The interaction of these rights leads to a stochastic *game* between the two counterparties, with the buyer being the minimizing agent, and the seller the maximizing agent. Valuation of such two-sided swing options is related to Dynkin games as explained in [5].

Gas Storage

The owner of a gas storage facility attempts to maximize revenue by injecting gas into storage when prices are low and withdrawing gas when prices are high. Exercise of such flexibility can be

viewed as a nomination swing option where k_n is allowed to be both positive and negative (to model withdrawals). The volume constraints $S_{\min} \leq S_n \leq S_{\max}$ then correspond to the physical size constraints of the facility. Gas storage valuation is complicated by additional storage costs, which add further path dependency to the problem; (see [4, 8]).

Tolling Agreements

In a *tolling agreement*, the buyer is granted use of a commodity-generating facility for a fixed period of time. The buyer can then run the facility when commodity price is high and keep it shut down when prices are low. This is again a nomination swing option with additional engineering constraints, such as fixed switching costs. Modeling of such contracts fits within the more general framework of optimal switching problems [7, 18].

Employee Stock Options

Company employees are often granted multiple stock options in their employers (see **Employee Stock Options**). Partial exercise of such options by risk-averse employees is similar to a swing call option, as multiple exercises must be considered [26].

References

- [1] Aj Bardou, O., Bouthemy, S. & Pagés, G. (2007). *Pricing Swing Options Using Optimal Quantization*, available at <http://www.arxiv.org/0705.2110>.
- [2] Baldick, R., Kolos, S. & Tompaidis, S. (2006). Interruptible electricity contracts from an electricity retailer's point of view: valuation and optimal interruption, *Operations Research* **54**, 627–642.
- [3] Barbieri, A. & Garman, M.B. (1996). Understanding the valuation of swing contracts, *Energy and Power Risk Management* **1**.
- [4] Barrera-Esteve, C., Bergeret, F., Dossal, C., Gobet, E., Meziou, A., Munos, R. & Reboul-Salze, D. (2006). Numerical methods for the pricing of swing options: A stochastic control approach, *Methodology and Computing in Applied Probability* **8**(4), 517–540.
- [5] Carmona, R. (2007). *Monte Carlo American Exercises, 2007*. Lecture Notes, Sixth Winter School in Financial Mathematics Jan. 22–24, 2007 Lunteren, Holland.
- [6] Carmona, R. & Dayanik, S. (2008). Optimal multiple stopping of linear diffusions, *Mathematics of Operations Research* **33**(2), 446–460.
- [7] Carmona, R. & Ludkovski, M. (2008). Pricing asset scheduling flexibility using optimal switching, *Applied Mathematical Finance* **15**(6), 405–447.
- [8] Carmona, R. & Ludkovski, M. (2009). Valuation of energy storage: an optimal switching approach, *Quantitative Finance* (To appear).
- [9] Carmona, R. & Touzi, N. (2008). Optimal multiple stopping and valuation of swing options, *Mathematical Finance* **18**(2), 239–268.
- [10] Cartea, A. & Williams, T. (2008). UK gas markets: the market price of risk and applications to multiple interruptible supply contracts, *Energy Economics* **30**(3), 829–846.
- [11] Clewlow, L., Strickland, C. & Kaminski, V. (2001). Valuation of swing contracts in trees, *Energy and Power Risk Management* **6**, 33–34.
- [12] Clewlow, L., Strickland, C. & Kaminski, V. (2001). Risk analysis of swing contracts, *Energy and Power Risk Management* **6**.
- [13] Dahlgren, M. (2005). A continuous time model to price commodity-based swing options, *Review of Derivatives Research* **8**(1), 27–47.
- [14] Davison, M. & Anderson, L. (2003). *A Recursive Framework for Approximating Early Exercise Boundaries of Electricity Swing Options*.
- [15] Eydeland, A. & Wolyniec, K. (2003). *Energy and Power Risk Management: New Developments in Modeling, Pricing and Hedging*. John Wiley & Sons, Hoboken, NJ.
- [16] Figueroa, M.G. (2006). *Pricing Multiple Interruptible-Swing Contracts*. Birkbeck Working Papers in Economics and Finance 0606. School of Economics, Mathematics & Statistics, Birkbeck, June 2006.
- [17] Haarbrücker, G. & Kuhn, D. (2009). Valuation of electricity swing options by multistage stochastic programming, *Automatica* **45** (To appear).
- [18] Hamadène, S. & Jeanblanc, M. (2007). On the starting and stopping problem: application in reversible investments, *Mathematics of Operations Research* **32**(1), 182–192.
- [19] Hambly, B., Howison, S. & Kluge, T. (2007). *Modelling Spikes and Pricing Swing Options in the Electricity Markets, 2007*. Working paper, University of Oxford.
- [20] Ibáñez, A. (2004). Valuation by simulation of contingent claims with multiple exercise opportunities, *Mathematical Finance* **14**(2), 223–248.
- [21] Jaillet, P., Ronn, E.I. & Tompaidis, S. (2004). Valuation of commodity-based swing options, *Management Science* **50**(7), 909–921.
- [22] Kamat, R. & Oren, S.S. (2002). Exotic options for interruptible electricity supply contracts, *Operations Research* **50**(5), 835–850.
- [23] Keppo, J. (2004). Pricing of electricity swing options, *Journal of Derivatives* **11**, 26–43.
- [24] Kjaer, M. (2008). Pricing of swing options in a mean reverting model with jumps, *Applied Mathematical Finance* **15**(6), 479–502.

- [25] Lari-Lavassani, A., Simchi, M. & Ware, A. (2001). A discrete valuation of swing options, *The Canadian Applied Mathematics Quarterly* **9**(1), 35–73.
- [26] Leung, T. & Sircar, R. (2009). Accounting for risk aversion, vesting, job termination risk and multiple exercises in valuation of employee stock options, *Mathematical Finance* **19**(1), 99–128.
- [27] Meinshausen, N. & Hambly, B. (2004). Monte Carlo methods for the valuation of multiple exercise options, *Mathematical Finance* **14**, 557–583.
- [28] Pflug, G.C. & Broussev, N. (2009). Electricity swing options: Behavioral models and pricing, *European Journal of Operational Research*.
- [29] Ross, S.M. & Zhu, Z. (2008). On the structure of a swing contract's optimal value and optimal strategy, *Journal of Applied Probability* **45**(1), 1–15.
- [30] Thompson, A.C. (1995). Valuation of path-dependent contingent claims with multiple exercise decisions over time: the case of take or pay, *Journal of Financial and Quantitative Analysis* **30**, 271–293.
- [31] Winter, C. & Wilhelm, M. (2008). Finite element valuation of swing options, *Journal Of Computational Finance* **11**(3), 107–132.
- [32] Zeghal, A.B. & Mnif, M. (2006). Optimal multiple stopping and valuation of swing options in Lévy models, *International Journal of Theoretical and Applied Finance* **9**(8), 1267–1297.

Related Articles

American Options; Bermudan Swaptions and Callable Libor Exotics.

RENÉ CARMONA & MICHAEL LUDKOVSKI

Emissions Trading

There are numerous forms of environmental impact associated with human activity. The best known in the energy sector are emissions to air, particularly carbon dioxide, sulfur dioxide (SO_x), nitrogen oxides (NO_x), and other emissions such as mercury. Regulatory intervention can limit these emissions in a number of ways.

Our aim here is to show how management under the “cap-and-trade” regulatory mechanism for emission limitation can be partially modeled in terms of standard instruments for financial derivatives. We discuss briefly how similar modeling techniques can be applied to more complex environmental factors such as emission of particulates, impact on water, and impact on local amenity. The modeled entity is principally the price of the cap-and-trade allowance, and we also briefly consider the shadow price of other forms of environmental limits, and other features, such as the emissions cap and its allocation.

The politics, economics, and science of emissions impact and emission abatement, and the various regulatory schemes, are well covered in other texts and are not covered here. We offer a caveat here; while we characterize the microeconomics of emission abatement using supply and demand curves, we should be aware that the long-term nature of emission impacts are, in some aspects, more amenable to a classical economic analysis than the neoclassical analysis that standard quantitative approaches to derivative instruments necessarily entail.^a Beyond the politics and economics, the essential challenge in the modeling of allowance prices in particular, is the capture of the essential features of the allowance, so that the pricing problem can be framed in terms of standard quantitative analysis of derivatives.

Regulatory Interventions

The key methods for regulatory intervention can be divided into (i) technology prescription; (ii) instantaneous limits which are usually measured by instrumentation; (iii) annual limits which are commonly measured by estimation using the average composition of the fuel; and (iv) taxes and subsidies. Commonly, more than one intervention will apply.

An efficient form of regulatory intervention to optimize emission abatement is a tax at the marginal damage rate, if known accurately. See [6] this may be regional, periodic or otherwise varying across time, and stochastic according to the evolution of information. This is currently impractical for a variety of reasons including agreement on current benchmarks, definitions, and included sectors, changing purchasing power of international currency in different countries, and the difficulty in securing agreements.

In the pure cap system, each registered source (or collection of sources under one company) of emission has an annual limit that may not be exceeded and which is allocated, with or without cost, to emitters. Relative to the pure tax system, this system has the practical benefit of easier implementation, limited step change, and relatively straightforward definition of source, which can change over time. So, for example, international air travel can be included in the scheme after its inception. The total cap should, in theory, be increased as the emission comes within the scope of the scheme.

In the cap-and-trade system, each source is assigned an emission allowance for the compliance period.^b The source can emit more if it purchases allowances from other sources, and can sell allowances if it can reduce its emissions below the cap. Initial allowances are commonly related to past emissions, and allocations are reduced over time. Ultimately, the free allocation of allowances can be reduced to zero, with the central administrator selling a fixed (i.e., maintaining a set amount of emissions) or variable (maintaining a set tax rate) amount of allowances.

Here, we pay attention to the quantitative analysis of allowance prices in the cap-and-trade system.

Framework

From an analytic perspective, the most important features of the regulatory framework are as follows:

- The compliance periods—length and phase.
- Banking and borrowing—regulatory exchange rate between allowances in different compliance periods.^c
- The use of “grandfathering” for allocation—the extent to which caps are dependent on historic emission.

- Allocation to new facilities and withdrawal for closing facilities.
- Buyout mechanisms—a penalty mechanism within the overall scheme, or government withholding of their allowance allocations, allowing a penalty mechanism within the country without affecting the global cap.
- Fines—for emitting without allowance.
- Relationship between schemes—allowed exchange between schemes, double count of emissions, and emissions missing from schemes.
- Recycling of fines, buyout, and other revenues to allowance “producers”.
- Sequestration allowances for sinks.
- Coverage—countries, sectors, producers, and consumers.
- Restrictions and other scheme details—for example, in the United Kingdom, Climate Change Levy Exempt Certificates can be borrowed for only five periods in a row.

Using Supply Stacks

Let us first assume that the only way to reduce the emission is to “switch on” a series of installed abatement technologies at the expense of a (upward sloping and curving) marginal cost. Assuming steady demand and a constant generation mix, the abatement “stack” can be simply represented. The effect of a rate limit in tonnes per day on the “production stack” of allowances is linearly related to the residual time in the compliance period (see Figure 1).

In the absence of banking and borrowing of allowances (which we treat as a 0% and $\infty\%$ inter-compliance period exchange rate, respectively), then it is easy to see not only why allowance prices become

extremely volatile as the end of the compliance period arrives but also how this volatility is itself volatile according to the prevailing use of technology (i.e., whether we are operating at the higher or lower cost part of the allowance production curve).

When considering the installation of new technology to produce allowances, we can treat each technology as having a capital and fixed cost (the option premium) and a marginal cost (the strike price for the technology owner). We expect our technology to endure across compliance periods and hence this creates a shadow exchange rate and option inter-dependence between periods. If we apply a simple rule that ensures an exact recovery of total fixed and variable costs in each compliance period, then the marginal price offered is stochastic according to the past and planned fixed cost recovery within the compliance period. So, for example, the offer cost from a technology may be lower if it has run at a high load factor during the compliance period and thereby recouped most of its fixed costs, or as a result of high forward prices within the compliance period the planned load factor is high and therefore the offer cost can be relatively low.

Allowances can be “produced” in a number of other ways including demand reduction, sequestration and offset, and banking and borrowing.

Sources of Uncertainties

There are three principal sources of uncertainty:

- production cost stack
- demand for energy—high energy demand creates high allowance demand
- regulatory change.

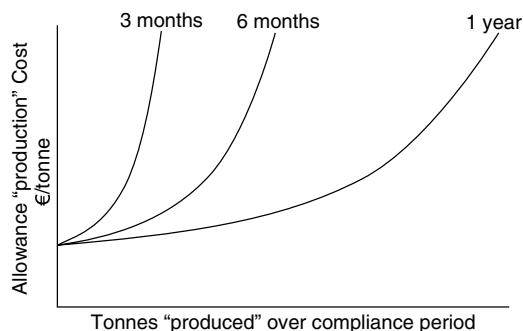


Figure 1 The cost curve for allowances in relation to the horizon to the end of the compliance period

Regulatory change uncertainties are (i) change in framework; (ii) change in coverage; and (iii) changes in the figures under the framework (cap, buyout price, banking and borrowing rates, etc.). These changes are themselves influenced by changing knowledge about damage rates, sector outputs, and political effects.

Price Characterization

If there is no banking or borrowing across a compliance period, then the steepening of the abatement stack that can be applied to the residual period causes an increase in volatility. Since the compliance period is set in calendar time, rather than having a rolling horizon, the volatility term structure is therefore not stationary. Similarly for any nearby compliance period, there is no equilibrium price to revert to, although there are a number of reference prices, such as the short-run marginal cost of abatement, the fully loaded cost of abatement (loading the fixed costs onto the short-run marginal costs to ensure precise cost recovery over the compliance period), value of lost load, buyout prices, and gaming prices. An example of a gaming price is the price outcome from Cournot competition between allowance producers.

In the same way that monthly futures can be regarded as derivatives of forward price curves with daily resolution, compliance period allowances can be regarded as derivatives of the daily allowance price curve, which itself is constructed from the abatement stack. In common with futures, allowances can be regarded as “cheapest to deliver”. There is never merit in submitting the allowance prior to the end of the compliance period, but due to the limit on the “production rate” of allowances, the rate of physical production of allowances can have complex structures within the compliance period. This is a standard optimization problem used, for example, in hydro reservoir optimization. The cost of capital should be taken into account when optimizing. The forward price curve for allowances in each compliance period is a discrete vector, with as many forward allowance prices as there are actual and expected compliance periods. To a degree, this allows us to treat the long-term problem as a “Bermudan” rather than “American” one, although the latent price structure within the compliance period is important.

- Long-term prices—since the economic theory is that the regulatory cap should be adjusted such that the cost of marginal abatement enacted is equal to the marginal damage rate, there is, in theory, a long-term equilibrium price. Indeed, since climate change has such universal impact, we may regard the optimal long-term allowance price as a , and possibly *the*, international numeraire asset.^d However, the uncertainties surrounding the marginal damage rate, the difficulty of denominating the allowance in currency, and the difference between practically achievable allowance prices and the “ideal” price are such that the long-term price has a very high degree of uncertainty associated with it, and the long-term price can change on an instantaneous basis.^e Hence we should regard the long-term price as highly volatile.
- Negative prices—market prices for allowances cannot have negative values, and hence allowances can be regarded as assets in this sense. However, under a pure cap regime, the shadow prices for allowances can be negative since ongoing allocations are related to past emissions, and production below the allocation risks an allowance reduction in future.
- Embedded options—schemes such as the Renewables Obligation scheme in the United Kingdom can have buyout prices, which allows the retail supplier to pay the buyout price instead of procuring allowances. So the cost is capped for retail suppliers, while the Renewable Obligation Certificate (ROC) price, created by the recycling of buyout revenues, is higher and volatile. However, if the cost of ROC production falls below the buyout price, suppliers will buy ROCs instead of paying the buyout. Therefore retail suppliers have embedded put options in their compliance costs. The compliance cost therefore has a stochastic element even if the buyout price is known.
- Government influence—there are a number of mechanisms that can be used by the government to adjust prices. For example, the government can allocate to sectors less than the national limit, leaving it with “hot air”. This increases the cost of compliance. The government can then release the hot air at low or high prices to sectors of its choice. The impact on price depends on the effect that discretionary allocation has on physical production of allowances. Given that the market

solution in theory delivers the neoclassical optimum, then from a modeling perspective, government influence in allocation should be regarded as increasing allowance prices.

- Path dependence—the two key drivers to path dependence are the cost structures of emission abatement, and the effect of the fixed cap on emissions under conditions of stochastic energy consumption. We have already seen that if “allowance producers” recover their full costs in each compliance period, that their offer price above marginal costs depends on the fixed cost recovery to date. So, for them, high historic compliance prices have a depressing effect on forward allowance prices within the compliance period. The fixed volume of the cap has the opposite effect. If energy demand has been high in the early part of the compliance period, then emissions will have been high, and the residual cap volume will have been depleted, which has an elevating effect on allowance prices. The modeling of this is very similar to that for an energy complex, which has an element of fixed energy amount. An example might be a complex with hydro reservoirs and combined cycle gas turbines. We return to the swing modeling in an electricity context below.
- Cost of risk—the cost of risk^f for traded commodities is the cost of risk of the marginal player. As the compliance period draws to a close, each player ensures that they are not short and places decreasing reliance on the market to close short positions. This may manifest itself as a substantial negative cost of risk (i.e., forward price above expectation price for the last submission date of the compliance period) in the medium short term, with a possibly positive cost of risk in the very short term as emission uncertainty decreases and installations hold (and can sell) long allowance positions. There are pressures for long-term risk to be in either direction. Allowance producers (who have positive cost of long-term allowance price risk) are commonly the same entities as the emitters (who have positive costs of long-term energy price risk).
- Hidden variables—the cost of abatement is highly dependent on the power generation technology mix operating. So, for example, low gas prices will cause combined cycle gas turbines to displace coal-fired power stations,

and the abatement stack will be dominated by technologies at coal plant. The allowance price depends on the price paths of the major fossil fuels (coal, oil, gas). A critical hidden variable is the amount of cap remaining in the compliance period.

Instrument Pricing

Here we consider that the underlying commodity is the emissions allowance for some examples of exchanges, see Nordpool, European Climate Exchange and EEX, or its index price.

- Modeling—we have the usual mixture of static optimization, stochastic dynamic programming using multidimensional trees/forests, and semianalytic techniques available. Owing to the high degree of importance of hidden nonstationary variables^g in influencing the forward allowance price, the application of one-factor Monte Carlo techniques should be undertaken with some caution.
- Options—emissions options are traded, with the main instrument being the European option. They are important because of the effect that allowance prices have on the operation of energy installations. This is best illustrated by using the cap-no-trade system with a single power plant as an example. Consider a simplified power station that is operationally flexible, never fails, has no variability in its emission abatement, and has a pure cap-no-trade regime. It can be regarded equivalently as a family of European fuel-plus-allowance-to-power options with the internal market allowance price (i.e., the shadow price) adjusting such that the compliance period limit is exactly reached, or a family of European fuel-to-power options with the number of exercises limited by the cap (variously called *swing*, *flexi-cap*, or *take or pay*). We note in each case that the option value is path dependent. If the key factor prices for electricity (coal, oil, gas, etc.) and the price of electricity are all regarded as exogenous to allowance prices, then the allowance price can be modeled using standard techniques for hydro reservoirs (the cap being treated as a reservoir, with limitations on physical production rate). In reality, these factors have interdependency that is high, periodic, complex, and variable. Across the

entire global energy complex, the emission limit, net of sequestration, is reached, but the modeling of this is excessively complex.

At any instant, optimization methods provide the efficient marginal and total cost of abatement to meet the cap. Dynamic marginal prices arise from exogenous shocks to the system, such as shock or to the cap limit. This visualization is particularly useful in assessing the codependence between the key factors and, in fact, the only truly exogenous factor in the early stages of the scheme is the cap itself. In the later stage, if the scheme is efficient, the exogenous factor is actually the allowance price.

- Intercompliance period swaps—the swap rate is determined by the caps, the abatement demand (caused by energy consumption), the abatement cost stacks in the two compliance periods, and the rules such as those on banking and borrowing and on interscheme fungibility such as the clean development mechanism for greenhouse gases.
- Compliance period swaptions—in the absence of banking and borrowing within the scheme, the swaption volatility is a result of (i) shared economics between the periods and (ii) activities that can “produce” allowances in one period at the expense of the other. In economics terminology, these are complements and substitutes, respectively. An example of the former is as follows: if a high allowance price in the near-compliance period makes an abatement scheme attractive, then while the factors that made the near-period allowance price high might have a similar effect on the later period, the enactment of the abatement scheme might have a depressing effect on the allowance prices in the later period. An example of the latter is the use of physical and commercial storage in the energy complex. For example, if a full hydro reservoir is emptied in the first compliance period, then the carbon intensity of production will be low in this period and high in the next. The ability to adjust the release of water therefore has the economic result of storage of allowances, equivalent to banking and borrowing.

Intercommodity Price and Price Movement Relationships

First consider a simple system in which we have one coal- and one gas-fired power stations in country, an in-country, cap-no-trade SO_x regime for the coal-fired stations, exogenously determined nonseasonal coal and gas prices, exogenously determined inelastic domestic demand for power, and an international cap-and-trade CO₂ scheme. Let us simplify the problem by ignoring fixed costs and assuming that the power stations offer power at the variable cost of fuel and allowances. Finally let us begin by assuming that coal is cheaper than gas on an efficiency adjusted basis.

The number of hours that the power price is set by the coal price is set by the minimum of (i) the number of hours for which the coal-fired capacity is sufficient to satisfy demand; (ii) the number of hours that the coal-fired generation can run under the SO_x cap; and (iii) the number of hours for which the coal price is less than the gas price.

Figure 2(a) shows this. First for a nonseasonal gas price, we can see that for the hours on the right, the CO₂ price affects the power price by the effect on coal-fired costs (roughly speaking 0.9 t CO₂/MWh), and for the hours on the left the CO₂ price affects the power price by the effect on gas-fired costs (roughly speaking 0.4 t CO₂/MWh). In the middle, the power price is raised as a result of the result on coal-fired hours, but this is mitigated by the lower carbon costs for gas-fired generation.

Figure 2(b) shows the impact of seasonality in gas prices. This can either increase or decrease the amount of gas-fired generation.

Here, we have taken the CO₂ price as exogenously determined and examined the effect of power prices. If we instead wish to view the effect of demand variation, demand elasticity, gas prices, coal prices, SO_x prices (or shadow prices), then we allow these to change and examine the effect on the volume of CO₂ allowances consumed, and thence construct the stack of amount of CO₂ allowances consumed in relation to CO₂ prices to find the price elasticity of CO₂ allowance demand.

Correlation and Cointegration

Correlation term structures—we have noted a number of features that reduce the tractability of correlation analysis. These are as follows:

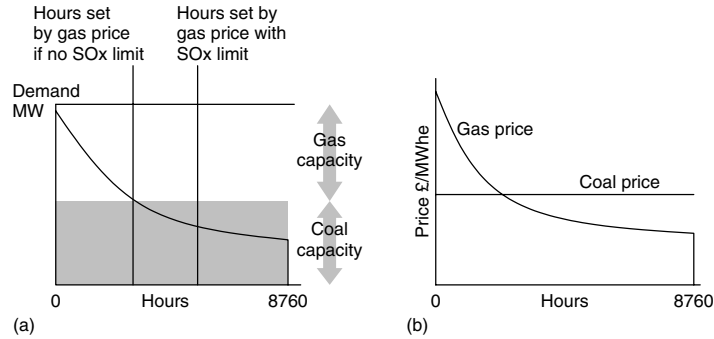


Figure 2 (a) Impact of restriction on coal-fired hours on price setting plant. (b) Effect of gas price seasonality on the number of hours for which gas price is cheaper than coal

1. Path dependence.
2. The “fundamental” long-term “value” of CO₂ allowances, associated with the uncertainty of this value and the difficulty of driving allowance prices to this value through adjustment of caps.
3. Complex co-integration relationships in the energy complex.
4. Nonstationarity of allowance price term structures.
5. Complex periodicity (daily, weekly, annual) in relation to the compliance periods.

While there are price or price return correlation measures that can be constructed, such as between commodities in the current or next compliance period, or between compliance periods, the nonstationarity and periodicity issues are such that it is very difficult

at this stage to apply this in a useful way for derivative pricing.

Other Kinds of Limits

When running a power station fleet or other physical complex, with multiple commodity inputs and environmental and amenity impact outputs, it is helpful to construct shadow prices to optimize the profitability of the operation. The theoretical approach is to use Lagrange multipliers, and the equivalent effect can be modeled by optimizing the operation subject to constraints, which may vary. For example, the combustion of a coal that is low in sulfur, and therefore SO_x, may entail high NO_x, particulates, carbon in ash, and ash. Figure 3 shows this in generic form, where the limit for factor X could be of various forms,

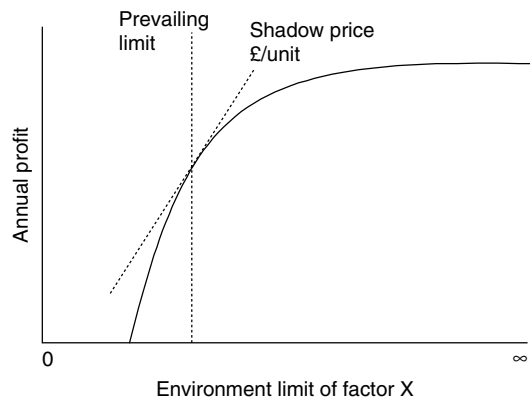


Figure 3 The shadow cost of complex environmental limits

such as height or temperature of water body, road traffic for bulk transport, or local air quality.

Conclusions

1. For small complexes we can limit the number of exogenous variables. For example in a country with a SOx cap-and-trade, we initially model all technologies and costs as fixed. We then model the SOx price change from a shock to the cap, or fuel prices, or both.
2. For a large complex such as the European Emission Trading Scheme under the Kyoto protocol, the same methods apply in principle, but the exogenous variables are more complex. In the early stages of the scheme, the cap is the exogenous variable. In the later stage it is the monetized marginal damage rate.

End Notes

^aThere are two particularly important considerations here. First that the ambient trading price of allowances that results from the cap is not exogenously determined but is closely related to the societal damage cost of emissions, and secondly that the interest rate used for the damage cost is driven not by the rolled-up money market account in the Arrow–Debreu market economy but the intergenerational interest rate in the overlapping generations economy.

^bFor an early description of the sulfur dioxide scheme in the USA, see [3].

^cThis is socially efficient and refer to [5].

^dAs mentioned in the beginning, we should be aware that the concept of a numeraire asset is less easily applied in classical economic analysis.

^eFor fitting of jump modeling to EEX data, see [2].

^fThis can be expressed in terms of convenience yield—see [1].

^gFor empirical fits using GARCH modelling, see [4].

References

- [1] Borak, S., Härdle, W., Trück, S. & Weron, R. (2006). *Convenience Yields for CO₂ Emission Allowance Futures Contracts*, SFN Discussion Paper 2006-076. University of Berlin.
- [2] Daskalakis, G., Pychoyios, D. & Markellos, R. (2006). *Modeling CO₂ Emission Allowance Prices and Derivatives: Evidence from the European Trading Scheme*, Athens University and Manchester Business School.
- [3] Joskow, P., Schmalensee, R. & Bailey, E. (1998). The market for sulfur dioxide emissions, *American Economic Review* **88**, 669–685.
- [4] Paolella, M.S. & Taschini, L. (2006). *An Econometric Analysis of Emission Trading Allowances*. Finrisk Working Paper 341.
- [5] Rubin, J. (1996). A model of intertemporal emissions trading, banking and borrowing, *Journal of Environmental Economics and Management* **31**, 269–286.
- [6] Weitzman, M. L. (1974). Prices vs. quantities, *The Review of Economic Studies* **41**(4), 477–491.

Further Reading

European Climate Exchange www.europeanclimateexchange.com, 2008.
 European Energy Exchange www.eex.com, 2008.
 International Emissions Trading Association www.ieta.org, 2008.
 Nordpool www.nordpool.com.
 Powernext www.powernext.fr, 2008.

CHRIS HARRIS

Weather Derivatives

Weather is not only an environmental issue but also a key economic factor, as recognized by the former US Commerce Secretary, William Daley, in 1998, when he stated that at least USD 1 trillion of the world economy is weather sensitive. The risk exposure is not homogeneous across the globe and some countries, usually those heavily dependent on agriculture, are more sensitive than others. It also includes a large range of phenomena such as modifications in temperature, wind, rainfall, and snowfall.

Weather risk has some specificities compared to other sources of economic risk: in particular, it is a local geographical risk that cannot be controlled. The impact of weather is also very predictable: the same causes will always lead to the same effects. Moreover, weather risk is often referred to as a *volumetric risk*, its potential impacts being mainly on the volume and not (at least directly) on the price. This explains why hedging of weather risk *via* the trading of commodities futures is difficult and imperfect. For example, oil futures price does not depend solely on demand (cold winter) and can be high even if the demand is low as in the case of a war, for instance. Volumetric risk is imprecisely compensated by the price variation in the futures position.

Usually, it is possible to hedge against risk by contracting some insurance policies. However, this is not really a possibility for weather risk for two main reasons: first, it is more a high-frequency, low-severity risk and, second, the same weather event can generate economic losses for some agents and some gains for others. To deal with this risk, some financial contracts depending on weather conditions (temperature, rainfall, snowfall, etc.) were created and introduced on the financial market 10 years ago. They are called *weather derivatives*.

Weather Derivatives

Weather derivatives are financial instruments whose value and/or cash flows depend on the occurrence of some meteorological events that are easily measurable, can be independently authenticated, and are sufficiently transparent to act as triggering underlyings for financial contracts. Typically, location is

clearly identified and measurement is provided by independent and reliable sources. The underlying meteorological events can be considered as noncatastrophic.

According to the Weather Risk Management Association (WRMA), the financial market related to weather has two main facets: the management of the financial consequences of adverse weather for those with natural exposure to weather, and the commercial trading of weather risk, both in its own right and in conjunction with a variety of commodities.

The first weather transactions took place in 1997 between Enron and Koch Industries. These transactions were the result of a long thinking process by Koch, Willis, and Enron aimed at finding a means of transferring the risk of adverse weather. These deals followed the deregulation of the energy market in the United States (transition from an oligopolistic position to a status of mere participant in a competitive market) and were based upon some temperature indices to compensate the energy producer in case of a mild winter.

The most common underlying is related to the notion of degree day, which is expressed as the difference between a reference level temperature (65°F or 18°C) and the average daily temperature T . The average is computed between the maximum and minimum recorded temperature over a particular day. A heating degree day (HDD) is defined as follows: $\text{HDD} = (65^{\circ}\text{F} - T)^+$, where $(.)^+ = \max(., 0)$. The bigger the HDD is, the lower the temperature is, and as a consequence the larger the demand for heating will be. Similarly, a cooling degree day (CDD) is defined as follows: $\text{CDD} = (T - 65^{\circ}\text{F})^+$.

The definition of a temperature index in those terms reflects the close relationship between the energy sector and the weather derivative market. Daily results are often cumulated to give a total on a given period, such as a week (Xmas–New Year's Eve, a sport event), a month (sales period, harvest period), or a quarter (summer holidays, opening season of a resort).

The Weather Derivatives Market

Most of the transactions are tailor-made and part of the OTC (over the counter) market. Usually, OTC transactions are realized within the ISDA standards (master agreement standards of the International Swap and Derivatives Association), which

provide standardized contracts aimed at easing OTC transaction processes. Some transactions go through specialized brokers but most of them are done without any intermediaries. It is not surprising that these tailor-made structures better suit the management of weather risk and the needs of the various players on this market. The organized markets are however rather successful, mainly because of the transparency, liquidity, and security they offer. Among them, the most predominant one is the Chicago Mercantile Exchange (CME), which offers several types of contracts.

Some attempts were made in Europe to launch an independent organized market for weather derivatives. In November 2005, Powernext, which is a European energy exchange based in France, and Meteo-France launched the quotation of national temperature indices for nine European countries and different types of indices. This initiative was further developed with the launch in June 2007 of Metnext, a joint venture between Meteo-France and Euronext, specialized in indices for weather risk management.

Besides organized markets, since the beginning of the weather market in the late 1990s, electronic trading platforms have always played an important role in the development of the market, especially Enron's platform, in the early days. Many big market players have such a platform, and, in particular, Spectron is rather important in Europe.

Microinsurance and Weather Derivatives

One of the most spectacular developments of weather derivatives lies, however, in the specificities of this widespread risk, with among the most exposed ones being those that rely on agriculture. Microinsurance, that is, a microfinance service that allows poor individuals to insure possessions, such as livestock or crops, has already seen the fantastic potentiality offered by weather derivatives. Therefore, it seems natural to see the World Bank as one of the players in this market. Among the various structures of the World Bank, the International Force on Commodity Risk Management (CRM, [4]) has, as its main objective, to deal with the agricultural risk in developing countries, where agricultural risk is defined as "negative outcomes stemming from imperfectly predictable biological, climatic and price variables". The economic and social consequences of this risk being

so huge in some countries, it seems logical to seek some form of protection against this risk. However, even if insurance companies were to write insurance policies against this risk, they would be typically too expensive for small farmers, and the compensation would take too long to be effective, partly because of the claim-checking procedure. The CRM has developed several projects throughout the globe in order to deal with agricultural risk transfer using weather derivatives. Among these projects, the pilot program conducted in India between 2003 and 2006 has been particularly successful. Microinsurance is a growing sector and more projects have been implemented across the globe to protect small- and medium-size farmers from weather risk. Some reinsurance companies have been particularly active in these programs by helping local institutions to implement the insurance schemes.

Pricing of Weather Derivatives

Given the uncertainty and the flexibility in their accounting classification, and also their relative illiquidity, several pricing methods have been suggested for weather derivatives. They can be classified into three main categories: actuarial, financial, and economic. In the following, we briefly present these various approaches, focusing on the (forward) pricing rule of a weather derivative with a payoff H at a future time. Considering forward price allows us to simplify the problem in terms of interest rates and focus on the pricing rule itself.

The actuarial method (*see Actuarial Premium Principles*) uses the fair value corrected by some margin as pricing rule. More precisely, denoting by P the statistical probability measure used as prior probability measure, the (forward) price of the derivative can be obtained as $\pi(H) = E_P[H] + \lambda \sigma_P[H]$. Different authors have studied the impact of the choice of the probability measure on pricing (*see, e.g., [8, 9]*).

The financial methods often assume the weather derivative market to be complete and therefore use the risk-neutral pricing rule $\pi(H) = E_Q[H]$ where the expectation is taken under a unique martingale measure Q . A milder argument consists in assuming the absence of arbitrage opportunities only. In an incomplete market framework, there exist many different methods to price a contingent claim, without

creating any arbitrage opportunity. A rather standard approach involves utility maximization. Any individual wants to maximize the expected utility of his or her terminal wealth in this framework. The maximum price that he or she is ready to pay for the weather derivative is therefore the price such that he or she is indifferent, from his or her utility point of view, between buying it or not. For this reason, the price obtained by utility maximization techniques is called *indifference price* (many references exist on this subject, see, for example, [7] and **Arbitrage Strategy**).

Denoting by u the utility function of the agent we consider, and assuming that there are no interest rates (for the sake of simplicity), the indifference buyer price of H , $\pi^b(H)$ is determined as $E_P[u(W_0 + H - \pi^b(H))] = E_P[u(W_0)]$, where W_0 is the initial wealth (which may be random). The price $\pi^b(H)$, which theoretically depends on the initial wealth and on the utility function, is not (necessarily) the price at which the transaction will take place. This gives an upper bound to the price the agent is ready to pay for the payoff H . The agent will accept to buy the contract at any price below $\pi^b(H)$. In the characterization of the indifference price, there is no question on the volume of the transaction. The potential buyer has two options: either buying one contract or not buying it. There is another possible approach, which consists of determining the price of the contract such that agreeing to buy a small quantity has a neutral effect on the expected utility of the agent. This notion of fair price in that context was first introduced by Davis in [5, 6] and corresponds to the zero marginal rate of substitution price. More precisely, the fair price p is determined such that

$$\frac{\partial E_P[u(W_0 + \theta H - p)]}{\partial \theta} \Big|_{\theta=0} = 0 \quad (1)$$

Finally, the economic approach relies on a transaction price, which is an equilibrium price, either between the seller and the buyer only, or between the different players in the market. Note that a transaction will take place only if the indifference buyer price is higher than the indifference seller's price, which gives a lower bound to the price the seller is ready to accept for the contract. This is a necessary condition for a transaction, and more generally for a market equilibrium, which characterizes the situation where all agents in the market maximize their expected utility at the same time by exchanging their risks (such an equilibrium is also called *Pareto optimal*). Such

an equilibrium approach has been adopted by different authors, for example, [3, 8]. While the first authors look at how weather forecasts can influence the demand for weather derivatives, and hence their price, the latter consider the problem of pricing in an incomplete market, with a finite number of agents willing to exchange their weather risk exposure. The price of the contract is obtained as that of the Pareto-optimal equilibrium.

Design Issues

As shown by various failed attempts of weather risk securitization, in particular the failed issue of a weather bond by Enron in 1999 (see [1, 2] for a detailed study), the design of the new securities appears as an extremely important feature in the transaction. It may be the difference between success and failure. More precisely, as previously mentioned, the high level of illiquidity, deriving partly from the fact that the underlying asset is not traded on financial markets, makes these new products difficult to evaluate and to use. The characterization of their price is very interesting as it questions the logic of these contracts itself. Moreover, the determination of the contract structure is a problem in itself: on one hand, the underlying market related to these risks is extremely illiquid, but, on the other hand, the logic of these products itself is closer to that of an insurance policy. Consequently, the question of the product design, which is unusual in finance, is raised.

Acknowledgments

Author Olivier Scaillet thanks the Swiss NSF for financial support through the NCCR Finrisk.

References

- [1] Barrieu, P. & El Karoui, N. (2002). Reinsuring climatic risk using optimally designed weather bonds, *Geneva Papers Risk and Insurance Theory* **27**, 87–113.
- [2] Barrieu, P. & El Karoui, N. (2005). Inf-convolution of risk measures and optimal risk transfer, *Finance and Stochastics* **9**, 269–298.
- [3] Campbell, S. & Diebold, F.X. (2005). Weather forecasting for weather derivatives, *Journal of the American Statistical Association* **100**, 6–16.
- [4] Commodity Risk Management Group (2005). *Progress report on Weather and Price Risk Management Work*,

- Agriculture and Rural Development*, The World Bank.
<http://www.itf-commrisk.org/>.
- [5] Davis, M. (1998). Option pricing in incomplete markets, in *Mathematics of Derivative Securities*, M.A.H. Dempster & S.R. Pliska, eds, Cambridge University Press.
- [6] Davis, M. (2001). Pricing weather derivatives by marginal value, *Quantitative Finance* **1**, 305–308.
- [7] Hodges, S. & Neuberger, A. (1989). Optimal replication of contingent claims under transaction costs, *Review of Futures Markets* **8**, 222–239.
- [8] Jewson, S. & Brix, A. (2005). *Weather Derivative Valuation: The Meteorological, Statistical, Financial and Mathematical Foundations*, Cambridge University Press.
- [9] Roustant, O., Laurent, J.-P., Bay, X. & Carraro, L. (2005). A bootstrap approach to the pricing of weather derivatives, (2004) *Bulletin Francais d'Actuariat*, Vol. **6**, n° 12, 153–171.

Further Reading

Horst, U. & Mueller, M. (2007). On the spanning property of risk bonds priced by equilibrium, *Mathematics of Operations Research*, **32**(4), 784–807.

Related Articles

Actuarial Premium Principles; Arbitrage Strategy; Catastrophe Bonds; Insurance Derivatives.

PAULINE BARRIEU & OLIVIER SCAILLET

Liquidity

An asset is relatively liquid if it can be traded without significant movement in its price. We typically think of an asset for which large quantities can be traded without substantial price movement as being in a deep and liquid market. In effect, a liquid asset can be traded at modest execution cost, even potentially in substantial size. Market events and turmoil in 2007 and 2008 focused attention upon the nature of illiquidity and have led to renewed interest about the meaning of liquidity as the trading activity in many hard-to-value assets declined significantly and the costs of trading such assets rose substantially.

The Kyle Framework

The nature of liquidity is extensively highlighted in the seminal work of Pete Kyle [11], who developed a model of adverse selection and trading cost in a classic analysis of the bid–ask spread with a risk-neutral informed investor and market-makers (*see Kyle Model*).^a In this setting when the investor is informed, he/she possesses a normally distributed signal of the value of the asset, and when uninformed, the investor trades a normally distributed exogenously specified quantity. Under the assumption of a linear equilibrium, Kyle [11] derived the unique solution in which the informed investor's trade is a linear function of his/her signal (optimizing an underlying quadratic decision problem) and the pricing at which a competitive market-maker trades is a linear function of the order flow (under the normality assumptions, this reflects a regression that predicts the expected informational cost conditional upon the order flow as the market-maker cannot distinguish an informed order from an uninformed one). The market-maker's (linear) pricing function is such that the cost of trading a position increases proportionately with its size because of the adverse informational consequences of facing a larger order given the decision rule of the informed agent.

The Kyle framework is consistent with the extent of adverse selection being influenced by current market conditions; in the Kyle context, these conditions are reflected in the structural parameters of the model such as the variability of the signal of the informed and the volatility of the trade of the uninformed. For

example, a spike in adverse selection above normal levels, as illustrated by an increase in the variability of the private signal of the informed, implies that liquidity will decline and the price impact of an order of a specified size will increase.^b

The “No-trade” Theorem and “Lemons”

Another complementary perspective on the nature of illiquidity in an asset market is illustrated by models in which trade need not arise due to informational frictions. For example, the classic “no-trade theorem” in Milgrom and Stokey [13] offers fundamental insights about the nature of illiquidity. The central theorem in [13] is that at a Pareto optimal allocation private signals do not lead to trade. One very simple interpretation that summarizes this conclusion is the “Groucho Marx joke”: Groucho Marx declared that he would not join any club whose standards were low enough such that he would qualify for membership.^c This intuition is closely related to the “lemons” idea that underlies the classic work of Akerlof [1]. In the Akerlof [1] setting, potential buyers of a used car are discouraged from purchasing it in light of the seller's private information. Since the potential seller has more precise information on the quality of the vehicle than the buyer, the seller will only sell those types of vehicles that are less valuable to it than the price that the buyer is willing to pay. The buyer's valuation is not based upon the actual quality of car that he is purchasing, but on the distribution of relatively low types that the seller is willing to allow the potential buyer to purchase. In the absence of a substantial difference in preference parameters and valuation between the potential buyer and seller, the presence of private information will lead to a breakdown in this marketplace and the absence of trading and liquidity. This provides a form of the “no-trade” theorem. Instead, if the preference differences were sufficient, then some trading (or perhaps even a lot of trading) would arise. If the equilibrium were to support a small volume of trading, then only the lowest quality of cars would be traded. In this context, illiquidity is reflected in low volume (quantity) and low quality arising in the marketplace, that is, most investors do not sell their holdings because they regard the market price as inadequate. For example, the ability of the seller to select specific low-value mortgage-backed securities pools for sale (such as in the context of substantial

credit risk those pools not carefully underwritten or which have serious documentation issues) and the extent of adverse selection would have greatly limited the trading of these instruments during portions of the market crisis.

The example that lies at the heart of Akerlof's analysis at least indirectly points to mechanisms by which liquidity can be sustained. If the qualities of the goods being sold were known (transparent) so that there was not private information, then liquidity across the marketplace would be sustainable. Alternatively, if the seller could warrant or guarantee quality to a specified level or alternatively, the buyer could rely upon the seller's market incentives to maintain a reputation for providing an accurate description of quality, then trading would be viable across the quality spectrum. Finally, if the difference in preference valuation between the sellers and buyers were sufficiently great, then these very large potential gains to trade would overcome the adverse selection for at least a significant proportion of the population and lead to trade. Relative to the "no-trade" theorem, trade and liquidity can arise, provided we start at an allocation that is not Pareto optimal.

Trading Volume and Asset Value

An interesting and important dimension of the liquidity of a market is the extent of buying and selling interest. If an asset is relatively idiosyncratic (such as an individual family residence with quirky characteristics), then there is one seller and likely few potential buyers. On the other hand (still within a residential real estate context), some properties may have quite close substitutes (units with similar lay-outs in a high rise condominium or cooperative building) and there could be many potential buyers and sellers. In many market contexts, the level of volume is often viewed as an indicator of liquidity and in the case of futures contracts the extent of open interest would be another such indicator. The connection between volume and prices explicitly arises in several of the key foundational frameworks addressing liquidity issues. For example, in the Kyle [11] framework, a liquid asset is the one in which a modest volume will not move the asset's price very much, so a large volume would be required to move the price substantially. Akerlof's [1] "lemons" analysis associates low trading activity with illiquidity as well as low quality of the goods

being exchanged and low prices. More generally, liquid assets typically have high trading volume and turnover.

Relatively liquid (low trading cost) assets are especially valuable and are the natural assets to utilize in a situation in which the owners anticipate high turnover. For example, liquid assets (such as futures contracts in some contexts) are often used by portfolio managers to adjust or maintain their aggregate exposures (to the direction of the equity market or interest rates), potentially even at relatively high frequencies. Less liquid assets are more appropriate for buy-and-hold exposures. To the extent that an investor owns assets with a range of liquidities, we would expect that the most liquid assets would be sold most often. While lower trading costs would enhance value across the spectrum of potential holding periods, the benefits would be especially pronounced for the potentially most active assets. These effects point to important clientele differences among assets of varying liquidity, which are analogous to clientele effects that differentiate assets that are taxed distinctly. However, just as for all investors (except tax-exempt ones) lightly taxed assets are more valuable than heavily taxed assets, similarly more liquid assets are more valuable than less liquid ones. One interesting theme in empirical asset pricing is that liquidity has emerged as a "priced" factor and that assets with higher trading costs provide higher expected returns without excluding the costs.^d Much of the focus in market microstructure is in assessing liquidity, informational affects, and trading costs.^e

Increasing liquidity enhances the value of assets not just across instruments (i.e., cross sectionally) but also over time (i.e., in time series) on specific assets as market conditions evolve. When liquidity in the marketplace disappears, the value of assets declines. This is illustrated by a range of market crises—including most recently, the subprime mortgage crisis that began in 2007.

Investors as Liquidity Providers

Investors themselves can possess considerable liquidity if they have funds available for investment without incurring substantial costs from selling existing positions. In that sense investors can transfer liquidity among assets and liquidity can move from investor to investor (via the assets) through the trading process.

Some investors even specialize in providing liquidity as illustrated by investors engaged in the traditional market-making function (such as a securities dealer or stock market specialist). Even a retail investor who places a limit order in a limit-order book should be viewed as offering liquidity to the marketplace (investors who hit the book consume liquidity). To a degree the dichotomy between demanders and suppliers of liquidity has been reflected in empirical analyses that focus on the interaction between the order flow and limit-order book.^f

The Self-fulfilling Nature of Liquidity

Liquidity is often viewed as self-fulfilling in the sense that liquidity attracts liquidity.^g For example, even in government bond markets there can be instruments with similar features that attract very different trading interest. The “benchmark” government bond in Japan (and to a degree in the United States) has had a higher price (and lower prospective return), but considerably greater trading interest, than other bonds that have similar characteristics.^h The theme that liquidity attracts liquidity is nicely illustrated by many aspects of the subprime mortgage crisis. For example, interest in trading collateralized debt obligation (CDO) structures has waned partially due to the drying up of liquidity in many such markets—purchasing such instruments became substantially riskier because of the price impact and uncertainty of liquidating positions.ⁱ The breakdown of the auction reset mechanism for municipal bonds is another example of a breakdown of liquidity due to a coordination failure—in this context seemingly rather creditworthy borrowers could not roll over their reset instruments because of doubts about whether the borrower would be able to repay current short-term borrowings due to concerns about the financial status of municipal bond insurers.^j The “seizing up” or “freezing” of interbank lending (“LIBOR”) markets (e.g., in September 2008) is yet another example of how the collapse of liquidity can reflect a coordination failure.

During the spring of 2008, the effective collapse of Bear Stearns was viewed by some as resulting from a liquidity crisis in which the firm’s short-term funding sources dried up because of potential concerns about the firm’s liquidity in the future, spurred at least in part by excess reliance upon such short-term funding sources.^k While a liquidity crisis can occur without

an obvious reason, these also can result in the asset management context from new or renewed concerns about the creditworthiness of a player. The lack of transparency of holdings and even market pricing can heighten concerns about adverse selection and can lead to fragility of an equilibrium in which short-term credit plays a prominent role.^l With the benefit of hindsight, the solvency challenges facing financial institutions in 2008 became more apparent.

The coordination issue also illustrates the difficulty in starting new futures contracts (e.g., the difficulties in attempting to establish the Shiller macroeconomic housing contracts). Often the potential success of instruments is difficult to forecast—which futures instruments will ultimately prove viable and attract sufficient liquidity can be difficult to discern. In this context, it may even make sense for those with a vested interest in the success and viability of the contract to consider subsidizing trading to build up open interest and volume, thereby mitigating concerns about adverse selection and the eventual viability of the contract.

Liquidity and Contract and Market Design

The specification of a contract as well as the market structure in which it trades can influence the underlying liquidity of that instrument. For example, indexed contracts will face less adverse selection than contracts on individual securities because there is much less scope for private information on an indexed instrument.^m The market structure in which an instrument trades also influences the liquidity of the instrument. For concreteness, consider the use of a “call” auction, which aggregates considerable trading activity temporally vs a continuous trading mechanism under which trades are dispersed over time. The continuous mechanism could provide less liquidity than the call auction.

However, not only does the market design influence liquidity, but also the magnitude of potential liquidity can influence the market’s design. A simple real estate example illustrates this point. In selling a unique house with many idiosyncratic features, the use of a broker is likely to be advantageous. Yet in selling a condominium for which there are many substitutes and suitable ways for the seller to reach out to prospective buyers, it need not be optimal to hire a broker. In effect, certain properties are so “attractive”

that they do not need a broker. That a property can potentially be advantageously sold without a broker adds to its *ex ante* value.ⁿ This highlights a potentially important selection difference between the properties sold by brokers and those sold directly; specifically, those being sold through brokers will often be those that are relatively more difficult to sell.

Concluding Comments: The Transmission of Illiquidity

Several major financial market crises also illustrate historically some interesting facets of illiquidity from an asset management perspective. For example, illiquidity can spread across asset categories. This occurred in the Long-Term Capital Management (LTCM) crisis in 1998—part of LTCM’s problems arose from a reluctance to sell instruments that were particularly illiquid because those had the largest losses (by definition). Consequently, the asset sales spilled to more liquid assets in order to reduce leverage. A similar phenomenon in the subprime mortgage crisis caused the spreads to substantially widen for prime jumbo loans and for that market to “freeze.” More broadly, a by-product of the subprime mortgage crisis was the transmission of liquidity problems across both markets and investors around the globe. The resulting contraction in equity capital in turn led to a “credit crunch” and indeed, various actions of the Federal Reserve Board have been an attempt to cushion the resulting economic shocks.

Another example of the interdependence in liquidity was the fallout in September 2008 from the collapse of Lehman Brothers. The subsequent suspension of liquidity by a major money market fund then led to a “run” on money market funds more broadly (temporarily shrinking by about half a trillion dollars)^o and the collapse of the commercial paper market, resulting in both new government guarantee programs to restore “confidence” and a “run” to use open bank credit lines. The various examples in this conclusion illustrate that liquidity is interconnected among assets and can be an aggregate rather than just an asset-specific phenomenon. In fact, historically, at the heart of various financial market breakdowns (including the Great Depression) have been liquidity crises and crises of confidence.

Acknowledgments

This paper builds upon a presentation prepared for a panel discussion at the Federal Deposit Insurance Corporation’s (FDIC’s) “7th Annual Bank Research Conference: Liquidity and Liquidity Risk” in Arlington, Virginia, in September 2007.

End Notes

^aGlosten and Milgrom [8] provide a second classic treatment of the bid–ask spread under private information and rational expectations.

^bWithin the Kyle [11] model, the structural parameters are constant. The text of the current article describes the comparative static associated with a change in model parameters.

^cThis interpretation of the theorem was originally pointed out by Milton Harris (see [13], p. 21, footnote 3).

^dFor example, these themes are illustrated by Pastor and Stambaugh [14] and Amihud and Mendelson [2].

^eAn interesting recent survey of market microstructure is Biais, Glosten and Spatt [3].

^fThe earliest such studies using high-frequency order data were Biais, Hillion and Spatt [4] and Harris and Hasbrouck [10]. A by-product of distinguishing between quote innovations on both the bid and ask sides of the market in the empirical analysis of Biais *et al.* [4] is that one can potentially distinguish between orders that are motivated by liquidity provision and those that reflect informational trading.

^gDow [7] examines a model with a mix of informed and uninformed traders in which there are multiple equilibria and a tight spread reduces the fraction of informed traders and strengthens the available liquidity.

^hThe nature of the “benchmark bond” in Japan is discussed in Boudoukh and Whitelaw [5].

ⁱChanges in perceived fundamentals also led to reduced interest in CDOs.

^jThis theme is illustrated by Mackenzie [12].

^kSomewhat related to this, even the use of short-term secured repurchase agreements to obtain capital became more restricted because of concerns about potential pricing of the underlying collateral in the event of collapse of Bear Stearns, which was a major securities firm, and concerns about the protection afforded repurchase agreements in the bankruptcy process.

^lCox [6] suggested that the collapse of Bear Stearns resulted from a lack of market confidence in the firm rather than insolvency or the absence of capital.

^mThis theme is explored by Subrahmanyam [15] and Gorton and Pennacchi [9].

ⁿPete Kyle offered an interesting interpretation of this point at the 2007 FDIC conference, noting that when properties are advertised as “one of a kind” that need not enhance their value, since “one-of-a-kind” properties would tend to require greater costs to resell the property efficiently.

^oTo avoid slightly analogous problems that would result from “runs” by their investors, some hedge funds began to utilize “liquidity gates” that limit overall outflows during a redemption period. However, such gates can actually heighten the tendency toward a “run” (e.g., to avoid being shut out).

References

- [1] Akerlof, G. (1970). The market for ‘Lemons’: Qualitative uncertainty and the market mechanism, *Quarterly Journal of Economics* **84**, 488–500.
- [2] Amihud, Y. & Mendelson, H. (1986). Asset pricing and the bid-ask spread, *Journal of Financial Economics* **17**, 223–249.
- [3] Biais, B., Glosten, L. & Spatt, C. (2005). Market microstructure: A survey of microfoundations, empirical results, and policy implications, *Journal of Financial Markets* **8**, 217–264.
- [4] Biais, B., Hillion, P. & Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the Paris Bourse, *Journal of Finance* **50**, 1655–1689.
- [5] Boudoukh, J. & Whitelaw, R. (1993). Liquidity as a choice variable: A lesson from the Japanese Government Bond Market, *Review of Financial Studies* **6**, 265–292.
- [6] Cox, C. (2008). *Chairman Cox Letter to Basel Committee in Support of New Guidance on Liquidity Management*. Securities and Exchange Commission, March 20.
- [7] Dow, J. (2004). Is liquidity self-fulfilling? *Journal of Business* **77**, 895–908.
- [8] Glosten, L. & Milgrom, P. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* **14**, 71–100.
- [9] Gorton, G. & Pennacchi, G. (1993). Security baskets and index-linked securities, *Journal of Business* **66**, 1–29.
- [10] Harris, L. & Hasbrouck, J. (1996). Market vs. Limit Orders: The SuperDOT evidence on order submission strategy, *Journal of Financial and Quantitative Analysis* **31**, 213–231.
- [11] Kyle, A. (1985). Continuous auctions and insider trading, *Econometrica* **53**, 1315–1335.
- [12] Mackenzie, M. (2008). US medical centre shuns muni bonds, *Financial Times* February 19, p. 25.
- [13] Milgrom, P. & Stokey, N. (1982). Information, trade and common knowledge, *Journal of Economic Theory* **26**, 17–27.
- [14] Pastor, L. & Stambaugh, R. (2003). Liquidity risk and expected stock returns, *Journal of Political Economy* **111**, 642–685.
- [15] Subrahmanyam, A. (1991). A theory of trading in stock index futures, *Review of Financial Studies* **4**, 17–51.

Related Articles

Bid-Ask Spreads; Kyle Model; Liquidity Premium.

CHESTER S. SPATT

Execution Costs

Execution costs are the difference in value between an ideal trade and what was actually done. The execution cost of a single, completed trade is typically the difference between the final average trade price, including commissions, fees, and all other costs, and a suitable *benchmark* price representing a hypothetical perfectly executed trade. The sign is taken so that positive cost represents loss of value: buying for a higher price or selling for a lower price. If a trade is not completed either for endogenous reasons (e.g., the price moves away from an acceptable level) or for exogenous reasons (a trader gets sick or a system fails), then some value must be assigned to the unexecuted shares. The cost of a portfolio transaction, or a series of transactions, is computed as a suitably weighted average of the costs of the individual executions.

Some of the costs of trading are direct and predictable, such as broker commissions, taxes, and exchange fees. Although these costs can be significant, they are commonly not included in the quantitative analysis of execution costs. “Indirect” costs include all other sources of price discrepancy, such as limited liquidity (market impact) and price motion due to volatility. These are much more difficult to characterize and measure and are much more amenable to improvement. Since the technology of cost measurement is most advanced for equity trading, we shall use that language; cost measurement for other asset classes is also very interesting, but less developed.

The benchmark is commonly taken to be the *arrival price*, that is, the quoted market price in effect at the time the order was released to the trading desk. Using that benchmark is equivalent to saying that a perfect trade, one with zero execution cost, would be one that executed instantaneously at the arrival price. The cost measured using the arrival price benchmark is called the *implementation shortfall*, a term introduced by Perold [10]: “on paper you transact instantly, costlessly, and in unlimited quantities . . . simply look at the current bid and ask, and consider the deal done at the average of the two.”

For example, suppose that an overnight investment decision assumed that a large quantity of stock could be purchased at the previous day’s closing price of \$50 per share; if the trade was fully completed at

an average price of \$50.25, then the execution cost would be reported as 25 cents per share. However, execution costs are only part of the picture: if the stock closed that day at \$51, then the trade would be successful despite its positive cost; a naïve cost model might assign \$1 profit to the portfolio manager and \$0.25 cost to the trader. Execution costs can be negative, for example, if the price dropped in the course of a purchase program and the asset was acquired for a lower price than that anticipated; or, in the above example, if the benchmark price were the day’s closing price. The forecast execution costs on any particular order typically have a very high degree of uncertainty, due to market volatility and other random effects.

A well-calibrated model for execution costs is an important part of the quantitative investment process. At a minimum, it is a tool for the portfolio manager to evaluate the performance of his or her trading desk and external brokers: were the results achieved on a particular execution compatible with the costs estimated from the pretrade model? Furthermore, anticipated transaction costs should be a component of the portfolio formation decisions: turnover should be minimized, and expected transition costs should be incorporated in the portfolio construction model along with expected alpha. Grinold and Kahn [6] discuss in depth the use of transaction cost models in investment management.

This article is divided into three parts, corresponding to the order in which three aspects should be addressed in designing an investment process, although the order, from the point of view of a single trade, is chronologically the reverse. First, we shall look at posttrade cost reporting, then at optimal trading to minimize execution costs, and finally at pretrade cost estimation.

Posttrade Reporting

“If you can not measure it, you can not improve it,” said Lord Kelvin. The first step toward reducing execution costs in any program is to measure them systematically. For each trade executed, the cost should be reported relative to a collection of benchmarks. In addition, the cost statistics should be computed across all trades in a suitable time period (daily or weekly) and broken down by any relevant parameters: primary market, size of trade, market capitalization of stock, and so on.

As noted above, the most common benchmark is the pretrade arrival price. Another common choice is the “volume-weighted average price” (VWAP), taken across the time interval during which the trade was executed. Although this most likely does not correspond directly to an investment goal, it is a popular benchmark for assessing the quality of the execution, because it largely filters out the effects of volatility. One would typically also use a posttrade price, for example, the closing price on the day during which the trade was executed. Typical posttrade reporting systems display the execution price relative to all of these benchmarks: before, during, and after trading.

When aggregating cost numbers across a diverse variety of trades, the individual cost numbers should be weighted so that the result is representative of the overall change in portfolio value. If the individual costs are measured in cents per share, then they should be weighted using the number of shares in each individual trade. If the individual costs are measured in basis points, then they should be weighted using the dollar value of each trade.

If a pretrade cost model has been developed, then the realized costs should also be compared with the forecast values, both on the level of individual execution, and on the overall portfolio level. This will help to identify trades that may have been badly executed as well as in maintaining accurate calibration of the pretrade model.

In addition to the average, it is useful to report the standard deviation of the costs. This is useful as a reality check for the significance of the mean. For example, if the sample standard deviation were 25 basis points on 100 independent trades, then the expected error in the sample mean would be 2.5 basis points and a change of one or two basis points in the mean cost would not be significant. The standard deviation should also be weighted by trade size; the most reasonable weights are the same that are used for the average cost. The standard deviation does not have good properties under subdivision.

Ideally, aggregate cost numbers would be reported so that the result is indifferent to subdivision of trade blocks, but this is often not possible. For example, suppose that 100 000 shares are purchased throughout the day; a natural benchmark price would be the day’s opening price. However, if this block were considered as 40 000 shares in the morning and 60 000 shares in the afternoon, the arrival price for the second block would be the midday price, and the overall reported

cost would be lower. The choice can be made only with knowledge of the investor’s overall goals. This difficulty does not arise with a VWAP benchmark or with a benchmark price at a fixed time such as at the close.

Additional difficulties in cost reporting come from price limits and incomplete trades. Suppose that a buy order is put in for a stock currently trading at \$50, but a condition is imposed that no shares are to be bought at a price higher than \$50.05. It will then be certain that the price impact on this trade will be at most 5 cents per share, but it may be that only a small fraction of the requested shares are actually executed. If the limit is below the initial price, the situation may be even more extreme. Similar difficulties arise if the trade is halted for other reasons.

The solution to this difficulty rests on the valuation of unexecuted shares, but there is no simple rule. The most straightforward would be to imagine that the unexecuted shares were purchased at the day’s closing price. In practice, this rule is far too stringent and does not take account of the variety of possible reasons the trade might have been halted. Trade cost reporting must take account of the entire investment process.

Optimal Trading

Once a system is in place for measuring and reporting trade costs, and after it has been agreed what criteria define a good trade or collection of trades, then one can design optimal strategies to meet investment goals. “Best execution” is not only advantageous to investors but also is required by regulation in most markets. The most important goal is always to reduce mean execution cost, which is largely due to market impact and uncaptured short-term alpha. One also often desires to reduce the standard deviation of trade costs in order to reduce the overall investment volatility [5].

Reducing costs is achieved by searching for liquidity in “space” as well as in time, accessing as many pools of potential liquidity and as many potential counterparties as possible for each piece of the trade. This includes routing to nonexchange trading venues such as “dark pools” and block crossing services when possible. Rapid changes in market structure make this increasingly complicated (see [7, Appendix] for a discussion of the US equities markets).

Accessing liquidity in time means being willing to slow trading to give potential counterparties the opportunity to appear in the market. If short-term price drift is not expected to be significant, then average trading costs can generally be reduced by trading more slowly.

Other aspects of trading push for rapid trading, most significantly anticipated price drift and market volatility: “To trade a list of stocks efficiently, investors must balance opportunity costs and execution price risk against market impact costs. Trading each stock quickly minimizes lost alpha and price uncertainty due to delay, but impatient trading incurs maximum impact. In contrast, trading more patiently over a longer period reduces market impact but incurs larger opportunity costs and short-term execution price risk” [1]. The actual trade schedule is determined by using a quantitative balance of all of these factors. This schedule may need to be dynamically adapted depending on the observed liquidity or other market components. For a portfolio, correlation between the assets should be included in addition to the volatility of each one, and the schedule may need to maintain strict neutrality to a market index or other risk factors.

In practice, optimal strategies for a mean–variance trade-off, and strategies that are calculated to optimize expected cost in the presence of short-term drift, are generally similar: at the beginning of each execution, rapid trading reduces the exposure to

volatility or alpha, then the trading slows to reduce overall impact costs (Figure 1, from [2]). The most important question is the overall speed of the execution: should it be completed in minutes, hours, or days? Regardless of the model, an approximate quantitative model for execution costs is an essential element of this decision.

Pretrade Cost Estimation

Once a measurement system has run long enough to generate a useful amount of data, one can begin to think about developing an analytic model. The goal of the model is to forecast the execution cost to be expected on any anticipated trade, along with an estimate of the uncertainty in the forecast. In its most general form, such a model must contain a full explanation of how trading in markets affects prices. This is a rich area of research with a large literature (see [9] for an extensive survey; see [4, 8] for more subtle models). Thus, one typically poses the problem in the more prosaic terms of estimating the future in terms of past data.

The goal is to give the predicted cost C (in cents per share or basis points) in terms of input variables, including, at a minimum, the number of shares traded X and the daily volume V of the stock (either historical average or volume on the day of trading), and the effective duration T over which the trade was executed as in Figure 1. To normalize across many different stocks, one should also include properties such as volatility σ , and, perhaps, market capitalization, primary exchange or country, and other economic parameters. In addition, one might include trade information such as the anticipated short-term alpha, and the algorithm or other means that were used to trade the stock.

For any proposed trade, the pretrade model predicts the cost on the basis of cost that was measured on “similar” trades in the past. “Supervised learning” algorithms [11] could, in theory, be used to carry out this classification. In practice, because the dimension of the parameter space is so large, there may be no, or very few, trades that are sufficiently similar to the proposed ones in all variables. This argues for a regression approach in which a specific functional form is proposed, calibrated to data, and evaluated using standard statistical tests of significance.

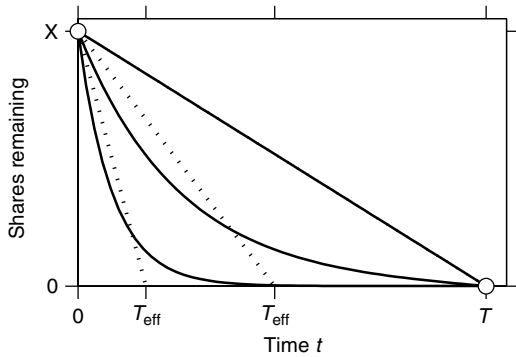


Figure 1 Optimal trading trajectories with varying degrees of urgency (solid curves). Regardless of the exact analytic form of the curve, the most important variable is the approximate effective duration T_{eff} corresponding to a linear approximation of the trajectory and represented here by a broken line

The most detailed of such a study was carried out by Almgren *et al.* [3]. This model is based on separation of the cost into a “permanent” and a “temporary” component. Although both components contribute to the realized cost on every trade, it is useful to calibrate each separately. The calibration relies on three observed prices: (i) S_{pre} denotes the arrival price, that is, a market price just before the order begins trading; typically, this would be the bid–ask midpoint before the first execution; (ii) S_{exec} is the actual average price for which the order finally executes, which is, of course, the quantity of most interest to the trader; and (iii) S_{post} denotes a market price just after the order has finished executing, possibly with a short time lag to allow transient effects to dissipate.

The model of Almgren *et al.* [3] measures two quantities I and J , for permanent and temporary impact, respectively. For a buy order, these are

$$I = \log \frac{S_{\text{post}}}{S_{\text{pre}}} \approx \frac{S_{\text{post}} - S_{\text{pre}}}{S_{\text{pre}}} \quad (1)$$

$$J = \log \frac{S_{\text{exec}}}{S_{\text{pre}}} \approx \frac{S_{\text{exec}} - S_{\text{pre}}}{S_{\text{pre}}} \quad (2)$$

and for a sell order, the signs would be reversed. The approximate equalities are valid for moderate-sized orders, for which the price does not move more than a few percent during execution. These quantities represent price changes relative to the pretrade price, as fractions of that benchmark.

The permanent component of cost is interpreted as the net displacement of the market due to the buy–sell imbalance introduced by this particular trade. The simplest model sets

$$I = \gamma \sigma \frac{X}{V} + \langle \text{noise} \rangle \quad (3)$$

In this expression, we normalize the trade size X by the day’s volume V , so X/V is the trade size as a fraction of a typical day’s flow. We also normalize the impact I by the volatility σ , so that we represent the impact as a fraction of a typical amount that the stock moves without our trading. The coefficient γ is taken to be constant across all stocks, with widely varying daily volumes and volatilities. The noise term arises from intraday volatility: the actual cost experienced on any particular trade may be very different from the mean prediction from the model.

In this form of the model, the permanent impact is linear in the total number of shares traded. This model is particularly convenient for theoretical modeling but must be justified by empirical analysis of real-trade data. Almgren *et al.* [3] found a reasonable agreement with their data, but other recent empirical studies have suggested that the linear form might not be justified.

The temporary impact component of cost is interpreted as the additional premium that must be paid for execution in a finite time, above a suitably prorated fraction of the permanent cost. We model it as

$$J = \frac{I}{2} + \eta \sigma \left(\frac{X}{VT} \right)^\beta + \langle \text{noise} \rangle \quad (4)$$

where T represents the effective duration of the trade as a fraction of the trading day. Thus, X/VT represents the “participation rate” or the fraction of market flow that this trade constitutes during the time that it is active. The factor $1/2$ on the permanent cost component represents the fraction of posttrade impact that is paid on the trade itself. As for the permanent cost, the impact cost is expressed as a fraction of typical volatility; the coefficient η is taken constant across all stocks. Other factors such as bid–ask spread and market capitalization were not determined to be significant in US equity markets.

Almgren *et al.* [3] calibrate this model to a large sample of the US equity trades using a two-step procedure. First, the permanent impact term I is calibrated, testing the hypothesis of linear impact and estimating the value of γ . Next, the temporary impact term J is calibrated, determining values for the exponent β and the coefficient η . A key aspect of the verification is characterization of the error terms to verify that volatility is an adequate explanation for the residuals. The result has $\beta \approx 0.6$, which is roughly compatible with the earlier square-root models [12], as well as values for the coefficients γ and η . For trades that are a few percent of daily volume executed across several hours, the predicted impact costs are tens of basis points.

A number of factors make such models at best very approximate. The first is the extremely low values of R^2 in the regression, typically around a few percent at best. That is, the ability of the model to predict the cost of any single trade is very poor, because market volatility due to other trading activity is usually very large. The second, is

the difficulty of the model in distinguishing market impact from alpha: did the price move up because the buy program impacted the price or was the trade executed because the manager correctly anticipated a price rise? A third difficulty is difference in behavior between small and large trades: although the model claims to be universally valid, in practice, a model developed for small trades gives poor results on large trades.

In any particular application, the model should be critically evaluated by the user and recalibrated and extended as necessary. Despite its intrinsic difficulties and limitations, this model and extensions are extremely useful to give approximate anticipated cost values for pretrade planning, optimal trade scheduling, and posttrade evaluation.

References

- [1] Alford, A., Jones, R. & Lim, T. (2003). Equity portfolio management, in *Modern Investment Management: An Equilibrium Approach*, B., Litterman, ed, Wiley.
- [2] Almgren, R. & Chriss, N. (2000). Optimal execution of portfolio transactions, *Journal of Risk* **3**(2), 5–39.
- [3] Almgren, R., Thum, C., Hauptmann, E. & Li, H. (2005). Equity market impact, *Risk* **18**, 57–62.
- [4] Bouchaud, J.-P., Gefen, Y., Potters, M. & Wyart, M. (2004). Fluctuations and response in financial markets: the subtle nature of ‘random’ price changes, *Quantitative Finance* **4**(2), 176–190.
- [5] Engle, R. & Ferstenberg, R. (2007). Execution risk: it’s the same as investment risk, *Journal of Portfolio Management* **33**(2), 34–44.
- [6] Grinold, R.C. & Kahn, R.N. (1999). *Active Portfolio Management*, 2nd Edition, McGraw-Hill.
- [7] Hasbrouck, J. (2007). *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*, Oxford University Press.
- [8] Lillo, F., Farmer, J.D. & Mantegna, R.N. (2003). Master curve for price-impact function, *Nature* **421**, 129–130.
- [9] Madhavan, A. (2000). Market microstructure: A survey, *Journal of Financial Markets* **3**, 205–258.
- [10] Perold, A.F. (1988). The implementation shortfall: paper versus reality, *Journal of Portfolio Management* **14**(3), 4–9.
- [11] Poggio, T. & Smale, S. (2003). The mathematics of learning: dealing with data, *Notices of the American Mathematical Society* **50**(5), 537–544.
- [12] Torre, N. (1997). *Market Impact Model Handbook*, Barra, Berkeley.

ROBERT ALMGREN

Bid-Ask Spreads

The literature on asset pricing often assumes an ideal world where security prices are set by market participants in *frictionless* financial markets. However, a variety of market frictions, including trading costs and constraints on short selling, exist in actual markets. It is important for market participants to accurately estimate and incorporate the impact of trading costs. For portfolio managers and investors, implementing investment decisions is costly and will typically lead to a shortfall in investment performance [26] relative to that theoretically attainable in frictionless markets. Decisions need to be conditioned not only on the fundamental soundness of potential investments but also on the anticipated costs of implementing the required trades. Estimates of trading costs also serve as a measure of market quality, allowing policy makers to assess the impact of regulatory reforms and exchange officials to assess the effects of trading rule and market structure changes, and informing corporate managers' decisions on where to list their shares. This article provides an overview of several issues related to measuring trading costs in financial markets. It focuses on three related measures of trading costs: quoted spreads, effective spreads, and realized spreads.

The Nature of Trading Costs

A fundamental issue in trading is the asynchronous arrival of buyers and sellers [9]. This creates uncertainty as to the amount of time that will be required to locate a counterparty, and regarding the market price that will prevail at the time a trading partner is located. This uncertainty can be mitigated by the continual presence of "liquidity suppliers", who stand ready to serve as counterparties, thereby providing immediacy of trade execution, that is, "liquidity". Liquidity suppliers often take the form of designated market makers or dealers, but liquidity can also be provided by traders in the form of limit orders.

Liquidity providers need to be compensated for the costs involved. Besides order processing costs, dealers incur inventory holding and adverse selection costs. Accommodating investors' order flows generally leaves dealers holding inventory positions that are not optimal in terms of the diversification of risk

[16] (see **Inventory Effects**). Further, dealers need to be compensated for the possibility that some buy or sell orders can originate with traders possessing superior information regarding security value. Dealers, on average, lose money on transactions with better-informed traders [13, 22] (see **Glosten-Milgrom Models; Kyle Model**), who sell to dealers ahead of price declines and buy from dealers ahead of price increases.

Dealers recover these costs by purchasing at a lower "bid" price, while selling at a higher "ask" (or "offer") price. The bid and ask prices are the dealer's quoted prices or "quotes", and the difference between the two is the bid-ask or quoted spread, a measure of trading cost. The corresponding quantities offered by the dealers at the quoted prices are referred to as the *quoted depth*, that is, the *bid depth* and the *ask depth*.

In most financial markets, including the New York Stock Exchange (NYSE) and the Nasdaq Stock Market, liquidity provision by designated dealers is augmented by standing limit orders submitted by public traders. A limit order to buy sets a maximum price to be paid, while a limit order to sell sets a minimum price that will be accepted. A limit order may be viewed as a one-sided quote. Some markets (e.g., The Hong Kong Stock Exchange) operate without designated dealers, in which case, the bid-ask spread is determined by the most aggressively priced unexecuted limit orders. In summary, the bid-ask spread, which measures trading costs for investors buying at the ask and selling at the bid, arises to compensate liquidity providers for order processing, inventory, and adverse selection costs.

Measures of Trading Cost

To estimate trading costs, empirical methodologies rely on the simple intuition that transactions would occur at the true underlying security value in the absence of trading costs. Hence, the deviation between transaction price and an estimate of the true underlying security value is an estimate of trading cost. For buyer (seller) initiated trades, the traded price is expected to be higher (lower) than the true security value, the difference being the estimate of trading cost.

The Quoted Spread

The simplest measure of trading cost is the quoted spread (QS), which is defined as the difference between the bid and ask prices. The quoted spread measures the cost of completing a round trip (buy and sell), if trades are executed at the quoted prices. Execution costs for a single trade are often measured as half the spread, described on a percentage basis by equation (1):

$$\text{Quoted half spread} = QS_{it} = 100 \times \frac{(A_{it} - B_{it})}{(2 \times M_{it})} \quad (1)$$

where A_{it} and B_{it} are the posted ask price and bid price for security i at time t , respectively, and M_{it} , the quote midpoint or mean of A_{it} and B_{it} , is a proxy for the true underlying security value.

The Effective Spread

In many dealer markets, including those that trade fixed-income securities and foreign exchange, the quoted prices are simply a starting point for negotiations between customers and dealers, and transactions frequently occur at prices other than the quotes. Also, in some markets, including those relying on trading floors, there may be *latent liquidity* not reflected in the quotes. On the NYSE, for example, market orders may execute at prices within the quotes when the specialist (the NYSE's designated dealer) or a floor broker elects to improve on the quote [28, 30] (see **Specialist Markets**). Many electronic exchanges allow traders to hide some or all of the order size, implying that limit orders offering more attractive prices than the quotes may exist on the book [4] (see **Order Types**). Further, quoted prices pertain only to the quoted depth; large orders might exhaust the depth at the quote and "walk up the book", executing against limit orders with less attractive prices and leading to a weighted-average trade price outside the quotes.

When trades occur either within or outside the quotes, a better measure of trading costs is the percentage effective half spread, which is based on the actual trade price, and is computed on a percentage basis as described in equation (2):

$$\text{Effective half spread} = ES_{it} = 100 \times D_{it} \times \frac{(P_{it} - V_{it})}{V_{it}} \quad (2)$$

where P_{it} is the transaction price for security i at time t , D_{it} is an indicator variable that equals 1 for customer buy orders and -1 for customer sell orders, and V_{it} is an observable proxy for the true underlying value of security i at time t . The effective spread is based on the deviation between the execution price and the true underlying value of the security, and can be viewed as an estimate of the execution cost actually paid by the trader and the gross revenue earned by the liquidity provider.

It is also possible to distinguish between the noninformational (inventory and order processing) and informational (adverse selection) components of trading costs, on the basis of the behavior of prices *subsequent* to a transaction. The seminal paper using this approach is [21]. The intuition is that noninformational transaction costs should result only in a *temporary* deviation of price from value, evidenced by a price reversal after the trade. In other words, while customer purchases (sales) should occur at prices above (below) pretrade value, we should subsequently observe a partial reversal of the price change. The price reversal is partial rather than full because the informational component of trading costs is, on average, associated with a *permanent* increase (decrease) in security value after buys (sells). The informational component can be measured by the change in the estimate of security value, while the noninformational component can be measured by the reversal from trade price to posttrade value.

Estimating the Informational Component: The Price Impact of Trades

The possible presence of informed traders is revealed to liquidity providers in a noisy manner by the order flow imbalance, that is, the difference between quantities of buy *versus* sell orders, which will tend to be positive when the security is undervalued and negative when the security is overvalued. Market makers incorporate the information in order flow imbalances by adjusting quotes upward (downward) after buy (sell) orders. These price adjustments reflect both the proportion of the informed traders *versus* liquidity traders in the market, and the extent of superior information about security value held by the informed traders. The private information contained in trades, or equivalently the amount of adverse selection cost incurred by the liquidity provider, can

be estimated using equation (3):

$$\text{Price impact of trade} = PI_{it} = 100 \times D_{it} \times \frac{(V_{it+n} - V_{it})}{V_{it}} \quad (3)$$

where V_{it+n} denotes the security's true underlying value n periods after the transaction. The price adjustment from V_{it} to V_{it+n} reflects the markets' assessment of the private information conveyed by the trade [3, 17] (see **Price Impact**). Research has documented that markets are particularly sensitive to order flow imbalances ahead of anticipated news disclosures, such as earnings announcements by corporations [23], as price impacts are larger than on nonannouncement days. The price impact of trades is an often used empirical proxy in the corporate finance literature for the degree of information asymmetry regarding security value across traders.

Estimating the Noninformational Component: Realized Spreads

The presence of informed traders will cause market prices, on average, to rise after customer buys and to fall after customer sells. Owing to these adverse price movements, market makers earn less than the effective spreads for their services. Market making revenue net of losses to better-informed traders can be measured by the *reversal* from the trade price to the posttrade value. The realized spread captures the extent of reversal, computed using equation (4):

$$\begin{aligned} \text{Realized spread} = RS_{it} &= 100 \times D_{it} \times \frac{(P_{it} - V_{it+n})}{V_{it}} \\ &= \text{Effective Spread}_{it} - \text{Price Impact}_{it} \end{aligned} \quad (4)$$

As noted above, some studies refer to price impact and realized spreads as the “permanent” and “temporary” price impacts of a trade [21, 25]. Some authors [29] have argued that trading costs and market quality are better measured by temporary price impacts (realized spreads) than by total price impacts (effective spreads).

Implementation Issues

To measure effective or realized spreads, researchers need to identify (i) whether the trade was initiated

by a buyer or a seller, and (ii) estimates of security value before and after the trade. As documented in recent research, the methodological choices regarding these issues are nontrivial and can significantly affect estimates of trade execution costs [27, 32].

Algorithms for Assigning Trade Direction

Some publicly available databases from international markets, for example, Euronext-Paris and the Toronto Stock Exchange, as well as some proprietary datasets on institutional trading, for example, those provided by consulting firms Abel/Noser or Plexus, contain information on the buy and sell orders submitted to the markets. In contrast, databases such as Trade and Quote (TAQ) (released by the NYSE) and Nastraq (released by Nasdaq), which are widely studied owing to their comprehensive inclusion of all trade and quote data for all listed US stocks, do not provide information on underlying buy and sell orders. As a consequence, the direction of the trade (i.e., whether the trade was initiated by a buyer or a seller) must be imperfectly inferred from the available data (see [2] for a detailed discussion).

In datasets where order level data is not available, the most widely used algorithm for assigning trade direction is that recommended by Lee and Ready [24]. Their algorithm assigns trades completed at prices above (below) the prevailing quote midpoint as buyer (seller) initiated trades. Trades executed at the quote midpoint are classified on the basis of a “tick test”. The tick test assigns a trade as a buy (sell) if the trade executes at a higher (lower) price as compared to the most recent trade at a different price. An alternative algorithm is proposed by Ellis, Michaely and O'Hara [11], who assign trades executed at the ask (bid) quote as buyer (seller) initiated, while using the tick test for all other trades. On the basis of proprietary order level data, prior research finds that the algorithm proposed by Lee and Ready [24] works fairly well, classifying about 85% of the trades correctly.^a Error rates are slightly lower for the algorithm proposed by Ellis, Michaely and O'Hara [11] (see [2] and [12]).

Research based on data from the early 1990s [14] finds that trade report times lagged actual trade times for NYSE stocks. As a consequence, studies such as [24] recommend adjusting the time stamps by 5 s when comparing trades and quotes. However, for recent data, Bessembinder [2] and Ellis, Michaely and

O'Hara [11] report that the allowance for reporting lags is not necessary. They recommend that trades are best compared to contemporaneous quotations for both NYSE and Nasdaq stocks when assigning trades as buyer or seller initiated.

Pretrade Benchmark Price

Pretrade price impact refers to implicit trading costs that may be incurred if prices move systematically away from the trader (rising before buys or dropping before sells) between the time of a trade decision and complete execution of the order [26]. Pretrade price impact will tend to arise when larger orders are broken into smaller orders and executed successively. Prices may also move away from a trader because his trading intentions are detected by market participants who then "front run" the order or engage in "predatory trading" [7], or simply infer information from the existence of the trading interest. In addition, prices will tend to move away from traders relying on momentum strategies. Executing a trading program too slowly exposes traders to the risk of larger pretrade price impacts. On the other hand, executing orders that are larger than quote sizes too quickly may result in larger effective spreads. The skill of a trader handling larger orders lies in balancing these effects. To capture the impact of trader's "timing" and "liquidity" decisions, Perold [26] recommends that the average of the prevailing bid and ask quote at the time of the trading decision be used as the pretrade benchmark price.

However, data on trade decision times is often not available, except in specialized proprietary datasets such as that studied by Conrad, Johnson and Wahal [8]. Studies using publicly available datasets, including [17] and [3] use the quote midpoint at the time of the trade as the pretrade benchmark price. Although adverse drift in prices ahead of trade executions is most obviously an issue for larger traders, recent evidence [2, 27, 32] indicates that prices move systematically and adversely in the seconds before even small trades are executed. Bessembinder [2] recommends that researchers use quotation midpoints in effect 5 s prior to the trade report time as a proxy for the true underlying price (V_{it}) when measuring effective spreads and price impact of trades.

Posttrade Benchmark Price

The posttrade benchmark price (V_{it+n}) should be measured when the market has had sufficient time to incorporate the information contained in the trade [17]. If the period after the trade time is too short, temporary price effects may still dominate, or alternately, the market may not have had a chance to assess the trade's likely information content. If the period is too long, the measure will become unnecessarily noisy because of the arrival of extraneous information. In the absence of theoretical guidance, studies have used different proxies for the posttrade benchmark price. Studies using institutional data have often used the closing price on the day of trade as the posttrade price [20]. The practitioner literature commonly relies on the volume weighted average price (VWAP) on the day of the trade. Among datasets with broad coverage, Huang and Stoll [17] use the first trade price both 5 and 30 min after the trade, while Bessembinder and Kaufman [3] use the quote midpoint 30 min and 24 h after the trade. Bessembinder [2] used the midquote in effect 30 min after the time of the reference quote, or the 4 p.m.-quotation for trades completed during the last half hour of trading. Werner [32] reports that realized spread measures obtained in large samples are relatively insensitive to the choice of the posttrade benchmark price.

Since September 2001, the Securities and Exchange Commission (SEC) has required each US stock "market center" to compile and disseminate on a monthly basis various standardized measures of execution quality in nearly all publicly traded securities [6]. The intent of SEC Rule 605 (formerly 11Ac1-5) is to provide traders with information on execution quality at different market centers. The execution quality measures that each market now reports include round-trip effective spreads, realized spreads, as well as average execution speed. The regulation provides something of an official validation of the effective spread and realized spread measures described above, and also codifies a specific methodology to estimate trading costs based on the order level data available to each market center. Specifically, the effective spread compares the traded price to the quote midpoint (as a proxy for V_{it}) at order arrival, while the realized spread is based on the quote midpoint (as a proxy for V_{it+n}) 5 min after the trade.

Evidence on Trading Costs

Jones [19] provides a detailed perspective on trading cost in US markets over the last century. He estimates that quoted spreads on Dow Jones stocks were in the range of 0.60% for sustained periods until the beginning of 1980s, with spikes in spreads observed during market downturns, such as the Great Depression. He documents that during the last two decades trading costs have fallen dramatically to around 0.20% for Dow stocks, partly facilitated by regulation but mainly by increased competition among market centers with the onset of electronic trading systems (see [15] for recent evidence).

Several recent studies have examined execution quality for comparable firms on the two major US market centers—the NYSE and the Nasdaq. In 2001, subsequent to regulatory changes, Bessembinder [1] reports that effective spreads are similar for comparable firms in the two markets. The most recent evidence available appears to be that reported Boehmer [6], using Rule 605 data over November 2001 to December 2003; he reports an average effective spread for NYSE stocks of 6.2 cents per share, *versus* 8.8 cents for comparable Nasdaq stocks.

Several studies have examined execution quality in market outside the United States. For example, Venkataraman [31] reports that effective spread for large firms traded on the Paris Bourse were 0.25% in 1997, relative to 0.21% for comparable firms on NYSE. Jain [18] reports on trading costs during the year 2000 for liquid firms on 51 stock exchanges around the world. The average effective (realized) spread across exchanges is 2.13% (2.15%). However, there is considerable variation in effective (realized) spreads across markets, ranging from 0.10% (0.25%) at the NYSE (Luxembourg) to 14.47% (14.6%) in Ukraine. Research indicates that differences in execution quality across markets are related to both exchange-design features, such as tick size and order handling rules, and the regulatory environment, such as the enforcement of insider trading laws and the protection of shareholder rights [5, 10].

Conclusions

Trade execution costs, which reflect the price concessions necessary to complete transactions quickly, are

important indicators of market quality and important determinants of traders' actual investment results. Execution costs arise because it is costly to provide liquidity, including order processing costs, inventory holding costs, and losses suffered to better-informed traders. Trading costs can be estimated on the basis of comparisons of trade prices with proxies for underlying security value, the most common proxy being quote midpoints. Comparisons can be of trade prices to midpoints at or before the time of the trade, as in effective spread measures, or to midpoints after the trade, as in realized spread measures. Further, the amount of asymmetric information present in a market can be estimated by assessing trades' price impact, measured as the difference between post- and pretrade estimates of security value. Recent research indicates that trade execution costs have declined steadily in US markets in recent years, particularly subsequent to decimalization in 2001, and documents substantial variation in average trading costs across international equity markets.

Acknowledgments

We would like to thank Charles Jones for his comments.

End Notes

^a. However, the accuracy of the algorithms in postdecimalization data, which is characterized by substantial increases in the number of quote revisions and trades, has, to our knowledge, not been assessed.

References

- [1] Bessembinder, H. (2003). Trade execution costs and market quality after decimalization, *Journal of Financial and Quantitative Analysis* **38**(4), 747–778.
- [2] Bessembinder, H. (2003). Issues in assessing trade execution costs, *Journal of Financial Markets* **3**, 233–257.
- [3] Bessembinder, H. & Kaufman, H.M. (1997). A comparison of trade execution costs for NYSE and NASDAQ-listed stocks, *Journal of Financial and Quantitative Analysis* **32**, 287–310.
- [4] Bessembinder, H., Panayides, M. & Venkataraman, K. (2009). Hidden liquidity: an analysis of order exposure strategies in electronic stock markets, *Journal of Financial Economics*, forthcoming.
- [5] Bhattacharya, U. & Daouk, H. (2002). The world price of insider trading, *Journal of Finance* **57**, 75–108.

- [6] Boehmer, E. (2005). Dimensions of execution quality: recent evidence for U.S. equity markets, *Journal of Financial Economics* **78**, 463–704.
- [7] Brunnermeier, M. & Pedersen, L. (2005). Predatory trading, *Journal of Finance* **60**(4), 1825–1863.
- [8] Conrad, J., Johnson, K. & Wahal, S. (2003). Institutional trading and alternative trading systems, *Journal of Financial Economics* **70**, 99–134.
- [9] Demsetz, H. (1968). The cost of transacting, *Quarterly Journal of Economics* **82**, 33–53.
- [10] Eleswarapu, V. & Venkataraman, K. (2006). The impact of legal and political institutions on equity trading costs: a cross-country analysis, *Review of Financial Studies* **19**(3), 1081–1111.
- [11] Ellis, K., Michaely, R. & O'Hara, M. (2000). The accuracy of trade classification rules: evidence from Nasdaq, *Journal of Financial and Quantitative Analysis* **35**, 529–551.
- [12] Finucane, T. (2000). A direct test of methods for inferring trade direction from intra-day data, *Journal of Financial and Quantitative Analysis* **35**, 553–576.
- [13] Glosten, L. & Milgrom, P. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* **14**, 71–100.
- [14] Hasbrouck, J., Sofianos, G. & Sosebee, D. (1993). *New York Stock Exchange Systems and Trading Procedures*, NYSE Working Paper, 93–01.
- [15] Hendershott, T., Jones, C. & Menkveld, A. (2007). *Does Algorithmic Trading Improve Liquidity?* Working Paper, University of California, Berkeley.
- [16] Ho, T. & Stoll, H. (1983). On dealer markets under competition, *Journal of Finance* **35**, 259–267.
- [17] Huang, R. & Stoll, H. (1996). Dealer versus auction markets: a paired comparison of execution costs on NASDAQ and NYSE, *Journal of Financial Economics* **41**, 313–357.
- [18] Jain, P. (2002). *Institutional Design and Liquidity at Stock Exchanges Around the World*, Working Paper, University of Memphis.
- [19] Jones, C. (2002). *A Century of Stock Market Liquidity and Trading Costs*, Working Paper, Columbia University.
- [20] Keim, D. & Madhavan, A. (1996). The upstairs market for large block transactions: analysis and measurement of price effects, *Review of Financial Studies* **9**, 1–36.
- [21] Kraus, A. & Stoll, H. (1972). Price impacts of block trading on the New York Stock Exchange, *Journal of Finance* **27**, 569–588.
- [22] Kyle, A. (1985). Continuous auctions and insider trading, *Econometrica* **53**, 13–32.
- [23] Lee, C., Mucklow, B. & Ready, M.J. (1993). Spreads, depths, and the impact of earnings information: an intraday analysis, *Review of Financial Studies* **6**, 345–374.
- [24] Lee, C. & Ready, M.J. (1991). Inferring trade directions from intraday data, *Journal of Finance* **46**, 733–746.
- [25] Madhavan, A. & Cheng, M. (1997). In search of liquidity: block trades in the upstairs and downstairs market, *Review of Financial Studies* **10**, 175–203.
- [26] Perold, A.F. (1988). The implementation shortfall: paper vs. reality, *Journal of Portfolio Management* **14**, 4–9.
- [27] Peterson, M. & Sirri, E. (2003). Evaluation of the biases in execution cost estimation using trade and quote data, *Journal of Financial Markets* **6**(3), 259–280.
- [28] Ready, M.J. (1999). The specialist's discretion: stopped orders and price improvement, *Review of Financial Studies* **12**, 1075–1112.
- [29] Seppi, D. (1997). Liquidity provision with limit orders and a strategic specialist, *Review of Financial Studies* **10**, 103–150.
- [30] Sofianos, G. & Werner, I. (2003). The trades of NYSE floor brokers, *Journal of Financial Markets* **6**, 139–176.
- [31] Venkataraman, K. (2001). Automated versus floor trading: an analysis of execution costs on the Paris and New York exchanges, *Journal of Finance* **56**, 1445–1885.
- [32] Werner, I. (2003). NYSE order flow, spreads, and information, *Journal of Financial Markets* **6**, 309–335.

Related Articles

Glosten–Milgrom Models; Kyle Model; Order Types; Price Impact; Specialist Markets.

HENDRIK BESSEMBINDER & KUMAR
VENKATARAMAN

Liquidity Premium

An asset is liquid if it can be traded quickly and at low cost. In addition to risk, liquidity affects asset prices and returns. Because investors want to be compensated for bearing the costs of illiquidity, *asset returns are increasing in illiquidity*. Thus, asset prices should depend on two asset characteristics: risk and liquidity.

Illiquidity reflects the costs of executing a transaction in the capital markets. These costs include three components:

1. *Price-impact costs* reflect the price concession that a buyer or seller makes when trading, or the price of immediacy: a discount when selling or a premium when buying. For small orders, the market impact is confined to the *bid-ask spread* (“spread”), which is the difference between the buying and selling price currently quoted by dealers or market-makers. For large orders, the price impact exceeds the spread and generally increases in the order size. *Depth* is the largest order size for which the price-impact cost is confined to the bid-ask spread.
2. *Search and delay costs* are incurred when a trader searches for better prices than those quoted in the market or wishes to reduce the price impact of his/her order. This often occurs with block orders, where traders prefer to search for counterparties rather than “dump” the entire order on the market. While reducing price-impact costs, the trader bears search and delay costs because the trade is not executed immediately. In particular, the trader incurs opportunity costs and risk while holding the position as the order awaits execution. For example, for a trader who wishes to sell a stock, its price may decline while he/she is searching for a counterparty. The trader then trades off the benefit of a lower price-impact cost against the risk of the market turning against him/her.
3. *Direct trading costs* include exchange fees, taxes, and brokerage commissions. These are also subject to trade-offs: For example, a trader may ask a dealer to liquidate a block, with the dealer bearing the search and delay cost while the trader pays a larger commission.

The three components of transaction costs are highly correlated: Assets with high price-impact costs or high bid-ask spread often have high search and delay costs and high brokerage commissions. Other measures of liquidity are correlated with these three components. For example, more liquid stocks (by the above measures) usually have greater turnover or trading volume and larger market capitalization.

How Does Liquidity Affect Asset Prices?

We have developed a model that shows how liquidity affects asset prices [4]. The model characterizes assets by their transaction costs and investors by their investment horizons. Each investor maximizes the expected present value of the cash flows that his/her assets generate considering the costs of transacting. In equilibrium, the return on an asset is an increasing function of its illiquidity costs because investors require a compensation for bearing these costs. The relation between illiquidity and return is increasing and concave, that is, the increase in return is mitigated for less liquid assets, which are held by investors with longer investment horizons who can depreciate their transaction costs over a longer period of time. The positive effect of illiquidity on stock returns is more prominent for liquid assets, which trade—and hence bear the transaction costs—more frequently.

Although the illiquidity costs of a single transaction are low relative to the asset price (for most publicly traded securities, they are a fraction of a percent), their *cumulative* effect on value is large because they are incurred repeatedly over the security’s life. Thus, the impact of illiquidity costs should equal *at least* the present value of all costs incurred currently and in the future, which can be substantial relative to the value of a stock, whose life is infinite. Further, less liquid securities produce in equilibrium higher after-transaction-costs (net) returns to long-term investors. Higher returns are necessary to induce long-term investors to hold illiquid securities rather than compete with short-term investors for the more liquid securities.

We present empirical evidence on the effects of liquidity on asset prices and expected returns. For stocks, bonds, and other financial instruments, the lower the liquidity, the higher the return or the lower the price (after controlling for other characteristics, such as risk). The evidence further shows that risk-averse investors price illiquidity risk.

Liquidity and Stock Returns

We tested the return–illiquidity relationship on NYSE-AMEX stocks during 1960–1980 [4, 6]. We divided stocks into seven portfolios by their spread, a common measure of illiquidity, and into seven portfolios based on their β coefficient, the capital asset pricing model (CAPM)-based measure of risk, amounting to a total of 49 (7×7) portfolios. In a cross-sectional estimation of the average return on each portfolio as a function of the spread and unsystematic risk, we found that average portfolio returns were significantly higher for stocks with higher spreads. The function was increasing and concave, as predicted by our model. The following formula and chart summarize our results:

$$R_j = 0.0065 + 0.0010\beta_j + 0.0021 \ln(S_j) \quad (1)$$

where R_j is the return on portfolio j and S_j is the spread (as a fraction of price). To illustrate, increasing the spread from 1% to 2% increased the average monthly return by 0.3% ($= 0.21 \cdot \ln(2)$) or about 3.6% per annum. The spread effect remained significant after controlling for firm size.

Subsequent studies support the theory that higher illiquidity entails higher expected returns, after controlling for risk and other characteristics, using a number of alternative measures of liquidity. Brennan and Subramanyam [13] use each stock's price-impact cost (per unit of order size) as well as the fixed cost associated with the spread, finding a strong positive relationship between average stock returns and both measures of illiquidity. Datar *et al.* [17] measure liquidity using stock turnover (the ratio of trading volume to shares outstanding), and [12] use trading volume to measure liquidity. Both studies find that the higher the liquidity, the lower the stock's return after controlling for risk and other characteristics. This relationship, which is consistent with the theory, is found to be very robust.

The illiquidity discount is evident in the pricing of restricted stock, that is, stocks with trading restrictions^a issued by companies that already have identical stocks traded in the public market. Thus, we observe two securities—the publicly traded stock and the same-company restricted stock—with the only difference between them being their liquidity. Consistent with the theory, Silber [21] finds that the average price of the restricted stock is 34% lower than that of the corresponding publicly traded

stock (after controlling for size, earnings, and the relationship between the restricted stockholders and the company).

Liquidity and Bond Yields

We tested the liquidity effect by examining the differences between US Treasury bills and notes with less than six months to maturity [7]. For these maturities, both securities are discount instruments and when maturities are matched, they are completely identical, except that Treasury bills are much more liquid than notes. The average spread on bills in our sample is 0.00775% compared to 0.0303% for notes. The brokerage fees are \$12.5–\$25 per \$1 000 000 value for bills and \$78.125 per \$1 000 000 for notes. Similarly, notes have a higher price-impact cost than bills. Because notes are less liquid than bills with the same maturity, our theory predicts that their yields should be higher. We tested the liquidity effect for 37 randomly selected days between April and November of 1987, analyzing 489 matches of notes and bills with the same maturities. As predicted, we found that the annual yield to maturity on notes was 0.43% higher than on bills with the same maturity. Kamara [18] finds that the note–bill yield differential holds after controlling for a number of security characteristics. A similar pattern is observed in the bond market in the yield differential between on-the-run bonds, which are most recently issued, and their less liquid off-the-run counterparts, whose maturity is only slightly shorter since they were issued earlier. A number of studies document a lower yield to maturity for on-the-run bonds compared to their off-the-run counterpart [19, 22]. In addition, the note–bill yield spread can be viewed as the difference between off- and on-the-run securities.

Corporate bonds have higher yields than similar-maturity Government bonds, and within corporate bonds, lower rated bonds have yet higher yields. This yield differential has traditionally been attributed to differences in the risk of default, which is surely higher for corporate bonds and yet higher for lower rated bonds. However, if illiquidity costs are higher on corporate bonds and on riskier bonds, part of the yield differential is due to illiquidity. This issue is studied by Chen *et al.* [16]. They estimate the illiquidity cost of corporate bonds and find that it generally increases as the bond rating declines. Then,

they estimate the effect of illiquidity on bond yields across bonds, controlling for risk, issuer characteristics, and the bond's special features. They find that illiquidity has a strong positive and robust effect on bond yields. Further, *changes* in bonds' illiquidity costs lead to changes in bond yields, as predicted. In the United States, Rule 144A bonds, which are less liquid corporate bonds whose trading is restricted to "qualified investors," exhibit a significant discount. Chaplinsky and Ramchand [15] find that the yield on Rule 144A bonds is 0.49% higher than on unrestricted bonds with similar characteristics. The differential is particularly large for investment-grade bonds.

Liquidity Changes Over Time and Liquidity Risk

Just as liquidity affects asset prices across securities, liquidity *changes* over time bring about *changes* in asset prices. One example is the October 19, 1987 stock market crash in the United States. Amihud *et al.* [11] propose that the crash resulted partly from a decline in investors' perception of the overall market liquidity compared to the precrash level. Consequently, investors priced securities lower, which led to the crash. Studying NYSE stocks in the S&P 500 list, they find that on October 19th, the dollar spread increased by more than 63% compared to its precrash level, and the quoted depth also declined significantly. The sharp decline in market liquidity followed a period when investors had believed the market could absorb large order flows with a small effect on prices. The study finds that across stocks, those whose price declined the most on the crash day were those with the greatest deterioration in liquidity, and the stocks that recovered most by the end of October 1987 were those whose liquidity improved the most.

Stock liquidity changes when a market trading system is improved. Around the world, markets have been moving from daily call-auction trading to more liquid, higher frequency trading. Amihud *et al.* [9] study the effect of moving stocks in the Israeli stock market from once-a-day call auction to a higher frequency trading regime. They show that stocks that moved to the improved trading system enjoyed greater liquidity (greater turnover and lower volatility-to-volume ratio) as well as a sharp increase in value of at least 6%. Subsequent studies obtain

similar results for other markets that improved their trading systems and became more liquid.

Market liquidity changes not only in response to singular events but also over time for economic reasons. The question that arises is whether this affects stock prices over time. Amihud [2] finds that when market illiquidity rises unexpectedly, stock prices fall and subsequent expected return rises, consistent with our theory. The effect of liquidity shocks on realized current returns and on subsequent expected return is particularly strong for small, illiquid stocks, because of a "flight to liquidity," namely, to larger stocks that are less vulnerable to liquidity shocks.

It follows that for risk-averse investors, a security's *liquidity risk*, measured by the exposure of its value to liquidity shocks, should affect its price and expected return. Pastor and Stambaugh [20] and Acharya and Pedersen [1] find that stocks whose return is more sensitive to market illiquidity shocks have, on average, higher average return (controlling for other stock characteristics). In addition, Acharya and Pedersen find that, within a CAPM model that allows for liquidity risk, higher sensitivity of stock illiquidity to market illiquidity as well as higher sensitivity of stock illiquidity to the market return also bring about higher expected stock returns. That is, systematic liquidity^b risk is priced.

Some Investment Implications

The theory and empirical results suggest that liquidity is priced: for any given level of risk, more illiquid securities have lower price or higher expected return or yield. Risk, too, lowers securities' prices, but there is a large difference between what investors can do to mitigate the costs of risk and illiquidity. To reduce risk, investors can diversify their holdings. Since most of securities' risk is idiosyncratic (uncorrelated across securities), a diversified portfolio enables investors to get rid of most of the risk, exposing investors only to systematic, diversifiable risk. And, while the total economy's risk is not diversifiable, an investor can hedge the market risk using appropriate market instruments (e.g., derivatives). In contrast, the cost of illiquidity cannot be diversified away. It is additive, so a portfolio of illiquid stock remains illiquid. Still, illiquidity can be managed by constructing a portfolio of securities with

different liquidity, ranging from cash-equivalents (the most liquid) to small, infrequently traded stocks and bonds. The objective is to reduce the frequency of trading in the least-liquid securities, thus enjoying their higher return while not bearing the full cost. Instruments that deal with the lower liquidity of small or risky stocks are closed-end funds and exchange-traded funds (ETFs) of such stocks. Trading these relatively traded funds absolves the trader from bearing the higher transaction costs of the underlying stocks. Consistent with the theory, Chan *et al.* [14] find that liquid closed-end funds of countries with low-liquidity stocks trade at a premium (or lower discount).

Asset managers usually focus on stock selection and market timing. However, the manager's skill in buying and selling the positions they manage has a substantial effect on fund performance. In particular, managing the trading so as to minimize transaction costs helps improve fund performance. For example, an alternative to a sale of a large quantity of shares may be to split the order and sell it in pieces in a way that provides an optimal trade-off between lowering the cost of trading and increasing the risk that results from spreading the execution over time.

Because systematic liquidity risk is also priced in addition to the ordinary CAPM β , portfolio optimization must take this risk into account. As transaction costs affect the *net return* on stocks, reducing the risk of net portfolio return calls for portfolio strategies that reduce the exposure of the stock's return and liquidity costs to the market return and to market liquidity shocks.

Some Corporate Finance Implications

The effect of liquidity on the return required by investors means that a company can reduce its cost of capital by increasing the liquidity of its stocks or bonds. The company can thus increase its market value for any given cash flows that it generates, by employing liquidity-enhancing strategies. We suggested [5, 8] a number of such strategies. One strategy is increasing the company's investor base, especially by adding small individual investors, increases liquidity and raises the stock price. Amihud *et al.* [10] find that when Japanese companies made their stocks more accessible to small investors by reducing the minimum trading unit, their investor bases increased,

resulting in an increase in stock liquidity and price. In general, the investor base can be increased by making it easier to trade the company's stocks and bonds (e.g., inducing more dealers to make a market in the company's bonds) and advertising the company so that individual investors are familiar with it.

Liquidity is also enhanced by providing voluntarily more information about the company, since it reduces the asymmetry of information about its value. This can be done by providing transparent financial reporting, making prompt announcements of new information, and facilitating the provision of information to the public, for example, through analysts.

Liquidity is also improved by reducing fragmentation in trading the stocks and bonds of a company. Amihud *et al.* [3] find that when two virtually identical securities are traded in the market—in that case, stock and deep-in-the-money warrants of the same company—the stock's value rises when the warrant is exercised, which eliminates fragmentation and makes the stock more liquid (since its outstanding value increases). In general, the larger the stock or bond issue, the greater its liquidity and the higher its value. This means that the benefits of having multiple classes of stock or small series of bonds should be considered against the resulting reduced liquidity, which raises the corporate cost of capital.

Concluding Remarks

Expected asset returns depend on their liquidity (or marketability) in addition to risk. For both bonds and stocks, the less liquid the asset, the higher its return (after controlling for risk). Further, the effects of liquidity on asset values and returns are larger than one might naively expect, because the costs of illiquidity are incurred repeatedly, whenever the asset is traded.

These results have important implications for investments, corporate financial decisions, and public policy. Securities analysis should incorporate, in addition to cash flows and risk considerations, the liquidity of the security and possible changes in it. Trading and portfolio construction strategies should incorporate liquidity considerations. Companies that issue securities should opt to enhance their liquidity in order to increase their price. And companies should employ strategies that make their publicly traded securities more liquid. Finally, there is value in public

policy that increases market liquidity, which reduces the corporate cost of capital and improves overall economic performance.

End Notes

^aTypically, restricted stock cannot be traded in the public markets over a period of one or two years.

^bAcharya and Pedersen [1] use three liquidity β s to measure systematic liquidity risk, and show these β s are priced.

References

- [1] Acharya, V.V. & Pedersen, L.H. (2005). Asset pricing with liquidity risk, *Journal of Financial Economics* **77**, 375–410.
- [2] Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time series effects, *Journal of Financial Markets* **5**, 31–56.
- [3] Amihud, Y., Lauterbach, B. & Mendelson, H. (2003). The value of trading consolidation: evidence from the exercise of warrants, *Journal of Financial and Quantitative Analysis* **38**, 829–846.
- [4] Amihud, Y. & Mendelson, H. (1986). Asset pricing and the bid-ask spread, *Journal of Financial Economics* **17**, 223–249.
- [5] Amihud, Y. & Mendelson, H. (1988). Liquidity and asset prices: financial management implications, *Financial Management* **17** (Spring), 5–15.
- [6] Amihud, Y. & Mendelson, H. (1989). The effects of beta, bid-ask spread, residual risk and size on stock returns, *Journal of Finance* **44**, 479–486.
- [7] Amihud, Y. & Mendelson, H. (1991). Liquidity, maturity and the yields on U.S. government securities, *Journal of Finance* **46**, 1411–1426.
- [8] Amihud, Y. & Mendelson, H. (2000). The liquidity route to a lower cost of capital, *Journal of Applied Corporate Finance* **12**(4, Winter), 8–25.
- [9] Amihud, Y., Mendelson, H. & Lauterbach, B. (1997). Market microstructure and securities values: evidence from the Tel Aviv Exchange, *Journal of Financial Economics* **45**, 365–390.
- [10] Amihud, Y., Mendelson, H. & Uno, J. (1999). Number of shareholders and stock prices: evidence from Japan, *Journal of Finance* **54**, 1169–1184.
- [11] Amihud, Y., Mendelson, H. & Wood, R. (1990). Liquidity and the 1987 stock market crash, *Journal of Portfolio Management* **16**, 65–69.
- [12] Brennan, M.J., Chordia, T. & Subrahmanyam, A. (1998). Alternative factor specifications, security characteristics, and the cross-section of expected stock returns, *Journal of Financial Economics* **49**, 345–373.
- [13] Brennan, M.J. & Subrahmanyam, A. (1996). Market microstructure and asset pricing: on the compensation for illiquidity in stock returns, *Journal of Financial Economics* **41**, 441–464.
- [14] Chan, J.S.P., Jain, R. & Xia, Y. (2008). Market segmentation, liquidity spillover, and closed-end country fund discounts, *Journal of Financial Markets* **11**, 377–399.
- [15] Chaplinsky, S. & Ramchand, L. (2004). The borrowing costs of international issuers: SEC Rule 144A, *The Journal of Business* **77**, 1073–1097.
- [16] Chen, L., Lesmond, D.A. & Wei, J.Z. (2007). Corporate yield spreads and bond liquidity, *Journal of Finance* **62**, 119–149.
- [17] Datar, V.T., Naik, N.Y. & Radcliffe, R. (1998). Liquidity and stock returns: an alternative test, *Journal of Financial Markets* **1**, 205–219.
- [18] Kamara, A. (1994). Liquidity, taxes, and short-term treasury yields, *Journal of Financial and Quantitative Analysis* **29**, 403–416.
- [19] Krishnamurthi, A. (2002). The bond/old-bond spread, *Journal of Financial Economics* **66**, 463–506.
- [20] Pastor, L. & Stambaugh, R. (2003). Liquidity risk and expected stock returns, *Journal of Political Economy* **111**, 642–685.
- [21] Silber, W.L. (1991). Discounts on restricted stock: the impact of illiquidity on stock prices, *Financial Analysts Journal* **47**, 60–64.
- [22] Warga, A. (1992). Bond returns, liquidity, and missing data, *Journal of Financial and Quantitative Analysis* **27**, 605–617.

Related Articles

Bid-Ask Spreads; Kyle Model; Liquidity.

YAKOV AMIHU & HAIM MENDELSON

Adverse Selection

Adverse selection is potentially an important issue in a financial market whenever any trader has private information related to the value of an asset. An informed trader will strategically adjust his or her trade across states in a manner that reduces the payoff of an uninformed counterparty. The latter is subject to adverse selection: the likelihood of a transaction increases precisely in those states of the world in which the payoff from the transaction is lower.

Adverse Selection in Quote-driven Markets

Consider a single-period binomial model for one stock. In standard fashion, at time 1, the value of the stock is v_h with probability π and v_ℓ with probability $(1 - \pi)$ (see Figure 1). Suppose the risk-free rate is normalized to zero. Then, at time 0, a risk-neutral agent is indifferent between buying and selling any quantity of the stock at the expected value $v_0 = \pi v_h + (1 - \pi)v_\ell$.

Now, suppose our risk-neutral agent at time 0 is asked to trade with an omniscient “informed” counterparty who knows the time 1 value of the stock. If our “uninformed” agent trades at a price $v_0 = \pi v_h + (1 - \pi)v_\ell$, he or she will always lose money: the counterparty buys the asset only if the value will be v_h at time 1, and sells the asset only if the value will be v_ℓ at time 1.

In equilibrium, of course, the uninformed trader rationally anticipates such adverse selection. Thus, if the informed counterparty wishes to buy the stock,

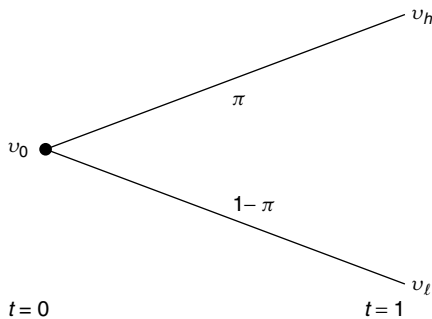


Figure 1 Single-period binomial tree

the uninformed trader is only willing to sell it at v_h . Similarly, the uninformed trader is only willing to buy the stock at v_ℓ . The possibility of adverse selection thus manifests itself as a spread between the ask price (the price at which he or she sells the stock) and the bid price (the price at which he or she buys the stock), thereby imposing a cost on all traders in the market.

The binomial example is perforce stark: we assumed that the counterparty was perfectly informed. More generally, suppose the counterparty is perfectly informed with some probability $\gamma > 0$, and uninformed with probability $1 - \gamma$. Then, we recover the essence of the Glosten and Milgrom [7] model (see **Glosten–Milgrom Models**). The bid–ask spread is easily seen to be increasing in γ .

To take the argument one step further, suppose there are a continuum of states at time 1, and the quantity an informed trader wishes to trade at time 0 is strictly increasing in value at time 1, v_1 , holding fixed the price of the asset. Then, the uninformed agent should infer that v_1 is higher when the net buy order is higher, so the bid–ask spread is naturally increasing in the quantity of shares traded. This intuition lies behind the Kyle [11] model (see **Kyle Model**).

Adverse Selection in Limit Order Markets

There is no market-maker in a limit order market; rather, traders transact directly with each other by posting orders (or price–quantity pairs) in a limit order book, and accepting the orders of other traders. The passage of time since a limit order was submitted allows for new information to arrive at the market. Thus, as Copeland and Galai [3] demonstrate, a limit order is inherently subject to adverse selection.

Suppose v_s , the expected value of the asset at time t_s , given all information available to that time, follows some stochastic process. Two kinds of agents are willing to trade in the market: liquidity or noise traders, who are uninformed, and informed traders. Liquidity traders are motivated to trade by unmodeled hedging or portfolio considerations. Each liquidity buyer i has a reservation value v_i^b , which comes from some distribution F , and is willing to buy at any price less than v_i^b . Informed traders, on the other hand, will buy only at a price less than v_s and sell only at a price above v_s . Each arriving trader is a liquidity trader with probability q_ℓ and an informed trader with probability $1 - q_\ell$.

Now, consider an agent who places a sell order at time t_0 , at an ask price $p_0 > v_0$. Suppose a new trader arrives in the market at time $t_1 > t_0$. There are two possibilities: (i) The buyer is an uninformed trader. Then, with probability $1 - F(p_0)$, no trade occurs. A transaction occurs with probability $F(p_0)$, with the limit order seller earning a payoff $p_0 - E(v_1 | v_0)$. (ii) The buyer is an informed trader. Then, with probability $\Pr(v_1 < p_0)$, no trade occurs. With probability $\Pr(v_1 \geq p_0)$, a transaction occurs, and the payoff to the limit order seller is $p_0 - E(v_1 | v_0, v_1 > p_0) < 0$.

Adverse selection results in the overall transaction probability, $q_\ell F(p_0) + (1 - q_\ell) \Pr(v_1 \geq p_0)$, being higher when $v_1 > p_0$. As Glosten [5] points out, the knowledge that the seller traded with an informed buyer raises the expected value of the asset from $E(v_1 | v_0)$ to the upper tail expectation $E(v_1 | v_0, v_1 > p_0)$.

The price at which the sell limit order is posted, p_0 , is chosen to maximize the submitter's overall expected payoff. Thus, even a relatively simple model of limit orders with asymmetric information involves a complicated decision on the part of a limit order submitter.

Informed Traders Submitting Limit Orders

To take this one step further, an informed trader may submit a limit order. Then, we have double-sided adverse selection: both the trader at t_0 and the trader at t_1 may find themselves trading with a counterparty who possesses better information about the asset's value.

In equilibrium in such a model, will an informed trader submit limit or market orders? This question is posed by Goettler *et al.* [8]. In their infinite horizon model, a Poisson process determines trader arrival. An uninformed trader observes v with a lag Δ . Traders may reenter the market with some delay. Let $\omega_s = \{L_s, (\hat{x}_s, \hat{p}_s), v_{s-\Delta}, v_s\}$ denote the state at time t_s ; the state includes the current limit order book L_s , whether the most recent transaction resulted from a market buy ($\hat{x}_s = 1$) or sell ($\hat{x}_s = -1$) order, the price $\hat{p}_s$ at which the most recent transaction occurred, and the lagged ($v_{s-\Delta}$) and current (v_s) expectations of asset value. An informed trader observes the state at the time he or she enters the market. An uninformed trader observes a subset of the state, $\{L_s, (\hat{x}_s, \hat{p}_s), v_{s-\Delta}\}$. A trader i who executes an order

(x, p) when the asset value is v obtains a payoff $x(\alpha_i + v - p)$, where $x = \pm 1$ denotes the direction of the order, p the price at which it is submitted, and α_i is a private benefit to trade (in the absence of such a private benefit, a no-trade theorem applies, and no transactions occur). A limit order changes the book to L'_s , thus affecting the states viewed by future traders.

Let $\phi(v, \tau | \omega, x, p)$ be the probability that a trader executes at time $t_{s+\tau}$ when the value of the asset is v , if he or she takes action (x, p) in state ω . For an informed trader, adverse selection is manifested in the $\phi(\cdot)$ function, which, in equilibrium, is increasing in v for sell orders and decreasing in v for buy orders. In other words, sell orders are more likely to execute when the asset value has increased (so the sell price is too low) and buy orders when the asset value has decreased (so the buy price is too high). An uninformed trader must take a further expectation over the true state ω and thus faces an added layer of adverse selection.

The model is solved numerically. Given the parameters chosen, in equilibrium, over 50% of the time, informed traders have limit orders at the best quotes, the exact proportion varying with asset characteristics. As a result, both the bid and ask prices and order depths from the limit order book are informative about the value of the asset. Uninformed traders therefore update their beliefs about value on observing the book, which mitigates adverse selection.

Thus, in a limit order market, we can expect both limit order submitters and market order submitters in a limit order market to face adverse selection. This result is consistent with the earlier analytic work of Chakravarty and Holden [2], who predict that an informed trader will use both market and limit orders. Bloomfield *et al.* [1] find that informed traders do use both market and limit orders in an experimental asset market.

Empirical Evidence

Quantitative estimates of the adverse selection component of the bid-ask spread vary by sample. Glosten and Harris [6] propose using the direction of trade to estimate the adverse selection component of the bid-ask spread in a quote-driven market. Their basic

regression equation may be written as

$$\Delta P_t = c_0(Q_t - Q_{t-1}) + \alpha Q_t V_t + \epsilon_t \quad (1)$$

where ΔP_t is the change in transaction price from $t-1$ to t , Q_t is a trade direction indicator (+1 for a market buy and -1 for a market sell), V_t is the size of the order, and ϵ_t is the change in price due to public information. Here, $c_0 + \alpha$ equals the half-spread, with c_0 representing the transitory component due to market-maker costs of supplying liquidity and α the permanent component due to adverse selection. On a sample of 250 NYSE stocks over a 14-month period from December 1981 to January 1983, they estimate that the adverse selection component is on average about 18% of the bid-ask spread.

Madhavan *et al.* [12] extend the basic model to allow for $E[Q_t | Q_{t-1} = 1] = \rho$, the first-order autocorrelation in order flow. The adverse selection component depends on the innovation in order flow. They use generalized method of moments to estimate the model on 274 NYSE stocks over the year 1990, and find that the spread and the adverse selection component both vary by time of day, being largest in the initial half-hour of trading. Estimates of the adverse selection component vary from 51% of the spread in the first half-hour to 35–36% over the latter half of the day. Gibson *et al.* [4] conduct a spread decomposition in the postdecimalization era, and find that although overall spreads are lower, the proportion due to adverse selection and inventory holding costs has increased to 55–59% of the total spread.

A different approach to determining the effect of information on prices is taken by Hasbrouck [9], who specifies a vector autoregression for trade direction and the midpoint of the bid and ask quotes. The permanent price impact of an innovation in the trade process then reflects the effect of information in the market. He finds that the full price impact occurs with a protracted lag, suggesting that information is revealed slowly in the market.

Empirical work on adverse selection in limit order markets is based on Glosten's [5] model of an electronic limit order book. de Jong *et al.* [10] find that adverse selection accounts for 25% (small orders) to 60% (large orders) of the spread on the Paris Bourse over a 44-day period in 1991. Sandås [13] conducts a structural test of the model's restriction that bid prices be equal to the lower tail

expectation of asset value given the order size, and ask prices be equal to the corresponding upper tail expectation. The model is rejected, with the slope of the limit order book being too steep. This finding, interestingly, is consistent with noncompetitive limit order submission, suggesting that informed traders submit some proportion of limit orders.

Conclusion

Adverse selection is important to asset pricing for two reasons. First, in a frictionless market, the bid-ask spread should be zero, with all trades occurring at the price v_0 . However, because of the potential asymmetric information between traders, there is now a cost associated with changing a position in the asset. Indeed, if the adverse selection problem is sufficiently severe, the market may even close down.

Second, adverse selection represents a dimension of counterparty risk that is unhedgeable: one cannot directly trade on the event that the counterparty is informed, since such information is typically unverifiable in a court. Thus, we cannot create a replicating portfolio to hedge this risk, so that asymmetric information across traders implies that markets are fundamentally incomplete. This incompleteness may directly affect traders' valuation for an asset and therefore the risk-neutral probabilities themselves.

References

- [1] Bloomfield, R., O' Hara, M. & Saar, G. (2005). The 'make or take' decision in an electronic market: evidence on the evolution of liquidity, *Journal of Financial Economics* **75**, 165–199.
- [2] Chakravarty, S. & Holden, C. (1995). An integrated model of market and limit orders, *Journal of Financial Intermediation* **4**, 213–241.
- [3] Copeland, T. & Galai, D. (1983). Information effects on the bid-ask spread, *Journal of Finance* **38**(5), 1457–1469.
- [4] de Jong, F., Nijman, T. & Röell, A. (1996). Price effects of trading and components of the bid-ask spread on the Paris Bourse, *Journal of Empirical Finance* **3**(2), 193–213.
- [5] Glosten, L. (1994). Is the electronic limit order book inevitable? *Journal of Finance* **49**, 1127–1161.
- [6] Gibson, S., Singh, R. & Yerramilli, V. (2003). The effects of decimalization on the components of the bid-ask spread, *Journal of Financial Intermediation* **12**, 121–148.

- [7] Glosten, L. & Harris, L. (1988). Estimating the components of the bid/ask spread, *Journal of Financial Economics* **21**(1), 123–142.
- [8] Glosten, L. & Milgrom, P. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* **14**, 71–100.
- [9] Goettler, R., Parlour, C. & Rajan, U. (2008). Informed traders and limit order markets, *Journal of Financial Economics*. Forthcoming.
- [10] Hasbrouck, J. (1991). Measuring the information content of stock trades, *Journal of Finance* **46**(1), 179–207.
- [11] Kyle, A. (1985). Continuous auctions and insider trading, *Econometrica* **53**, 1315–1335.
- [12] Madhavan, A., Richardson, M. & Roomans, M. (1997). Why do security prices change? A transaction-level analysis of NYSE stocks, *Review of Financial Studies* **10**(4), 1035–1064.
- [13] Sandås, P. (2001). Adverse selection and competitive market-making: empirical evidence from a limit order market, *Review of Financial Studies* **14**, 705–734.

CHRISTINE A. PARLOUR & UDAY RAJAN

Price Impact

Price impact refers to the correlation between an incoming order (to buy or to sell) and the subsequent price change. That a buy trade should push the price up seems at first sight obvious and is easily demonstrated empirically, as we will review below. It is also a dour reality for traders for whom price impact is tantamount to a cost: their second buy trade is on average more expensive than the first because of their own impact (and *vice versa* for sells). Monitoring and controlling impact is therefore one of the most active and rapidly expanding domains of research within trading firms. The most important questions are related to the volume dependence of impact (do larger trades impact prices more?), and the temporal behavior of impact (is the impact of a trade immediate and permanent, or is there some lag dependence of the impact?).

A moment of reflection suggests that the interpretation of price impact is far from trivial, and may even lead to contradictions—isn't a transaction a fair deal between a buyer and a seller? So why is there price impact? Three distinct possibilities come to mind:

1. *Agents successfully forecast short term price movements and trade accordingly.* This can result in measurable correlation between trades and price changes, even if the trades by themselves have absolutely no effect on prices at all. If an agent correctly forecasts price movements and if the price is about to rise, the agent is likely to buy in anticipation of it. But in this framework “noise induced” trades, based on no information at all, should have no price impact.
2. *The impact of trades reveal some private information.* The arrival of new private information causes trades, which cause other agents to update their valuations, leading to a price change. But if trades are anonymous and there is no easy way to distinguish informed traders from noninformed traders, then all trades will impact the price since other agents will believe that a fraction of these trades might contain some private information.
3. *Impact is a statistical effect due to order flow fluctuations.* Imagine, for example, a completely random order flow process that leads to a certain order book dynamics (see, e.g., [7] for a description of such models). Conditional to an extra buy

order, the price will on average move up if everything else is kept constant. Fluctuations in supply and demand can thus be completely random, unrelated to information, and still a well-defined notion of price impact will emerge. In this case, impact is a completely mechanical—or better, statistical—phenomenon.

All three of these mechanisms result in some “price impact”, that is, a positive correlation of trading volume and price movement, but they are conceptually very different. In the first two pictures, as emphasized in [13], “orders do not *impact* prices.” It is more accurate to say that orders *forecast* prices.” In the second picture, price impact is of paramount importance to understand how information gets incorporated into prices. If some traders really have an information on the “true” price at some time in the future, then the observation of an excess of buy trades allows the market to guess that the price will move up and to change the quotes accordingly (*see Glosten–Milgrom Models; Kyle Model*; for explicit implementations). Thus, while on one hand market impact is a friction, it should also be viewed as the mechanism that allows prices to adjust to new information.

But if the third, mechanical, interpretation is correct, correlation between price changes and order flow is a tautology. If prices move only because of trades, “information revelation” may merely be a self-fulfilling prophecy that would occur even if the fraction of informed traders is zero. Possible differences between these pictures may come about in the volume dependence and temporal behavior of impact, which we discuss below.

Linear and Permanent Impact: The Kyle Model

The simplest assumption is that impact is both linear in the traded volume and permanent in time. These assumptions can be partly justified within the Kyle model [17], where an insider trader and noise traders submit orders that are cleared by a market maker (MM) every time step Δt . In this model, the price adjustment rule Δp of the MM must be linear in the total signed volume ϵv , that is,^a

$$\Delta p = \lambda \epsilon v \quad (1)$$

where λ is a measure of impact, and is inversely proportional to the liquidity of the market. This price

adjustment is furthermore *permanent*, that is, the price change between time $t = 0$ and $t = T = N \Delta t$ is

$$p_T = p_0 + \sum_{n=0}^{N-1} \Delta p_n = p_0 + \lambda \sum_{n=0}^{N-1} \epsilon_n v_n \quad (2)$$

This formula assumes that the impact $\lambda \epsilon_n v_n$ of trades in the n th time interval persists unabated up to time T . Huberman and Stanzl [15] and Farmer [6] show that the linear price adjustment of the Kyle model is the only specification that does not allow price manipulations, *provided impact is permanent* (see below). From the above equation, it is clear that the sign of the trades must be serially uncorrelated if the price is to follow an unpredictable random walk. Within the setting of the Kyle model, the trading schedule of the insider is precisely such that the ϵ_n are uncorrelated [17]. But data from real markets reveal correlations of the sign of the traded volume over long timescales; we will come back to this important point below.

Measures of Price Impact

An interesting measure of price impact is the correlation between the price change between 0 and T and the sign of the trade at time 0, defined, in general, as

$$R(T) = E[(p_T - p_0) \cdot \epsilon_0] - E[(p_T - p_0)] E[\epsilon_0] \quad (3)$$

In the following, we will assume that drifts can be neglected so that only the first term in the right-hand side (RHS) is not zero. This is often the case, in practice, if T is chosen to be sufficiently small and/or if there is no strong buy/sell asymmetry ($E[\epsilon_0] = 0$).

Within the Kyle framework where the ϵ 's are uncorrelated, $R(T)$ is easily computed and is found to be

$$R(T) = \lambda E[v] \quad (4)$$

The impact function $R(T)$ is therefore *lag independent* in this model. One can, of course, also define a volume-dependent impact function by conditioning the above average on the incoming volume, that is,

$$R(T, v) = E[(p_T - p_0) \cdot \epsilon_0 | v_0 = v] \quad (5)$$

which, in the Kyle model, is equal to λv , for all T . Note that both $R(T)$ and $R(T, v)$ *a priori* depend

on the choice of the elementary timescale Δt , but we want to avoid heavy notations and skip this extra variable dependence.

Another commonly used measure of impact is the correlation $\rho(T)$ between the price change from 0 to T and the total signed volume in the same interval, or, more precisely,

$$\rho(T) = \frac{E\left[(p_T - p_0) \cdot \sum_{n=0}^{N-1} \epsilon_n v_n\right]}{\sqrt{E[(p_T - p_0)]^2 E\left[\left(\sum_{n=0}^{N-1} \epsilon_n v_n\right)^2\right]}} \quad (6)$$

In the Kyle model, this correlation is trivially equal to unity, but this correlation can decrease if liquidity fluctuates (λ becomes time dependent), or if other sources of volatility are taken into account. For example, the MM can revise his quotes in the absence of any trades, say, because of some incoming news. This can be modeled by adding an extra term in equation (2) above.

$$p_T = p_0 + \lambda \sum_{n=0}^{N-1} \epsilon_n v_n + \sum_{n=0}^{N-1} \eta_n \quad (7)$$

where the η_n 's are uncorrelated increments, representing causes of price variations not directly related to trading.

It is often useful to generalize the above definition of $\rho(T)$ by replacing the volumes v_n by v_n^ψ , where ψ is a certain exponent. As discussed below, the measured correlations are found to be stronger when $\psi < 1$.

Empirical Facts

Here we summarize briefly the salient empirical facts that emerged in the last 20 years, concerning the volume dependence and the temporal behavior of impact (see [3] for a recent review). Note that to characterize impact empirically, one has to specify two timescales: (i) the “elementary” timescale Δt over which trades are aggregated; (ii) the timescale T over which the impact of the initial trade is measured.

Concave Volume Dependence

The Kyle model assumes linear dependence of impact on the traded volume. One can measure

the volume-dependent *instantaneous* impact function $R(T = \Delta t, v)$ for different elementary timescales Δt , ranging from the average transaction time to several hours or days. One typically finds that the volume dependence of this impact is *sublinear* and well described by a power-law:

$$R(T = \Delta t, v) \propto v^{\psi(\Delta t)}; \quad \psi(\Delta t) \leq 1 \quad (8)$$

The exponent ψ increases with the elementary timescale, taking rather small values $\psi \simeq 0.1 \rightarrow 0.3$ for individual trades, and increasing toward $\psi = 1$ when Δt corresponds to several thousands of trades, with some concavity remaining for large volumes [14, 20]. The correlation coefficient $\rho(T = \Delta t)$ similarly increases as Δt increases, and reaches rather large values > 0.5 for daily returns of individual stocks, futures, or currencies [5].

Note that the small value of ψ at the individual trade level means that impact is only weakly dependent on volume, in line with many studies that show a stronger correlation of price changes with the *number of trades* rather than with the traded volume [16]. The small value of ψ is often interpreted in terms of discretionary trading: large market orders are only submitted when there is a large prevailing volume at the best quote, a conditioning that mitigates the impact of these large orders.

The above results are established using the total aggregated signed volumes, independent of the origin of the trades. Large brokerage or trading firms can also measure the price impact of their own trades; a concave impact function is usually observed with a value of ψ close to $1/2$ [1] (*see Execution Costs*). For example, the BARRA price impact model posits that, on a time interval Δt needed to complete a typical trade,

$$R(\Delta t, v) = A\sigma\sqrt{\frac{v}{V}} \quad (9)$$

where σ is the volatility per unit time and V the traded volume per unit time, and A a numerical coefficient of order unity [2]. Different theoretical justifications for this square-root impact law are given in [2, 11] and [8].

In the case of stocks, one can also study empirically the influence of the market capitalization M . One finds that when Δt corresponds to a single trade, the data can be approximately rescaled as [18]

$$R(\Delta t, v) \approx M^{-0.3} F\left(M^{0.3} \frac{v}{V}\right) \quad (10)$$

where $\bar{v}$ is the average volume per trade for a given stock, and $F(\cdot)$ a master function that behaves as a power law with exponent ψ .

Impact Cannot be Permanent

As we observed above, a permanent impact model leads to unpredictable price changes only if the signed volume is uncorrelated. However, empirical data show that on a large variety of markets the autocorrelation of the signs ϵ_n decays extremely slowly with time, over at least several days, representing thousands of trades or more [3]. The order flow is therefore found to be strongly persistent and predictable. This comes from the fact that even “highly liquid” markets only offer very small volumes for immediate execution. The fact that the outstanding liquidity is so small has an immediate consequence: trades must be fragmented, and need several hours, days, or even weeks to be completed. This clearly creates long memory in the sign of the order flow and shows that private information can only be *slowly* incorporated into prices. This observation, however, is incompatible with a permanent impact model such as equation (2), which would lead to trends, that is, strongly autocorrelated price changes.

Nonlinear Transient Impact

To reconcile persistent order flow with nearly unpredictable price changes, one can postulate a nonlinear, transient impact model that generalizes equation (2) above.

$$p_T = p_{-\infty} + \lambda \sum_{n=-\infty}^{N-1} G(N-n) \epsilon_n v_n^{\psi} \quad (11)$$

where $G(\ell)$ describes the temporal behavior of impact. One can show that it is possible to choose a certain *decaying shape* for the impact function $G(\ell)$ such as to offset exactly the autocorrelation of the order flow and ensure that the price changes are white noise.

Assume for simplicity that volumes are all equal: $v_n \equiv v, \forall n$. One can then show that if $C(\ell) = E[\epsilon_n \cdot \epsilon_{n+\ell}]$ decays for large ℓ as $\ell^{-\gamma}$ with $\gamma < 1$ (typically $\gamma \approx 0.5$ for stocks), then $G(\ell)$ should itself decay to zero as $\ell^{-\beta}$ with $\beta = (1 - \gamma)/2$ [4]. A permanent impact component $G(\ell \rightarrow \infty) > 0$ is only compatible with the random nature of prices if $C(\ell)$

decays fast enough ($\gamma > 1$). Within this transient impact model with fixed volume of trades, the relation between the price impact function $R(T)$, $G(\ell)$, and $C(\ell)$ reads [4] as follows:

$$\begin{aligned} R(T = \ell \Delta t) &= \lambda v^\psi \left[G(\ell) + \sum_{0 < j < \ell} G(\ell - j) C(j) \right. \\ &\quad \left. + \sum_{j > 0} [G_0(\ell + j) - G_0(j)] C(j) \right] \end{aligned} \quad (12)$$

In other words, the impact $G(\ell)$ of a single trade in isolation is different from the directly measurable impact $R(T)$, which picks up contributions from the fact that trades tend to repeat themselves in the same direction.

Note that even if the impact $G(\ell)$ of an individual trade decays with time, one can show that both the total impact $R(T)$ and the correlation $\rho(T)$ defined by equation (6) tend to a nonzero limit when T is large, whenever the relation $\beta = (1 - \gamma)/2$ holds.

Finally, we mentioned above that the linear impact model, corresponding to $\psi = 1$, is the only choice consistent with the absence of price manipulation strategies if impact is permanent ($\beta = 0$). But if impact is transient, other values of $\psi \leq 1$ become acceptable. Gatheral has recently shown that if $\beta + \psi \geq 1$, price manipulation is not possible [9].

Surprise in the Order Flow

If one insists *a priori* that prices must follow a strict random walk, then only the *surprise* in the order flow can impact the price. In other words, the impact of the trades in the n th interval of time Δt should read (neglecting volume fluctuations) as follows:

$$\Delta p_n = \lambda v^\psi (\epsilon_n - E[\epsilon_n | I_{n-1}]) \quad (13)$$

where I_{n-1} is the information set available just before the n th interval of time. If the ϵ_n 's are independent, then $E[\epsilon_n | I_{n-1}] = 0$ and one recovers the specification of the Kyle model. If, on the other hand, the ϵ_n 's are correlated, then one can form a prediction for the next value of ϵ_n based on the past realizations $\epsilon_{m < n}$, such that the surprise component $\epsilon_n - E[\epsilon_n | I_{n-1}]$ is, by construction, uncorrelated for different times. Within this simplified framework,

there are only two possible outcomes: either the sign of the n th transaction matches the sign of the predictor $E[\epsilon_n | I_{n-1}]$, or the signs are opposite. It is easy to show that the more likely outcome, that is, $\epsilon_n = \text{sign}(E[\epsilon_n | I_{n-1}])$, has the smaller impact [10]. Because of the positive correlation in order flow, the impact of a buy following a buy should be less than the impact of a sell following a buy—otherwise trends would appear. The detailed microstructural mechanism for such an history-dependent asymmetric impact is a topic of research (see [3] and references therein).

Using a linear autoregression model for $E[\epsilon_n | I_{n-1}]$ allows one to identify the above surprise model, equation (13), with the transient impact model of the previous section. Following Hasbrouck's VAR model [12], one may assume that

$$E[\epsilon_n | I_{n-1}] = \sum_{j=1}^{\infty} a_j \epsilon_{n-j} \quad (14)$$

where the coefficients a_j can be computed from the sign autocorrelation $C(\ell)$ using standard methods in filtering theory. Then the above transient impact model is precisely recovered provided the following identification holds:

$$G(\ell) = 1 - \sum_{j=1}^{\ell-1} a_j \quad (15)$$

Spread and Impact are Two Sides of the Same Coin

Since MMs (or liquidity providers) cannot guess the surprise of the next trade, they post a bid price b_n and an ask price a_n given by

$$\begin{aligned} a_n &= p_{n-1} + \lambda v^\psi (1 - E[\epsilon_n | I_{n-1}]); \\ b_n &= p_{n-1} + \lambda v^\psi (-1 - E[\epsilon_n | I_{n-1}]) \end{aligned} \quad (16)$$

The above rule ensures no *ex post* regrets for the market maker: whatever the sign of the trade, the traded price is always the “right” one [19]. The bid-ask spread S is then given by:

$$S = a_n - b_n = 2\lambda v^\psi \quad (17)$$

showing that spread and impact are two sides of the same coin: liquidity providers must be compensated for the adverse impact of market order trades (*see Adverse Selection; Limit Order Markets*). Conversely, one sees that the impact of trades is expected to be proportional to the bid–ask spread, suggesting that the volatility per trade σ_1 is also proportional to the bid–ask spread. Such a correlation is well supported by empirical data [21]. The relation with the volatility *per unit time* σ involves the trading frequency f , through $\sigma = \sigma_1 \sqrt{f}$.

Conclusion

Although “price impact” seems to convey the idea of a forceful and intuitive mechanism, the story behind it might not be that simple. Empirical studies show that the correlation between signed order flow and price changes is indeed strong, but the impact of trades is neither linear in volume nor permanent, as assumed in several models, such as the Kyle model. Impact is rather found to be strongly concave in volume and transient, the latter property being a necessary consequence of the long-memory nature of the order flow. Only after averaging on a long timescale (on the order of days) may an *effective* linear and permanent model make sense. This is an important observation not only for monitoring and control of execution costs but also for building agent-based models of market activity that often posit linear and permanent impact. Coming back to Hasbrouck’s [13] comment –do trades *impact* prices or do they *forecast* future price changes? Since trading on modern electronic markets is anonymous, there cannot be any obvious difference between “informed” trades and “uninformed” trades if the strategies used for their execution are similar. Hence, the impact of any trade must statistically be the same, whether informed or not informed. Impact indeed allows private information to be reflected in prices, but by the same token, random fluctuations in order flow (induced by noise traders) must also contribute to the volatility of markets.

End Notes

^a $\epsilon = +1$ if the volume of buys v_b is larger than the volume of sells v_s in the time interval Δt , and $\epsilon = -1$ in the opposite case; and $v = |v_b - v_s|$.

References

- [1] Almgren, R., Thum, C., Hauptmann, H.L. & Li, H. (2005). Equity market impact, *Risk* **18**(7, July), 57–62.
- [2] BARRA (1997). *Market Impact Model Handbook*, Barra, Berkeley.
- [3] Bouchaud, J.-P., Farmer, J.D. & Lillo, F. (2009). How markets slowly digest changes in supply and demand, in *Handbook of Financial Markets: Dynamics and Evolution*, T. Hens & K. Schenk-Hoppe, eds, Elsevier, North-Holland.
- [4] Bouchaud, J.-P., Gefen, Y., Potters, M. & Wyart, M. (2004). Fluctuations and response in financial markets: the subtle nature of “random” price changes, *Quantitative Finance* **4**(2), 176–190.
- [5] Evans, M. & Lyons, R. (2002). Order flow and exchange rate dynamics, *Journal of Political Economy* **110**, 170–180.
- [6] Farmer, J.D. (2002). Market force, ecology and evolution, *Industrial and Corporate Changes* **11**, 895–953.
- [7] Farmer, J.D., Patelli, P. & Zovko, I. (2005). The predictive power of zero intelligence in financial markets, *Proceedings of National Academy of Sciences of USA* **102**, 2254–2259.
- [8] Gabaix, X., Gopikrishnan, P., Plerou, V. & Stanley, H. (2006). Institutional investors and stock market volatility, *Quarterly Journal of Economics* **121**, 461–504.
- [9] Gatheral, J. (2008). *No dynamic arbitrage and market impact*, Working Paper.
- [10] Gerig, A. (2007). *A Theory for Market Impact: How Order Flow Affects Stock Price*. PhD thesis, University of Illinois (available at: arXiv:0804.3818).
- [11] Grinold, R.C. & Kahn, R.N. (1999). *Active Portfolio Management*, McGraw-Hill.
- [12] Hasbrouck, J. (1991). Measuring the information-content of stock trades, *Journal of Finance* **46**, 179–207.
- [13] Hasbrouck, J. (2007). *Empirical Market Microstructure*, Oxford University Press.
- [14] Hasbrouck, J. & Seppi, D. (2001). Common factors in prices, order flows and liquidity, *Journal of Financial Economics* **59**, 388–411.
- [15] Huberman, G. & Stanzl, W. (2004). Price manipulation and quasi-arbitrage, *Econometrica* **74**, 1247–1276.
- [16] Jones, C., Kaul, G. & Lipson, M.L. (1994). Transactions, volume, and volatility, *Review of Financial Studies* **7**, 631–651.
- [17] Kyle, A. (1985). Continuous auctions and insider trading, *Econometrica* **53**, 1315–1335.
- [18] Lillo, F., Farmer, J.D. & Mantegna, R.N. (2003). Master curve for price impact function, *Nature* **421**, 129–130.
- [19] Madhavan, A., Richardson, M. & Roomans, M. (1997). Why do security prices fluctuate? A transaction-level analysis of NYSE stocks, *Review of Financial Studies* **10**, 1035–1064.

- [20] Plerou, V., Gopikrishnan, P., Gabaix, X. & Stanley, H.E. (2002). Quantifying stock price response to demand fluctuations, *Physical Review E* **66**, 027104.
- [21] Wyart, M., Bouchaud, J.-P., Kockelkoren, J., Potters, M. & Vettorazzo, M. (2008). Relation between bid-ask spread, impact and volatility in double auction markets, *Quantitative Finance* **8**, 41–57.

Related Articles

Algorithmic Trading; Inventory Effects; Kyle Model; Liquidity; Market Microstructure Effects.

JEAN-PHILIPPE BOUCHAUD

Inventory Effects

Traditional economic models of price-setting focus on call-auction markets in which all trading occurs simultaneously, at preestablished discrete times, with no market makers involved. Such models leave no role for any of the three sources of friction found in modern models of market liquidity: inventory, **order-processing costs**, and **adverse selection**. As market microstructure research has developed, researchers studying the links between inventories and liquidity have shed considerable light on how market makers resolve temporal imbalances in the continuous trading environment that characterizes most financial markets.

Linking Inventories and Liquidity

The study of how inventories affect pricing and liquidity in financial markets began with the work of Garman [6]. Garman models continuous trading with a market maker, or dealer, who smooths out temporary imbalances in other traders' orders to buy and sell securities. In Garman's model there is a single risk-neutral, monopolistic dealer who faces stochastic order arrival. The dealer has no ability to borrow either cash or securities, and he sets prices at which he is willing to sell (ask price) and buy (bid price) only once, before trading begins. The dealer's objective is to maximize his expected profit per unit of time subject to avoiding bankruptcy or failure. This formulation gives rise to a classic gambler's ruin problem. By setting his ask price higher than his bid price (a positive **bid–ask spread**), the dealer protects himself from certain failure, although he still faces a positive probability of failure. A key contribution of Garman's model is the idea that a dealer's inventory affects his viability. A limitation of his model is that all prices are set before trading begins; the dealer cannot adjust his prices as his inventory changes.

Inventory Models with No Possibility of Dealer Failure

Amihud and Mendelson [1] build on Garman's intuition to study how a dealer changes his prices as his inventory changes, explicitly incorporating inventory

into the dealer's pricing problem. As in Garman's model, there is a single risk-neutral, monopolistic dealer whose objective is to maximize expected profit, but Amihud and Mendelson assume that inventory is bounded above and below by exogenous parameters, eliminating the possibility of dealer failure. This allows a tighter focus on the dealer's optimization problem. This model incorporates a semi-**Markov process** in which inventory is the state variable. The decision variables (bid and ask prices) depend on the level of the state variable (inventory)—thus the bid and ask prices change over time as the level of inventory changes. Amihud and Mendelson's model predicts that the dealer's optimal ask price is above his bid price, but in this model the positive spread reflects the dealer's market power. (Extending the model to include competitive dealers would drive the spread to zero.) In addition, Amihud and Mendelson find that the dealer has a preferred or target inventory position, and he adjusts his prices to return to his target inventory. If the dealer is too long, to reduce inventory toward his target level he lowers both the bid and ask prices to induce other traders to buy. The reverse is true if the dealer is below his target inventory: To raise inventory toward its target level the market maker raises both the bid and ask prices to induce other traders to sell. Amihud and Mendelson's model thus predicts that the level of bid and ask prices is a monotonically decreasing function of the dealer's inventory. In the special case of linear supply and demand, higher inventory leads to wider spreads.

The models of Stoll [19] and Ho and Stoll [10, 11] maintain the assumption that the dealer cannot go bankrupt, but they focus on how a risk-averse dealer's inventory, **order-processing costs**, and **adverse selection** risk affect his pricing. The dealer is willing to alter his portfolio away from his desired position to accommodate the trading desires of others. The dealer is risk averse and cannot hedge his inventory exposure, so he must be compensated for bearing the risk of changing his portfolio from his target level. In these models the spread between bid and ask prices arises as compensation for the portfolio risk borne by a risk-averse dealer. Inventory causes the dealer to move both bid and ask prices up (if he is below his target portfolio) or down (if he is above his target portfolio) by the same amount, so inventory affects the level of bid and ask prices but not the magnitude of the spread between them. But

here the spread reflects dealer risk aversion, rather than market power as in [1]. Ho and Stoll [10, 11] demonstrate that the intuition of inventory affecting a dealer's quote levels but not the magnitude of his spread is robust to multiperiod and multidealer configurations. In a multidealer setting, Ho and Stoll [11] further predict that relative inventory positions among dealers determine the amount of interdealer trading.

Inventory Models with Capital Constraints

Most of the classical inventory models [1, 10, 11, 16, 19] assume either explicitly or implicitly that dealers have access to unlimited additional capital. Recent work by Brunnermeier and Pedersen [2] and Gromb and Vayanos [7] relaxes this assumption, examining how liquidity provision is affected by the inventory of dealers or arbitrageurs who face capital constraints.

Brunnermeier and Pedersen [2] construct a model linking an asset's market liquidity with market makers' funding, explicitly acknowledging that market makers generally cannot raise new capital instantly. In this model, market makers are risk neutral and their objective is to maximize final wealth, subject to the constraint that the total margin (the amount of capital required to finance their inventory positions) cannot exceed their capital. Brunnermeier and Pedersen show that when market makers' margin requirements approach their available capital, market makers provide less liquidity to the market and bid-ask spreads widen. This model thus predicts that very large inventory positions lead to wider spreads, in contrast to the capital-unconstrained models in which inventories have no effect on the magnitude of bid-ask spreads. Further, Brunnermeier and Pedersen predict that when large inventory positions push market makers close to their capital limits, market makers provide liquidity mostly in lower-risk securities, causing a "flight to quality" in which risky securities become especially illiquid.

Similar predictions can be obtained from the Gromb and Vayanos [7] capital-constraints model, although their main predictions concern the welfare implications of capital-constrained market makers.

Empirical Evidence on Inventory Effects

Models of how market-maker inventories affect pricing and liquidity in financial markets all predict that

a market maker will set his ask price above his bid price for reasons including bankruptcy risk (Garman [6]), market power (Amihud and Mendelson [1]), and risk aversion (Stoll [19], Ho and Stoll [10, 11]). This prediction of a positive bid-ask spread is borne out by observations in virtually all financial markets.

All of the inventory models also predict that inventories affect the level of bid and ask prices, with market makers lowering prices to reduce their inventory when they hold large positions and raising prices when their inventory positions are low or negative. Lyons [12] and Cao *et al.* [3] find evidence of such inventory-induced quote shading in the foreign exchange market, as do Garleanu *et al.* [5] in the S&P500 index options market. However, in the futures market, Manaster and Mann [15] find that futures market makers (who differ somewhat from the typical dealer in inventory models) actually raise their prices when they have larger inventory positions, in contradiction of the inventory models' prediction. Using London Stock Exchange (LSE) data, Hansch *et al.* [8], Reiss and Werner [18], and Naik and Yadav [17] find support for market makers' controlling risk by mean-reverting their inventory positions toward a target level. Hasbrouck and Sofianos [9] and Madhavan and Sofianos [14] find evidence that specialists on the New York Stock Exchange (NYSE) have preferred inventory positions and that their inventories affect prices, but Madhavan and Smidt [13] also find that specialist inventories deviate from their target levels for periods as long as several weeks.

Several researchers analyze the link between inventories and liquidity suggested by inventory models. Consistent with the predictions of Ho and Stoll [11] for a market with competitive dealers, Hansch *et al.* [8] and Reiss and Werner [18] find that differences in inventories across dealers on the LSE affect the extent of interdealer trading, but that individual dealer inventories do not affect the magnitude of each dealer's quoted spread intraday. On an interday level, Comerton-Forde *et al.* [4] find that large NYSE specialist inventories lead to wider spreads, consistent with the predictions of capital-constraints models such as those of Brunnermeier and Pedersen [2]. Comerton-Forde *et al.* [4] also find evidence that when NYSE specialist inventories are particularly high, spreads widen more for high-risk than for low-risk stocks, supporting Brunnermeier and Pedersen's [2] inventory explanation for flight to quality.

Summary

Dealer inventory was among the first market frictions studied by market microstructure theorists. Inventory models predict that dealers set ask prices above bid prices, that they lower their quotes when they have very large inventory positions, and that they may or may not change the magnitude of their quoted spreads as their inventory changes, depending on the particular modeling assumptions. Multidealer inventory models further predict that relative inventory positions give rise to interdealer trading and determine which dealers have the best (lowest ask or highest bid) quotes in a market. These predictions have been tested and largely borne out in empirical studies to date, but the difficulty of obtaining data on dealer inventories has limited researchers' ability to test the models in some of the largest dealer markets, such as the fixed income market.

References

- [1] Amihud, Y. & Mendelson, H. (1980). Dealership market: market making with inventory, *Journal of Financial Economics* **8**, 31–53.
- [2] Brunnermeier, M. & Pedersen, L. (2009). Market liquidity and funding liquidity, *Review of Financial Studies*, **22**, 2201–2238.
- [3] Cao, H., Evans, M. & Lyons, R. (2006). Inventory information, *Journal of Business* **79**, 325–363.
- [4] Comerton-Forde, C., Hendershott, T., Jones, C., Moulton, P. & Seasholes, M. (2009). Time variation in liquidity: the role of market maker inventories and liquidity, *Journal of Finance*, forthcoming.
- [5] Garleanu, N., Pedersen, L. & Poteshman, A. (2009). Demand-based option pricing, *Review of Financial Studies*, forthcoming.
- [6] Garman, M. (1976). Market microstructure, *Journal of Financial Economics* **3**, 257–275.
- [7] Gromb, D. & Vayanos, D. (2002). Equilibrium and welfare in markets with financially constrained arbitrageurs, *Journal of Financial Economics* **66**, 361–407.
- [8] Hansch, O., Naik, N. & Viswanathan, S. (1998). Do inventories matter in dealership markets? Evidence from the London Stock Exchange, *Journal of Finance* **53**, 1623–1655.
- [9] Hasbrouck, J. & Sofianos, G. (1993). The trades of market-makers: an analysis of NYSE specialists, *Journal of Finance* **48**, 1565–1594.
- [10] Ho, T. & Stoll, H. (1981). Optimal dealer pricing under transactions and return uncertainty, *Journal of Financial Economics* **9**, 47–73.
- [11] Ho, T. & Stoll, H. (1983). The dynamics of dealer markets under competition, *Journal of Finance* **38**, 1053–1074.
- [12] Lyons, R. (1995). Tests of microstructural hypotheses in the foreign exchange market, *Journal of Financial Economics* **39**, 321–351.
- [13] Madhavan, A. & Smidt, S. (1993). An analysis of daily changes in specialist inventories and quotations, *Journal of Finance* **48**, 1595–1628.
- [14] Madhavan, A. & Sofianos, G. (1998). An empirical analysis of NYSE specialist trading, *Journal of Financial Economics* **48**, 189–210.
- [15] Manaster, S. & Mann, S. (1996). Life in the pits: competitive market making and inventory control, *Review of Financial Studies* **9**, 953–975.
- [16] Mildeinstein, E. & Schleef, H. (1983). The optimal pricing policy of a monopolistic market maker in the equity market, *Journal of Finance* **38**, 218–231.
- [17] Naik, N. & Yadav, P. (2003). Do dealer firms manage inventory on a stock-by-stock or a portfolio basis? *Journal of Financial Economics* **69**, 325–353.
- [18] Reiss, P. & Werner, I. (1998). Does risk sharing motivate inter-dealer trading? *Journal of Finance* **53**, 1657–1704.
- [19] Stoll, H. (1978). The supply of dealer services in security markets, *Journal of Finance* **33**, 1133–1151.

Related Articles

Adverse Selection; Bid–Ask Spreads; Order Types; Order Flow; Specialist Markets.

PAMELA C. MOULTON

Intraday Price Efficiency

The efficient price of a security represents the present value of expected future cash flows conditional on current information. Although observed prices are estimates of this quantity, deviations can and will occur. Price efficiency measures summarize the extent of these deviations. The measures presented here focus on deviations that arise and are subsequently corrected over very short horizons within the day, and they facilitate comparisons through time as well as across securities and markets. As these measures are intended to evaluate the closeness of actual prices to efficient prices, they relate to transaction costs and describe an important dimension of overall market quality.

Measures^a

Measures of intraday price efficiency, either explicitly or implicitly, define the efficient price for a security as a random walk. New information may arrive continuously, and deviations from the random-walk benchmark can be considered to be inefficiencies. These inefficiencies arise from a variety of sources including discreteness, inventory control, and other non-information-based components of the bid-ask spread. In principle, inefficiencies could also arise from irrationality among market participants (e.g., overreacting to the size of a trade).

Hasbrouck [9] uses variance decomposition methodologies to assess inefficiencies. He defines the (log) transaction price, p_t , as the sum of a random-walk component (the efficient price) m_t , and a transitory, mean-zero pricing error, s_t , where t indexes transaction time:

$$p_t = m_t + s_t \quad (1)$$

The efficient price is unobservable and its innovations are new public information as well as the information content of order flow. Because the pricing error has a mean of zero, its standard deviation, σ_s , is a measure of its magnitude. Intuitively, σ_s summarizes how closely transaction prices follow the efficient price over time and serves as an inverse measure of price efficiency.

Hasbrouck proposes a general structure for the pricing error, where s_t is related to information in

trades, non-trade-related public information, and a disturbance. To connect this model with the data, he uses a vector autoregression (VAR) system over $\{r_t, x_t\}$, where r_t is the first difference of p_t and x_t is a vector of trading variables. This model is underidentified, so additional restrictions from the variance decomposition literature must be imposed. Specifically, the restriction of [1] effectively relates s_t only to the trading information contained in x_t , and the vector moving average representation is used to estimate σ_s . Because of the restriction, the estimated pricing error is a lower bound. Factors that would cause σ_s to exceed this quantity must be uncorrelated with lagged returns as well as the set of trade variables.

There is no formal model describing the conditioning information in x_t , but it should be constructed to explain meaningful variation in the true pricing error. Hasbrouck uses a 3×1 vector of the following trade variables: (i) a trade sign indicator, (ii) signed trading volume, and (iii) the signed square root of trading volume. This structure of x_t allows for a nonlinear relationship between pricing error and the trade series. For some perspective, Hasbrouck uses a sample of NYSE stocks from 1989 to demonstrate that this specification increases the estimate for σ_s by about 40% over that from a simple univariate estimation of equation (1); the added explanatory power of the x_t strengthens the lower bound and provides a more accurate estimate of σ_s .

In addition to the structure of x_t , numerous methodological considerations arise. First, the VAR system must be truncated at some lag length beyond which serial dependencies are assumed negligible. There is no theoretical guidance for this decision, but one could determine a cutoff point *ad hoc* by inspecting VAR estimates over various lags. Second, trading information in x_t will typically require the signing of trades. A common method compares prices to midpoint quotes [15]. Third, overnight returns are considered atypical due to the time delay as well as different trading mechanisms at the opening and therefore may be neglected in the analysis.

A number of empirical studies have used some variant of Hasbrouck's methodology in recent years. Kumar *et al.* [14] find that pricing errors decline for equities upon option listings. George and Hwang [6] extend the model to allow for heteroskedasticity in s_t , and they find that the pricing errors are similar at the opening and close of trading. Hotchkiss and Ronen

[12] show that the price efficiency of bonds resembles that of the underlying stocks. Boehmer *et al.* [3] find that the price efficiency of NYSE stocks improved in 2002 when the contents of the limit order book were made available to traders off the exchange. Finally, Boehmer and Kelley [2] find that pricing errors decline with more institutional ownership and trading.

As an alternative to the variance decompositions, one may also measure intraday efficiency using the autocorrelation properties of the return series. Consider once again an efficient price that, by definition, follows a random walk. For a simple implementation, one can construct a series of price changes over fixed increments in time (say, 30 min) and estimate the autocorrelation. Since either positive or negative autocorrelation is a deviation from a random walk, a summary (inverse) efficiency measure is the absolute value of the autocorrelation.

A slightly more complex implementation involves variance ratios. $VR(n, m)$ is the ratio of an m -period variance to an n -period variance, where both variances are scaled by the number of periods. For example, $VR(5 \text{ min}, 60 \text{ min})$ is the ratio of a 60-min return variance to a 5-min return variance, scaled by 60 and 5, respectively. Under a random walk, variance ratios will equal 1, so the summary (inverse) efficiency measure is $|1 - VR(n, m)|$. A variance ratio contains more information than a single autocorrelation, as $VR(n, m)$ can be written as a linear combination of the first $n - 1$ autocorrelations (see [4], p. 49).

The use of autocorrelations or variance ratios also requires some methodological consideration. Most importantly, and in contrast to Hasbrouck's methodology, one needs to remove the bid-ask bounce described in [17]. The best way to do this is by computing returns according to bid-ask midpoints. In addition, for some securities, stale quotes may be a problem. One may mitigate stale quote issues by ignoring intervals over which no new quote was posted.

As a summary measure of price efficiency, the Hasbrouck approach is favorable for a couple of reasons. First, the variance decomposition methods differentiate between permanent and transient price changes and attribute deviations from a random walk to the latter alone. Nonzero autocorrelation, and hence variance ratios deviating from unity, can arise from either efficient or inefficient pricing. For

example, order splitting by informed traders may lead to positive autocorrelation. Second, pricing errors are measured in trade time—the model is advanced by one period after each trade—whereas autocorrelations and variance ratios are based on returns computed in clock time. Measuring efficiency in trade time has the added benefit of granting more weight to periods of active information discovery. Moreover, the trade-time measure reflects the actual behavior of market participants and includes the information conveyed by each single trade.

It is also important to recognize that neither the variance decompositions nor the autocorrelation-based measures will pick up deviations from the efficient price persisting for more than a brief portion of the trading day, not to mention weeks or months. That is, only pricing errors that are quickly corrected will be measured. This approach ignores potential deviations from fundamentals in contexts such as long-term reversals as in [5], return momentum as in [13], or weekly return reversals as in [16].

Boehmer and Kelley [2] have estimated a variety of price-efficiency measures for all NYSE stocks for each quarter from 1983 to 2004. Figure 1 tracks the cross-sectional average for σ_s normalized by σ_p from equation (1) over the 22-year period. For comparison, the cross-sectional average of the absolute value of the first-order autocorrelation for a series of 30-min quote midpoint returns is also shown. Autocorrelations and variance ratios incorporating various other intraday horizons produce a similar picture. Both plots illustrate a general decline through time, which is interpreted as an overall *increase* in price efficiency. This matches a similar trend in other transactions costs such as execution costs. More importantly, an inspection of the plots reveals that both measures increase around the market crash of 1987, but there is a substantial drop in σ_s alone around the 16th pricing in 1997 and on decimalization in 2001—events that intuitively should lead to greater efficiency. Turning to the cross section, unreported tests show that the average cross-sectional correlation between the two measures is around 0.25, and that both measures generally decrease with firm size, execution costs, price, and trading activity.

The Evolution of the Efficient Price

While this article focuses on the extent to which transaction prices resemble an efficient price, similar

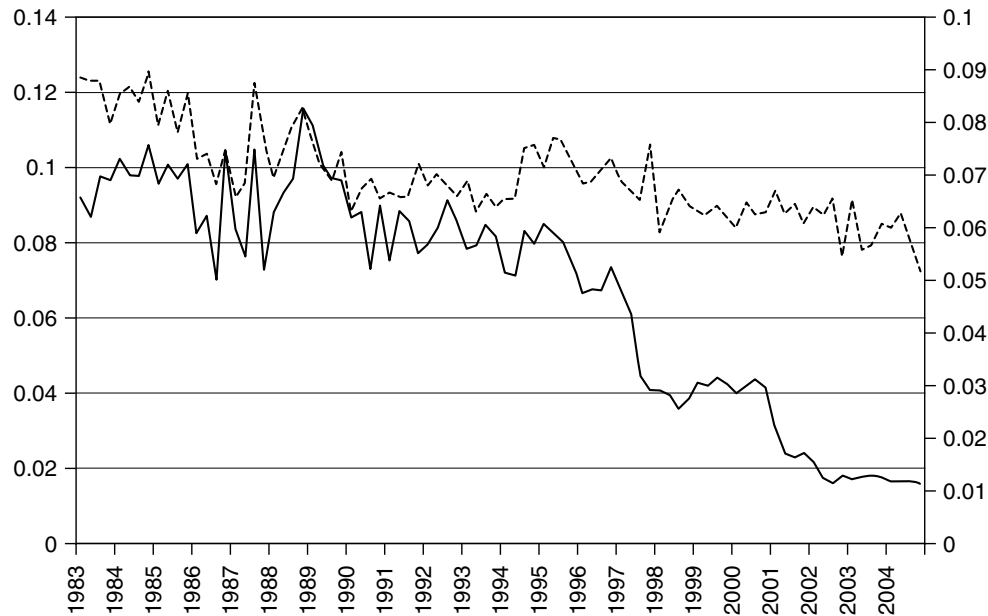


Figure 1 Average price efficiency through time. The sample is based on NYSE-listed securities between 1983 and 2004. The solid line tracks the cross-sectional mean of α_s/σ_p , which is the pricing error from Hasbrouck [9] divided by the standard deviation of transaction prices. The dashed line tracks the cross-sectional mean of $|AC30|$, which is the absolute value of the 30-min return autocorrelation. The left and right vertical axes are for α_s/σ_p and $|AC30|$, respectively

estimation techniques describe the evolution of the underlying efficient price itself.^b Specifically, Hasbrouck [7, 8] starts with equation (1) and considers a VAR over returns and trades. In this case, however, prices and returns are based on quote midpoints (rather than transaction prices) and the trade at time t precedes the quote change indexed by t . Thus, the quote change at time t reflects information in the trade, but not *vice versa*. Hasbrouck [7] uses the impulse response function as a measure of the information content of a trade innovation. He also [8] uses variance decompositions similar to the ones used by him in [9] to develop two *summary measures* of trade informativeness. The first is the component of the random-walk variance attributable to trades, which serves as an absolute measure of the private information revealed by trading. The second is the first measure scaled by the total random-walk variance. It is interpreted as the importance of trade-based information relative to all public information for price discovery. These measures are superior to other metrics such as the quoted spread or the permanent price impact (as in [7]), which condition on trade size. Finally, Hasbrouck [10] provides a multiple

market extension to assess the relative price discovery for each market in which a security trades. Here, there is a single random-walk efficient price and each market's information share is defined as the proportion of the random-walk variance attributable to its trades.

Summary

Intraday price efficiency can be measured using variance decomposition methods or by various autocorrelation-based measures. Because they explicitly characterize price changes as permanent or temporary, and they employ useful transaction-time models, variance decomposition methods are preferred. Several studies using these measures are also described.

End Notes

^a. Much of this section is adapted from the discussion in [2].

^b. See [11] for a survey.

References

- [1] Beveridge, S. & Nelson, C. (1981). A new approach to the decomposition of economic time series into permanent and transitory components with particular attention to the measurement of the 'business cycle', *Journal of Monetary Economics* **7**, 151–174.
- [2] Boehmer, E. & Kelley, E. (2009). Institutional investors and the informational efficiency of prices, *Review of Financial Studies* **22**, 3563–3594.
- [3] Boehmer, E., Saar, G. & Yu, L. (2005). Lifting the veil: an analysis of pre-trade transparency at the NYSE, *Journal of Finance* **60**, 783–815.
- [4] Campbell, J.Y., Lo, A.W. & MacKinlay, A.C. (1997). *The Econometrics of Financial Markets*, Princeton University Press, Princeton, NJ.
- [5] DeBondt, W.F.M. & Thaler, R. (1985). Does the stock market overreact? *Journal of Finance* **40**, 793–805.
- [6] George, T. & Hwang, C. (2001). Information flow and pricing errors: a unified approach to estimation and testing, *Review of Financial Studies* **14**, 979–1020.
- [7] Hasbrouck, J. (1991). Measuring the information content of stock trades, *Journal of Finance* **46**, 179–207.
- [8] Hasbrouck, J. (1991). The summary informativeness of stock trades: an econometric analysis, *Review of Financial Studies* **4**, 571–595.
- [9] Hasbrouck, J. (1993). Assessing the quality of a security market: a new approach to transaction-cost measurement, *Review of Financial Studies* **6**, 191–212.
- [10] Hasbrouck, J. (1995). One security, many markets: determining the contributions to price discovery, *Journal of Finance* **50**, 1175–1199.
- [11] Hasbrouck, J. (2002). Stalking the “efficient price” in market microstructure specifications: an overview, *Journal of Financial Markets* **5**, 329–339.
- [12] Hotchkiss, E. & Ronen, T. (2002). The informational efficiency of the corporate bond market: an intraday analysis, *Review of Financial Studies* **15**, 1325–1354.
- [13] Jegadeesh, N. & Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency, *Journal of Finance* **48**, 65–91.
- [14] Kumar, R., Sarin, A. & Shastri, K. (1998). The impact of options trading on the market quality of underlying securities: an empirical analysis, *Journal of Finance* **53**, 717–732.
- [15] Lee, C.M.C. & Ready, M. (1991). Inferring trade direction from intraday data, *Journal of Finance* **46**, 733–746.
- [16] Lehmann, B.N. (1990). Fads, martingales, and market efficiency, *Quarterly Journal of Economics* **105**, 1–28.
- [17] Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market, *Journal of Finance* **39**, 1127–1139.

Related Articles

Bid–Ask Spreads; Efficient Market Hypothesis; High-frequency Data; Liquidity; Price Impact; Roll Model.

ERIC K. KELLEY

High-frequency Data

High-frequency financial data usually refer to data sampled at a time horizon smaller than a trading day [5, 6, 8]. In this sense, high-frequency data are essentially intended as intraday data. However, the meaning of high frequency has changed over the years following the availability of more and more detailed information on the trading process. A stricter viewpoint considers those financial records where no kind of temporal aggregation is done as high-frequency data. The use of such data in finance is recent and dates to the late 1980s [16]. Among the first high-frequency databases worth mentioning are the Berkeley Option Data, containing bid–ask quotes and trade prices of option traded at the Chicago Board Option Exchange (CBOE), the TORQ database containing trades and quotes relative to 144 New York Stock Exchange (NYSE) stocks (1990–1991), the HFDF93 database of foreign exchange data released by Olsen & Associates (1992–1993), and the computerized trade reconstruction records maintained by the Commodities Futures Trading Commission (CFTC). In the following, we describe the most common types of datasets.

- **Tick-by-tick data**

The simplest form of financial data without any temporal aggregation is the so-called tick-by-tick data. In this type of data, all transactions are recorded. Typically, the dataset contains information on the time, price, and volume of each trade. Clearly, the time frequency at which data are recorded is not fixed, but can be different in different phases of a trading day (see below).

- **Trade and quote data**

A more refined type of dataset contains information on transactions and best quotes (Level 1 data). Transaction data usually contain the transaction prices, the time when the transaction occurred, and the exchanged volume. Quote data usually contain information about the best bid and the best ask price and the available volume at these prices. Quote data are usually updated every time there is a quote change either in the price or in the volume. The availability of trade and quote

data allows one to infer whether the transaction was buyer or seller initiated. The idea is to compare the trade price with the bid and ask prices of the prevailing quote, which is the most recent past quote. One of the most famous methods to assign signs to trades is by using the Lee and Ready algorithm [10]. However in some markets (e.g. NYSE), the signing process is not always straightforward because of some data and market subtleties; for a more recent discussion, see, for example, [14].

One of the most famous trade and quote database is the TAQ (Trade and Quote) database, maintained by NYSE since 1993. The trades and quotes data are available for a wide range of stock exchanges and markets worldwide.

- **Order book data**

The next level of high-frequency data resolution (Level 2 data) includes complete information about the order book in a given market. While the quote data contain information only on the best bid and best ask prices, in the order book data, all limit orders to buy or sell are recorded both when they are placed and when they expire or are canceled. This type of dataset allows one to reconstruct the limit order book at any instant of time and to follow the market dynamics more closely. A variant of this type of database contains the transaction data and the price and volume of a fixed number (usually 5 or 10) of best price levels on both sides of the order book. The use of such data dates back to the early 1990s [1]. One of the most popular databases of order book data is the Rebuild Order Book maintained by the London Stock Exchange (LSE). A different type of order book dataset provides a snapshot of the book with a fixed time frequency, rather than providing each individual event affecting the order-book dynamics. For example, the OpenBook database of the NYSE provides a snapshot of the order book refreshed every second.

- **Market member data**

Recently, a few markets have released financial datasets with another level of resolution. These datasets contain also the identity of the market members (and/or brokers, traders) involved

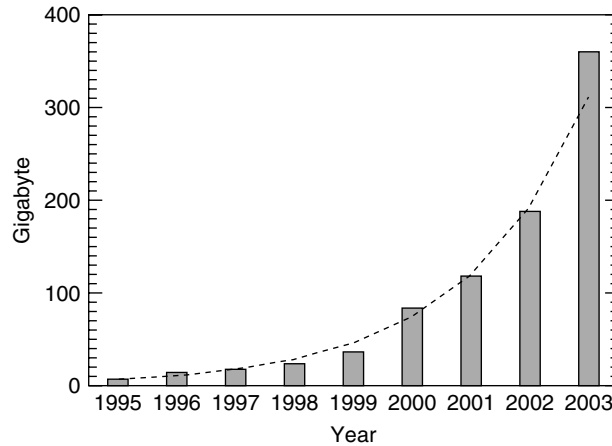


Figure 1 Size in gigabytes of the trade and quote (TAQ) database of the New York Stock Exchange during the period 1995–2003. The dashed line is a fit with an exponential function and the estimated doubling time is 1.5 years

in each transaction and sometimes of the market member who placed each limit order. For reasons of privacy, the identity of the market member is often coded, and, in some cases, the coding is different for different stocks or different time periods. This type of database allows one to identify the different strategies followed by different traders [12]. For example, the Spanish Stock Exchange releases such a database where the market member identity is completely transparent. In other markets, this information is recorded although it is not transparent to the agents during the market activity.

Two important aspects of high-frequency data are data size and time resolution. The size of high-frequency data has significantly increased in the last years because of the explosion of trading activity in financial markets. As an example, Figure 1 shows the approximate size (in gigabytes) of the TAQ database of the NYSE for the years from 1995 to 2003. The size has increased more than 50 times in magnitude, and the growth seems to be well fitted by an exponential law with a doubling time of about 1.5 years. More recently, the total volume of data related to United States large caps on a typical day of 2007 was 57 million lines, approximately a gigabyte of stored data.

The size of high-frequency data increases also when the level of data resolution is increased, for example, data on a high cap LSE stock in 2002

occupy roughly 12 kB for daily data, 15 MB for trade data, and more than 100 MB for order book data.

Concerning the issue of time resolution, very often, high-frequency data are recorded with a 1-s resolution. While this resolution was sufficient several years ago, nowadays, this resolution is insufficient to allow a correct time ordering of the market events. To give an example, at the end of 2007, Apple had on average 6 transactions per second, and on the order of 100 events per second affecting the order book. With this event frequency, it is very hard to infer the correct time ordering of events within a second. Of late, high-frequency databases are either released with a 1-ms resolution, or still have the 1-s resolution but the events are correctly sorted out within the second, according to the actual order of events.

Issues with High-frequency Data

Working with high- and ultrahigh-frequency financial data presents many additional issues when compared to the more traditional analysis of daily or weekly data. Besides the fact that high-frequency data analysis is computationally more demanding, the issues involved can be categorized into at least seven classes, which we have chosen as follows: data cleaning, irregular temporal spacing, discreteness, variable definition, daily periodicities, temporal correlations,

and market structure; we briefly review these in the following.

- **Data cleaning**

High-frequency data are often plagued by errors. Therefore it is very often necessary to perform a few preliminary steps to clean the data. A typical problem is the identification and removal of outliers. Such a cleaning procedure is quite problematic, given the arbitrariness in the definition of outliers. Typically, one considers a window of neighboring records, called the *filtering window*, and judges the plausibility of a record according to the statistical properties of the other records in the filtering window [2, 6].

- **Irregular temporal spacing**

Events such as trades, quotes, and limit order placement do not occur at regular intervals of time either in tick-by-tick data or in limit order-book data; rather they occur at irregularly spaced time intervals. In this framework, the time between events is itself a random variable with often nontrivial temporal correlations. Moreover, there are statistical dependencies between the time interval between two trades and the variables characterizing them. Figure 2 shows an example of the occurrence of irregular time spacing in a financial time series.

There are two possible approaches to the problem of irregular temporal spacing. In the first approach, one extracts a regularly spaced time series by using some kind of interpolation scheme. For example, the *previous tick interpolation* scheme uses the past most recent record as the value of the variable at a given instant of time. Alternatively, one can use a linear interpolation between the past most recent and the future most recent events to assign a value to a variable [6]. Interpolation schemes may introduce spurious correlations. The second approach makes use of models that explicitly take into account the irregular temporal spacing. Point processes [3] give a natural description of irregularly spaced events. There are several classes of models dealing with this issue, including the autoregressive conditional duration model [4] and the continuous time random walk model [13].

- **Discreteness**

When one starts dealing with events at the highest resolution level, the issue of variable discreteness cannot be neglected. For example, asset prices can only assume integer multiples of the tick size, which is the smallest allowable price change. Similarly, the volume of an order must be integer multiples of the minimum allowable number of shares in a single order. The discretization unit can also be variable in time. For example, in the last years US equities passed from a large tick

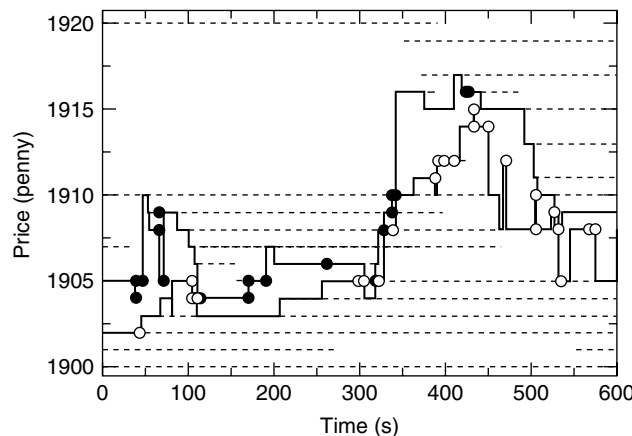


Figure 2 Graphical representation of the dynamics of the order book of the stock AstraZeneca traded at the London Stock Exchange during a 10-min interval on September 4, 2002. The filled (empty) circles are buyer- (seller-) initiated transactions. The thick lines are the bid and the ask price. The dashed lines are limit orders

size (1/8 of a dollar) to a small tick size (one cent, since 2001). The reduction of tick size has been related to several empirical facts, such as the decrease of the spreads and depths of the limit order book [7]. Many statistical models treat the variables as continuous variables, whereas from the above discussion, it is clear that at very high frequency, the variables should be treated as discrete.

- **Variable definition**

The availability of data with the resolution of quotes or of the order book raises the problem of the proper definition of variables. As an example, consider the price of an asset. At a given instant of time, there are many prices that can be considered as the relevant representative price. The most common one is the midprice, which is the mean of the best bid and the best ask prices. However, significantly different volumes may be available at the bid and at the ask, and one may wonder whether a volume-weighted average between the bid and the ask prices is a more proper definition

of price. If one wants to include not only the information on the volume at the best price but also the information of volume in part or in the entire order book, then the number of possible definitions of average price increases significantly.

- **Daily periodicity**

Financial time series at high frequency display strong daily periodicity. Among the variables having a daily periodicity, we mention volatility, trading volume, number of transactions, time between trades, and spread. The daily periodicity is due to the fact that the trading activity is not constant throughout the day. Typically, the trading activity is high around the opening, it decreases until midday, and finally increases toward the end of the trading day. This is sometimes referred to as a *U-shape pattern*. An example is shown in Figure 3. Figure 3(a)–(c) shows the mean intraday pattern for volume, volatility, and number of trades for General Electric at the NYSE. The volatility proxy is the absolute price

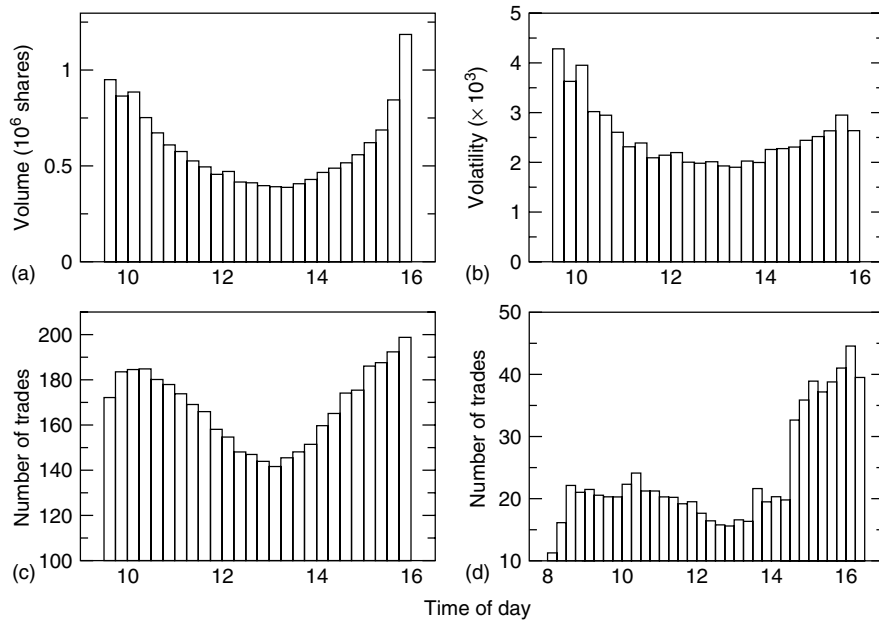


Figure 3 Examples of the intraday pattern. One bin corresponds to 15 min and the x -axis shows the hour of the day. The figure shows the traded volume (a), the absolute price change (b), and the number of trades (c) of General Electric at the NYSE during the period 2002–2003. (d) The number of trades of AstraZeneca exchanged in the electronic market (SET1) of the LSE during the period May 2000–December 2002

change here. The three quantities display a U-shape, even if the detailed shape of the U is quite different. Figure 3(d) shows the mean intraday pattern of the number of trades for AstraZeneca at the LSE. By comparing panel Figure 3(c) and (d), it is evident how the intraday pattern for a variable can be strongly dependent on the considered market. Note that in Figure 3(d) there is almost a doubling of the frequency of trades around 14:30. This is the London time corresponding to the opening of US markets. More generally, in many markets, one often observes peaks of activity or regime changes that correspond to the opening or closing of markets in other time zones.

One important consequence of daily periodicity is that financial time series are not stationary because the statistical properties of a variable depend on the time of the day. Moreover, the presence of the U-shape produces periodicities in financial time series. For these reasons, one is often interested in removing these “trivial” temporal dependencies from the series. This can be done, for example, by dividing the variable by its mean value observed in the sample at that time of the day. It is worth noting that other types of periodicity (e.g., weekly, yearly) are also present. For example, in US markets, the trading volume is larger at the middle of the week than at beginning and at the end. Usually, these periodicities are less intense than the daily periodicity.

Foreign exchange (FX) markets display different types of periodicity [6]. In fact, FX markets have no trading-hour limitations, which means that new bid–ask prices can arrive from all over the world at any time. The intraday patterns of FX variables are therefore not U-shaped. Typically, FX intraday patterns show three broad peaks that correspond to the working periods in America, Europe, and East Asia. Other seasonalities are observable in the FX data, such as the intraweek patterns that arise in connection with the fact that during the weekends the market activity is much smaller.

• Temporal correlations

Typically, financial variables sampled at high frequency display strong temporal correlations. Some temporal correlations are specific to the microstructure of financial markets at high frequency [9]. Other correlations are also present at a lower frequency but are more intense with intraday data. For example, Figure 4 shows the autocorrelation function of transaction price returns for the AstraZeneca (AZN) stock traded at LSE. The lags on the horizontal line are measured in event time. The autocorrelation function of transaction price returns is strongly negative at the first lag and then it rapidly decreases to zero. This is the well-known bid–ask bounce [15] and is mainly due to the presence of two trading

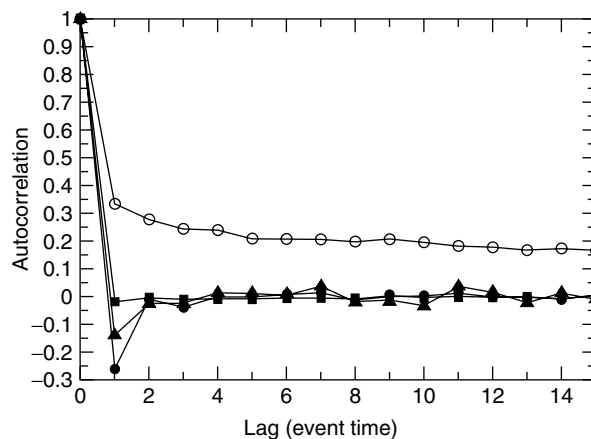


Figure 4 Autocorrelation function of transaction price returns (solid circles), absolute value of transaction price returns (open circles), returns of midprice sampled before each transaction (filled triangles), and returns of midprice sampled before each market event (filled squares) for the AstraZeneca (AZN) stock traded at LSE in 2002. The lags on the horizontal line are measured in event time

prices, one for buyer- and one for seller-initiated transactions. This negative autocorrelation disappears when one considers aggregate returns, and it is therefore a typical microstructural effect. Figure 4 shows also the autocorrelation function of the absolute value of trade price returns, considered here as a proxy of volatility. This autocorrelation function decays very slowly showing that persistency and volatility clustering can also be observed at the tick-by-tick timescale. It is important to stress that at this scale the sampling procedure plays an important role. As an example, Figure 4 shows the autocorrelation function of returns of the midprice sampled before each transaction and sampled before each market event (including limit order placements and cancellations). In the first case, a negative correlation at one lag is still present (though smaller than the one for trade price returns), whereas in the second case the negative correlation disappears. Finally, many financial variables at high frequency display strong temporal persistence that can often be modeled in terms of long memory processes. Examples include transaction rate, spread, sign of the order flow, and volume (*see Order Flow*).

• Market structure

When one works with high-frequency data, it is important to take into account the structure of the market under investigation. Many financial markets have a multiple structure composed of electronic markets, block markets, dark pools, and so on. The statistical properties of a financial variable can be quite different in different segments of the market. For example, trading volume distribution is fat tailed in the block market while it is usually thin tailed in the electronic segment [11]. Therefore, in the analysis of high-frequency data, it is important to take into account this market heterogeneity either by specializing the investigation to a given segment or by weighting the different statistical properties in a suitable way.

References

- [1] Biais, B., Hillion, P. & Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the Paris Bourse, *Journal of Finance* **50**, 1665–1690.
- [2] Brownlees, C.T. & Gallo, G.M. (2006). Financial econometric analysis at ultra-high frequency: data handling concerns, *Computational Statistics and Data Analysis* **51**, 2232–2245.
- [3] Cox, D.R. & Isham, V. (1980). *Point Processes*, Chapman & Hall.
- [4] Engle, R. & Russell, J. (1998). Autoregressive conditional duration: a new model for irregularly spaced data, *Econometrica* **66**, 1127–1162.
- [5] Engle, R.F. & Russell, J.R. (2006). Analysis of high frequency data, in *Handbook of Financial Econometrics*, Y. Ait-Sahalia & L.P. Hansen, eds, North-Holland.
- [6] Gençay, R., Dacorogna, M., Muller, U.A., Pictet, O. & Olsen, R. (2001). *An Introduction to High-Frequency Finance*, Academic Press.
- [7] Goldstein, M.A. & Kavajecz, K.A. (2000). Eighths, sixteenths, and market depth: changes in tick size and liquidity provision on the NYSE, *Journal of Financial Economics* **56**, 125–149.
- [8] Goodhart, C.A.E. & O'Hara, M. (1997). High frequency data in financial markets: issues and applications, *Journal of Empirical Finance* **4**, 73–114.
- [9] Hasbrouck, J. (2007). *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*, Oxford University Press.
- [10] Lee, C.M. & Ready, M.J. (1991). Inferring trade direction from intraday data, *Journal of Finance* **46**, 733–746.
- [11] Lillo, F., Mike, S. & Farmer, J.D. (2005). Theory for long memory in supply and demand, *Physical Review E* **71**, 066122.
- [12] Lillo, F., Moro, E., Vaglica, G. & Mantegna, R.N. (2008). Specialization and herding behavior of trading firms in a financial market, *New Journal of Physics* **10**, 043019.
- [13] Montroll, E.W. & Weiss, G.H. (1965). Random walks on lattices. II, *Journal of Mathematical Physics* **6**, 167–181.
- [14] Odders-White, E. (2000). On the occurrence and consequence of inaccurate trade classification, *Journal of Financial Markets* **3**, 259–286.
- [15] Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market, *Journal of Finance* **39**, 1127–1139.
- [16] Wood, R.A., McInish, T.H. & Ord, J.K. (1985). An investigation of transaction data for the NYSE stocks, *Journal of Finance* **40**, 723–741.

Related Articles

Automated Trading; Limit Order Markets; Market Microstructure Effects; Order Flow; Order Types; Price Impact.

FABRIZIO LILLO & SALVATORE MICCICHÈ

Order Flow

According to a common definition, “order flow is transaction volume with a sign” [1]. The sign of the volume conveys information on the initiator of the transaction. Conventionally, positive volumes are associated with buyer-initiated transactions, whereas negative volumes are associated with seller-initiated transactions. Under this definition, order flow is characterized only by orders that result in an immediate transaction. A more general definition of order flow should also include orders that do not result in an immediate transaction. More precisely, most modern financial markets work through a double auction mechanism. In these markets, impatient traders submit *market orders*, which are orders to buy or sell a given quantity of shares at the best available price,^a and are therefore immediately executed (*see Order Types*). More patient traders submit *limit orders*, which are orders to buy or sell a given amount of shares at a given price or better. If there are no counterparties willing to trade at this price, the order is added to a queue called the *limit order book*. A trader can decide to remove his/her limit order from the book whenever he/she wants and this event is called a *cancellation*. The order flow is the joint process of market and limit order placements and cancellations. The price formation process results from the interaction between these events and is the focus of market microstructure theory [1–3]. Order flow is related to supply and demand and to liquidity. More precisely, when one considers the sign of the orders, “Order flow, as used in microstructure finance, is a variant of a key term in economics excess demand” [1]. Whatever the order sign, the flow of limit orders increases the liquidity of the asset, while market orders and cancellations deplete liquidity.

Statistical Properties

The order flow is a complex stochastic process that takes place in continuous time. It has three components (limit orders, market orders, and cancellations), each component being characterized by a sign, a volume, and a price. The order flow is, furthermore, strongly dependent on the state of the limit order book. There is no available stochastic model able to

fully describe the dependence structure of this process. One usually restricts to a subpart of the order flow and attempts to describe it within a suitable reduced model.

In order to simplify the presentation, this article focuses on the statistical modeling of each component separately and refers mainly to stock markets.

Market Order Flow

The unconditional probability of the sign of a market order is very close to 1/2. The probability density function of the volume of market orders is often described by a decreasing function with a power-law tail (see Figure 1). The market order flow has very interesting temporal correlation properties. Consider, for simplicity, the symbolic time series obtained in event time by replacing buy market orders with +1 and sell market orders with −1, irrespective of the volume of the order. The autocorrelation function $C(\tau)$ of these strings of bits decays very slowly to zero and its asymptotic behavior is well fitted by a power-law function

$$C(\tau) \approx \frac{1}{\tau^\alpha} \quad (1)$$

where $\alpha < 1$ (typically $\alpha \cong 0.5$) [4, 5]. A process with such a nonsummable autocorrelation is called a *long-memory process* [6]. Panel (a) of Figure 2 shows the autocorrelation function of market order signs for AstraZeneca (AZN) traded at the London Stock Exchange (LSE). It is worth noting that the persistence in the market order flow is still statistically significant after many trading hours. The volume-weighted market order flow also shows strong temporal persistency, although weakened by the volume fluctuations (see Figure 2b). The long-memory properties of market order signs have been observed in different markets and can also be observed after temporal aggregation of the order flow. It has been proposed that the long memory is due to order splitting [7].

The persistence of market order flow raises the problem of how market efficiency is maintained. A positively correlated order flow and a fixed and permanent market impact imply positively correlated return time series, in contradiction with market efficiency. To reconcile a correlated order flow with efficiency, one has either to assume a fixed but temporary

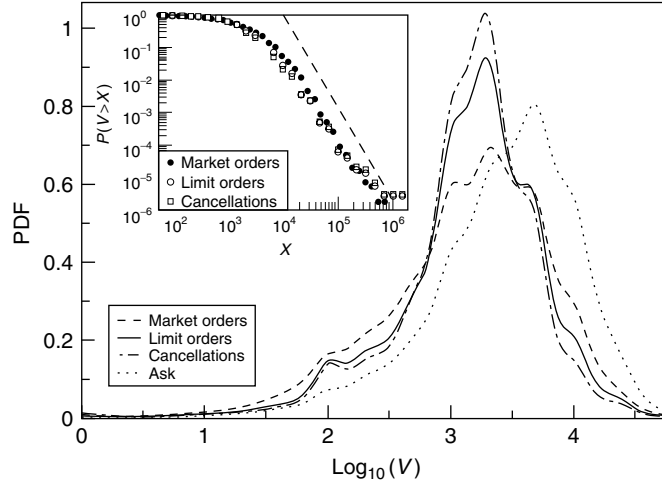


Figure 1 Probability density function of the log-volume of individual orders, for the three different components of the order flow. For comparison, the figure also shows the probability density function of decimal logarithm of the volume at the best (ask) price, sampled before each transaction. The inset shows in a double logarithmic scale the cumulative probability distribution of the volume for different types of orders. The dashed line is a power-law function with slope 2.8, which is the maximum likelihood estimator of the tail exponent for market orders. Data refer to the stock AZN traded at the LSE in the period 1999–2002

impact or a permanent but history-dependent impact [8] (*see Price Impact*).

The conditional properties of market order flow are also important. First, the size of the bid–ask spread is a key determinant of the market order flow. The rate of market orders submission, independently on their sign, decreases as the spread increases (see Figure 3). This behavior is intuitively expected since a large bid–ask spread is a strong disincentive to trade, given that the spread-related cost is large. A second conditional property concerns the volume of a market order, which is strongly dependent on the volume at the opposite best price. Specifically, it is rare that the volume of a market order exceeds the volume at the opposite best and thus that the trade penetrates more than one price level [9]. For example, in a set of 16 highly capitalized stocks traded at the LSE, approximately 86% of the market orders leading to an immediate price change have a volume exactly equal to the volume at the opposite best. Moreover, approximately 97% of the market orders leading to an immediate price change have a volume that is less than the total volume at the best and at the second-best opposite price. This indicates strong selective liquidity taking, meaning that agents condition the size of their market orders on

prevailing liquidity, making large transactions only when liquidity is high and small transactions when it is low.

Limit Order Flow

A limit order is characterized by its sign, volume, and limit price. Several statistical regularities of limit orders placement have also been observed. First, the unconditional distribution of limit order volume is similar to the one for market orders (see Figure 1). Second, the time series of limit order signs is also characterized by the long-memory property [5, 10]. Figure 2(c) shows the autocorrelation function of limit order signs for AZN. Third, several statistical regularities of limit order flow are observed when one also considers the limit price. The limit price characterizes the aggressiveness of the limit order and the patience of the trader since the probability that the order is executed in a given time interval depends on where it is placed. Limit orders in the spread are very aggressive and more likely to be executed, while limit orders placed in the book very far from the best price are less aggressive and their execution probability is low. For high cap stocks at LSE between 1999 and 2002, roughly one-third of

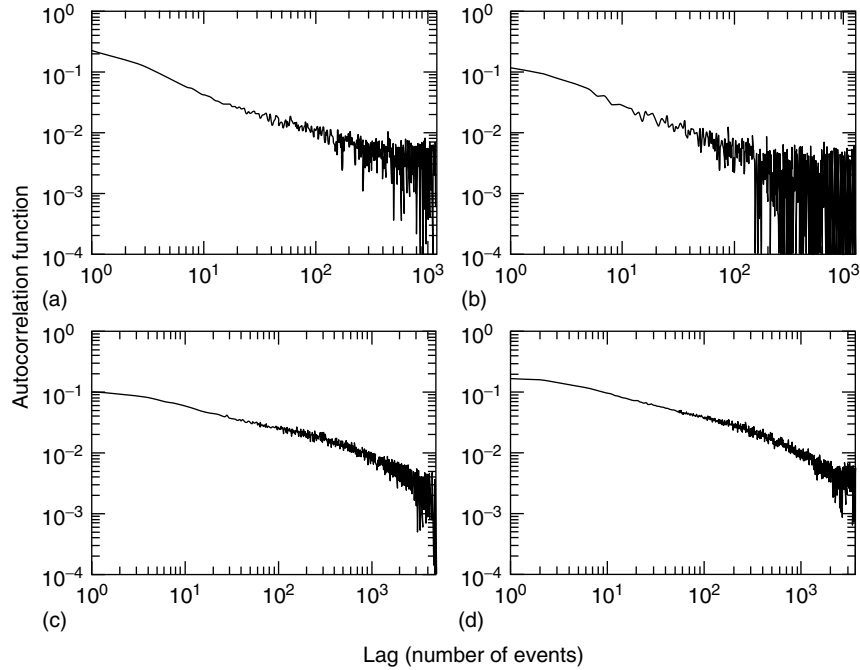


Figure 2 Autocorrelation of the order flow in event time plotted on a double logarithmic scale. Specifically, the panels show the autocorrelation function of market order sign (a), market order signed volume (b), limit order sign (c), and limit order signed volume (d). The span of the x -axis corresponds on average to two trading days. The data refer to the stock AZN traded at the LSE in the period 1999–2002

limit orders were placed inside the spread, another third at the best prices, and the last third inside the book. One of the statistical regularities recently observed is the power-law distribution of limit order price in continuous double-auction financial markets. Let $b(t) - \Delta$ denote the price of a new buy limit order and $a(t) + \Delta$ the price of a new sell limit order. Here $a(t)$ is the best ask price and $b(t)$ is the best bid price when the limit order is placed. The probability density function $\rho(\Delta)$ of the relative price Δ is very similar for buy and sell orders. Moreover, for large values of Δ , the probability density function is well fitted by a power-law function

$$\rho(\Delta) \approx \frac{1}{\Delta^{1+\mu}} \quad (2)$$

where $\mu \approx 1.5$ for LSE stocks [11], and $\mu \approx 0.6$ for stocks traded at the Paris Stock Exchange [12]. This power law extends from 1 tick to over 100 ticks (sometimes even 1000 ticks), corresponding to a relative change of price of 5 to 50%. Such a broad distribution of limit order prices tells us that the

opinion of market participants about the price of the stock in a near future could be anything from its present value to 50% above or below this value. Mike and Farmer [13] fit the distribution of relative limit prices Δ for LSE stocks with a Student distribution with 1.3 degrees of freedom corresponding to a value $\mu = 1.3$. Interestingly, the Student distribution fits relative limit order price both outside ($\Delta > 0$) and inside ($\Delta < 0$) the spread.

The rate of limit order placement strongly depends on spread. When the spread is large, the rate of limit order placement is also large (see Figure 3). This is true also if one considers only limit orders placed inside the spread (see Figure 3). The reason is that a large spread is an incentive to provide liquidity and to acquire price priority by placing a limit order in the spread. The limit order flow is also dependent on book thickness. Patient traders become more aggressive in their limit order submission when the own side book is thicker and when the opposite side book is thinner [14].

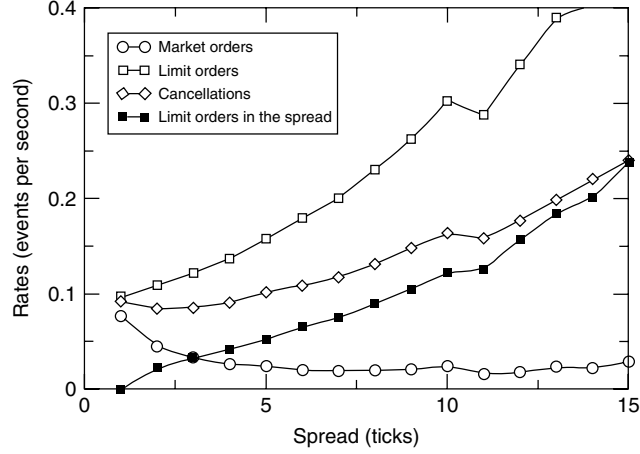


Figure 3 Rates of different types of events in the limit order book conditional to the size of the spread. The data refer to the stock AZN traded at the LSE in 2002 [Adapted from [20].]

Cancellation Flow

A limit order can be canceled or can expire and both events lead to a depletion of liquidity. Not surprisingly, the volume of canceled limit orders is distributed in a similar way to the volume of limit orders (see Figure 1). The time series of cancellation signs also displays the long-memory property (see Figure 2d). Moreover, the rate of cancellations increases with spread size (see Figure 3). This is probably due to the high limit order placement/cancellation activity observed when a temporary liquidity crisis triggers a competition between liquidity providers in order to find a new equilibrium value for the bid and ask prices. Finally, the lifetime of a given limit order (which is inversely related to the cancellation probability) increases as one moves away from the bid–ask spread [15]. This phenomenon can be explained by considering that far away orders are typically put in the market by patient investors that want to benefit from significant price swings in the medium term, while orders close to the bid and ask prices correspond to very active market participants that readjust their orders at a very high frequency. The cancellation probability at time t depends also on the ratio between the distance Δ of the limit price from the best at time t and at the time $t_0 < t$ when the order was placed. Mike and Farmer fit the cancellation probability with the functional form

$$p(t) \propto 1 - \exp[-\Delta(t)/\Delta(t_0)] \quad (3)$$

meaning that if the price moves in a direction opposite to the limit order, its cancellation probability increases [13].

Diagonal Effect

As discussed above, a common phenomenon observed in the statistics of order flow is a strong serial correlation of each component of the order flow. This phenomenon has been termed *diagonal effect* [16]. More precisely, the probability that a given type of event occurs is larger after that event has just occurred than it would be for the unconditional occurrence of the event [16]. This type of serial persistence is observed also for longer lags and it has been observed in different markets including foreign exchange markets [17]. The possible origins of these serial correlations are (i) strategic order splitting in order to reduce market impact; (ii) imitation behavior across traders that observe other participants and imitate their behavior; and (iii) traders reacting similarly to the same event. It is, in general, difficult to separate these alternative explanations. However, by using brokerage data, it has been shown that correlation in market order flow is mainly explained by order splitting [8]. Positive serial correlation of limit order flow is also due largely to order splitting and the undercutting behavior among limit order traders. Controlling for these two effects explains up to 72% of the positive serial correlation [10].

Commonality in Order Flow and Liquidity Provision

The above discussion has focused on the statistical properties of the order flow for an individual stock. However, the order flow of different stocks simultaneously traded in the same market can display mutual dependencies. Since the order flow determines the supply and demand of liquidity of a stock, this type of dependency is also called *commonality in liquidity*. When order flow is signed according to the buy/sell side, direct and indirect measures indicate that commonality in the order flows explains a consistent part (two-thirds as described in [18]) of the commonality in returns [18, 19]. By considering order flow with orders signed according to the liquidity supply or demand, empirical analysis shows that there are significant commonalities in liquidity and that a liquidity market factor exists [18, 19].

End Notes

^aIn some markets only limit orders are allowed and a limit order crossing the spread results in an immediate transaction. Thus a crossing limit order behaves as an effective market order.

References

- [1] Lyons, R.K. (2001). *The Microstructure Approach to Exchange Rates*, The MIT Press.
- [2] O'Hara, M. (1995). *Market Microstructure Theory*, Blackwell Publishing.
- [3] Hasbrouck, J. (2007). *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*, Oxford University Press.
- [4] Bouchaud, J.P., Gefen, Y., Potters, M. & Wyart, M. (2004). Fluctuations and response in financial markets: the subtle nature of "random" price changes, *Quantitative Finance* **4**, 176–190.
- [5] Lillo, F. & Farmer, J.D. (2004). The long memory of the efficient market, *Studies in Nonlinear Dynamics and Econometrics* **8**, 1.
- [6] Beran, J. (1994). *Statistics for Long-Memory Processes*, Chapman & Hall.
- [7] Lillo, F., Mike, S. & Farmer, J.D. (2005). Theory for long memory in supply and demand, *Physical Review E* **71**, 066122.
- [8] Bouchaud, J.P., Farmer, J.D. & Lillo, F. (2009). How markets slowly digest changes in supply and demand, in *Handbook of Finance: Dynamics and Evolution of Financial Markets*, K. Schenke-Hoppe & T. Hens, eds, North-Holland. Available at SSRN: <http://ssrn.com/abstract=1266681>.
- [9] Farmer, J.D., Gillemot, L., Lillo, F., Mike, S. & Sen, A. (2004). What really causes large price changes? *Quantitative Finance* **4**, 383–397.
- [10] Yeo, Y. (2006). *Persistence in Limit Order Flow*. Available at SSRN: <http://ssrn.com/abstract=923022>.
- [11] Zovko, I. & Farmer, J.D. (2002). The power of patience; a behavioral regularity in limit order placement, *Quantitative Finance* **2**, 387–392.
- [12] Bouchaud, J.P., Mézard, M. & Potters, M. (2002). Statistical properties of the stock order books: empirical results and models, *Quantitative Finance* **2**, 251–256.
- [13] Mike, S. & Farmer, J.D. (2008). An empirical behavioral model of liquidity and volatility, *Journal of Economic Dynamics and Control* **32**, 200–234.
- [14] Rinaldo, A. (2004). Order aggressiveness in limit order book markets, *Journal of Financial Markets* **7**, 53–74.
- [15] Potters, M. & Bouchaud, J.P. (2003). More statistical properties of order books and price impact, *Physica A* **324**, 133–140.
- [16] Biais, B., Hillion, P. & Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the Paris Bourse, *Journal of Finance* **50**, 1665–1690.
- [17] Danielsson, J. & Payne, R.G. (2001). Measuring and explaining liquidity on an electronic limit order book: evidence from Reuters D2000-2), *EFA 2001 Barcelona Meetings*. Barcelona. Available at SSRN: <http://ssrn.com/abstract=276541>.
- [18] Hasbrouck, J. & Seppi, D.J. (2001). Common factors in prices, order flows, and liquidity, *Journal of Financial Economics* **59**, 383–411.
- [19] Domowitz, I., Hansch, O. & Wang, X. (2005). Liquidity commonality and return co-movement, *Journal of Financial Markets* **8**, 351–376.
- [20] Ponzi, A., Lillo, F. & Mantegna, R.N. (2009). Market reaction to a bid-ask spread change: a power-law relaxation dynamics, *Physical Review E* **80**, 016112.

Related Articles

Algorithmic Trading; Bid-Ask Spreads; High-frequency Data; Inventory Effects; Limit Order Markets; Liquidity; Market Microstructure Effects; Order Types; Price Impact.

FABRIZIO LILLO

Glosten–Milgrom Models

Glosten–Milgrom models analyze transaction prices arising from quotes of competing risk-neutral dealers making a market in a single security and facing both privately informed and uninformed traders. The dealers must quote bid and offer prices at which they are willing to, respectively, buy and sell. A trader arrives and decides to buy, sell, or leave. Upon seeing the outcome of the trader’s decision, the dealers modify their quotes in preparation for the next arrival. Implementation of the model thus requires a specification of the arrival processes of informed and uninformed traders and the nature of public and private information. This specification, in turn, leads to the calculation of quotes and the time series of quotes and transaction prices.

Mathematical Structure and Results

The discussion follows the analysis in [7], and proofs can be found there. The model is designed to describe trade over a relatively small interval of time in which there is private information about the future cash flows of the security, but this private information will depreciate, either through information revealed in trade or *via* public announcements of the information. By the time all asymmetric information has been resolved, at T , the value of a share of the security will be V . Prior to T , this value is not known, but at time t , there has been a history of trade and public announcements. This information is summarized by the filtration H_t . Justified by the fact that a relatively short period of time is being analyzed, the time value of money and risk aversion can be reasonably ignored, and the value of a share of stock at time t is just its expected value conditional on time t information, $E[V|H_t]$.

During the trading period, potential traders arrive at the market one by one. An arriving trader at time t observes the bid and ask quotes B_t and A_t , respectively. If the trader is informed, with information represented by F_t , a refinement of the history H_t , then the trader is presumed to buy if $E[V|F_t]$ exceeds the ask and sell, if this expectation is below the bid. Otherwise, the trader does not trade. A trader who is uninformed is presumed to have a value M_t , which could be a function of the history H_t

as well as other features of the trader’s preferences. This private valuation should be thought of as a random variable independent of private information. The uninformed trader will buy if M_t exceeds the ask price, and sell if this idiosyncratic valuation is below the bid. Describing the arrival process by an indicator variable I_t , which is one if an arrival at time t is informed and zero if uninformed, we can describe the personal valuation of an arrival at time t by $Z_t = I_t E[V|F_t] + (1 - I_t)M_t$.

Consider the quoting problem of a market maker at time t facing a possible arrival. Should the arrival buy, the market maker will receive the ask price and give up a share of stock ultimately worth V . This will happen if Z_t exceeds the ask price. Similarly, if the arrival sells at the bid, the market maker will receive a share ultimately worth V and pay the bid price. Otherwise, there will be no transaction and no profit. Thus, the expected profit to a market maker quoting bid B and ask A , is

$$E[(A - V)\mathbf{1}\{Z_t > A\} + (V - B)\mathbf{1}\{Z_t < B\}|H_t] \quad (1)$$

Assuming that dealer competition drives the expected profit for both a buy and a sell to zero, the ask and bid quotes are the solutions, if they exist, to the two fixed-point problems:

$$\begin{aligned} A_t &= E[V|H_t, Z_t > A_t] \quad \text{and} \\ B_t &= E[V|H_t, Z_t < B_t] \end{aligned} \quad (2)$$

If solutions do not exist, the market is said to close down.

The zero-profit bid and ask prices are “no regret” prices in the sense that the market maker, after having sold to an order at the ask price will revise expectations and the revised expectation is precisely the ask price. That is, to quote the ask price the dealer must determine what the revised expectations will be in the event of a purchase.

The first result is immediate and intuitive: the ask price exceeds the expectation at the time of the quote that exceeds the bid. If a buy occurs, one possibility is that the arriving trader is informed and has a valuation higher than the ask price. It could never be the case that the arriving trader is informed and has a valuation less than the ask price and buys. This alone requires that the market increase the conditional expected value in response to a buy. This argument justifies

the term *adverse selection spread* as a description of the difference between the ask price and bid in the model (see **Adverse Selection**).

The second observation follows from the fact that transaction prices occur at the “no regret” quotes. The transaction price is a conditional expectation of the terminal value, V . A transaction at time t can be written, where H_{t+} indicates the information available following a transaction at time t :

$$\begin{aligned} P_t &= A_t \mathbf{1}\{Z_t > A_t\} + B_t \mathbf{1}\{Z_t < B_t\} \\ &= E[V|H_t, Z_t > A_t] \mathbf{1}\{Z_t > A_t\} \\ &\quad + E[V|H_t, Z_t < B_t] \mathbf{1}\{Z_t < B_t\} \\ &= E[V|H_{t+}] \end{aligned} \quad (3)$$

Thus, transaction prices form a martingale. From this observation, flow a number of immediate conclusions. Since reasonably behaved martingales such as this one converge, the absolute differences between transaction prices must get small. As long as trade continues, it must be the case that the spread gets small and that the informational difference between the informed traders and the market gets small.

A number of comparative statics results are readily obtained. At a point in time, an increase in the probability of informed arrival or an increase in the precision of private information increases the spread. As the elasticity of uninformed demand and supply increases, the spread increases as well. It is important to note, however, that these results hold only at a point in time—when the probability of informed trade increases, the rate at which trade reveals private information increases reducing spreads in the future.

The uninformed valuation plays a crucial role in the model. If the uninformed valuation varies little from the mean, then it is quite possible that the only bid and ask prices satisfying the condition are, respectively, the lowest and highest possible realizations of V . In this case, there is no trade and the market shuts down. On the other hand, if there is significant variation, so that the uninformed demand and supply are inelastic, then interior solutions for the quotes are guaranteed.

Canonical Example

The most common implementation of a Glosten–Milgrom model assumes that uninformed demand

and supply are perfectly inelastic and equally likely, the terminal value V has a two-point distribution and informed traders know what value of V will be realized. Without losing much generality, the two possible values of V can be taken to be zero and one. Define α to be the probability that an informed trader arrives in any time period. If p is the probability, conditional on past history, that the realization of V is one, then the conditional probability of a buy is $\alpha p + 0.5(1 - \alpha)$ and the conditional probability of a sell is $\alpha(1 - p) + 0.5(1 - \alpha)$. The ask and bid prices are thus

$$\begin{aligned} A &= P\{V = 1|\text{buy}\} = P\{\text{buy}|V = 1\}p / P\{\text{buy}\} \\ &= \frac{(\alpha + 0.5(1 - \alpha))p}{\alpha p + 0.5(1 - \alpha)} \\ B &= P\{V = 1|\text{sell}\} = \frac{0.5(1 - \alpha)p}{\alpha(1 - p) + 0.5(1 - \alpha)} \end{aligned} \quad (4)$$

The spread is thus $\alpha p(1 - p) / (P\{\text{buy}\}P\{\text{sell}\})$. Ignoring the denominator, which is approximately 0.25, the spread is increasing in the unresolved variance of V , $p(1 - p)$, and the probability of informed trade, α .

This model has particularly tractable transaction price dynamics. Recalling that transaction prices are updated expectations, it is readily seen that, denoting by p^+ the updated probability, and letting Q be a trade direction indicator (+1 for buy at ask and -1 for sell at bid), the following holds:

$$\frac{p^+}{1 - p^+} = \left(\frac{p}{1 - p} \right) \left(\frac{1 + \alpha}{1 - \alpha} \right)^Q \quad (5)$$

Assuming that, initially, it is equally likely that V be high or low, and denoting by p_t the revised probability that V is one, after t transactions, it is immediate that

$$\ln \left(\frac{p_t}{1 - p_t} \right) = \ln \left(\frac{1 + \alpha}{1 - \alpha} \right) \{Q_1 + \dots + Q_t\} \quad (6)$$

Conditional on V , the log odds ratio is proportional to a random walk with expected increment $\alpha(2V - 1)$. The larger the probability of informed trade is, the faster the log odds ratio increases in the case $V = 1$ and decreases in the case $V = 0$. The larger α is, the larger is the initial spread, but the faster the information gets impounded in prices and the faster the spread will fall.

In this model, expectations will not converge to one or zero in finite time. However, the convergence to almost one or zero can be quite fast. Let T_N be the first time that the log odds ratio hits N in the case $V = 1$ and $-N$ in the case $V = 0$, starting from $p_0 = 0.5$. Straightforward calculations show that the expectation of this stopping time is $N/[\alpha \ln((1 + \alpha)/(1 - \alpha))]$. For example, if the threshold odds ratio is 0.99/0.01, then N is less than 5. Setting $N = 5$, and letting α be 0.2, the expected first passage time is roughly 62 trades. The spread at these threshold odds and with α equal to 0.2 is 0.008. Setting α to 0.1 increases the expected number of trades to 250. The spread at these threshold odds and $\alpha = 0.1$ is 0.004. Thus, a stock with even moderate trading will reveal private information relatively rapidly, and the spread will decline quickly.

Relating the expected first passage time and the spread at the first passage time indicates an interesting trade-off. As the percentage of informed trading goes up, the initial spread increases, yet the speed with which the spread declines goes up as well. For $\alpha = 0.2$, the spread starts out at 0.2, but declines rapidly, becoming near zero with fewer than 100 expected transactions. In contrast, when $\alpha = 0.05$, the initial spread is 0.05, yet it takes an expected 300 trades to become minimal.

This raises intriguing questions about the welfare effects of informed trading; but these are not issues that this example of a Glosten–Milgrom model can handle. Since the uninformed trade, no matter what the spread, informed profit is merely uninformed loss and the total welfare is unaffected by how much informed trade there is. The spread is often related to a welfare measure, but this requires a questionable value judgment—informed are not entitled to their profits. To measure the welfare loss associated with informed trade, a model must incorporate elastic trade on the part of the uninformed.

Example with Elastic Uninformed Trade

The literature does not offer an analysis of transaction price dynamics with elastic uninformed trade. The following is a tractable model. As in the model above, the informed arrive with probability α , and know the terminal value. An uninformed trader arrives with private valuation, M , and buys if M exceeds the ask, sells if M is below the bid, and leaves otherwise.

Market makers do not know the private valuation, but know that if the current estimate of the security's value is p , then an uninformed trader's valuation has the following distribution:

$$P\{M < t\} = \frac{t(1 - p)}{t(1 - p) + p(1 - t)}; \quad 0 < t < 1, \quad 0 < p < 1 \quad (7)$$

In this case, the probability of a buy, when the current valuation is p and the offer is A is $\alpha p + (1 - \alpha)P\{M > A\}$. The probability of a buy given that V is one is $\alpha + (1 - \alpha)P\{M > A\}$ where $P\{M > A\} = p(1 - A)/[A(1 - p) + p(1 - A)]$ from equation (7). The ask, A , is the solution to

$$A = \frac{p(\alpha + (1 - \alpha)P\{M > A\})}{\alpha p + (1 - \alpha)P\{M > A\}} \quad (8)$$

$A = 1$ is one solution, but this will lead to no trade. Another solution is $A = p/(1 - 2\alpha(1 - p))$, as long as this is less than one and greater than zero, which is true if and only if $\alpha < 0.5$. Similarly, on the bid side, $B = p(1 - 2\alpha)/(1 - 2\alpha p)$ is the solution as long as $\alpha < 0.5$. If α exceeds 0.5, the market does not open since $A = 1$ and $B = 0$ are the only feasible solutions.

Because of assuming elastic uninformed trade, the initial spread is larger by the factor $1/(1 - \alpha)$. As the dealer increases the spread, he or she loses less to the informed, per trade, but as uninformed with less-extreme valuations drop out, the probability of transacting with an informed trader increases. This requires a greater widening of the spread.

In this model, not every arrival leads to a trade. The probability that an uninformed trader arrives and decides not to trade is α , precisely the probability of an informed trade. This model also has convenient dynamics. Conditional on V , the log odds ratio is $-\ln(1 - 2\alpha)$ times a random walk with drift $(2V - 1)\alpha$.

Assumptions and Extensions

One feature of the model leads to its immediate refutation—trades are for a single quantity. Several papers have extended the basic analysis to multiple trade sizes. Two possible trade sizes is considered by Easley and O'Hara [3], while Glosten [5] considers a continuum of trade sizes. Glosten [6] analyzes

continuous trade sizes in an electronic limit order book (see **Limit Order Markets**). Interestingly, there is evidence in [8] that the number of trades, as opposed to quantity is what contributes to volatility. Furthermore, [1] shows that a Kyle model ([9]; also see **Kyle Model**) with a continuum of allowable trades can, under some conditions, be thought of as a temporal aggregation of a dynamic Glosten–Milgrom model with a strategic informed trader.

One unreasonable assumption of the model is that the bid and offer are determined by a zero-profit assumption. The assumption implies that there is no compensation for the time and effort market making requires. The reason for the assumption is that the analysis can then focus on only the informational source of the spread. It is easy to extend the model to include compensation for market-maker opportunity (and other) cost by specifying that the bid, for example, satisfy $B = E[V|Z < B] - c$, where c is the economic cost that needs to be covered. With this more realistic specification, it is now apparent that the spread consists of two components—the cost component and the adverse selection component.

When there is no cost component, the martingale property of transaction prices implies that there will be no serial correlation in price changes. When there is only a cost component, however, it is easy to see that transaction prices will bounce back and forth between bid and ask, inducing negative serial correlation in price changes. That the different spread components lead to different transaction price dynamics informs empirical investigations of the spread (see **Adverse Selection**).

A strong assumption of the model is that the existence of informed traders is common knowledge. This is relaxed, first in [4] and then in a series of papers by Easley, O'Hara, and coauthors, in which the *probability of informed trade* (PIN) is theoretically derived and empirically investigated (see **Probability of Informed Trading**).

While the model is described in terms of competing market makers, it is somewhat more general than that. In particular, the model is not necessarily inconsistent with modeling trade on the New York Stock Exchange, or even an electronic limit order book. In the first case, the competing providers of liquidity are the specialist, floor brokers, and limit order submitters. In the second, limit order submitters com-

pete to provide the quote. Significantly, however, the model does not consider the trader choice between using a market or limit order, an important feature of trading on the NYSE and electronic limit order book markets (see **Adverse Selection; Limit Order Markets**).

Two survey papers on market microstructure, [10] and [2], are recommended. The first is stronger on a survey of empirical analyses, while the second is stronger on modeling issues.

References

- [1] Back, K. & Baruch, S. (2004). Information in securities markets: Kyle meets Glosten and Milgrom, *Econometrica* **72**, 433–465.
- [2] Biais, B., Glosten, L.R. & Spatt, C. (2005). Market microstructure: a survey of microfoundations, empirical results, and policy implications, *Journal of Financial Markets* **8**, 217–264.
- [3] Easley, D. & O'Hara, M. (1987). Price, trade size, and information in securities markets, *Journal of Financial Economics* **19**, 69–90.
- [4] Easley, D. & O'Hara, M. (1992). Time and the process of security price adjustment, *Journal of Finance* **47**, 576–605.
- [5] Glosten, L.R. (1989). Insider Trading, liquidity, and the role of the monopolist specialist, *Journal of Business* **62**, 211–236.
- [6] Glosten, L.R. (1994). Is the electronic open limit order book inevitable? *Journal of Finance* **49**, 1127–1161.
- [7] Glosten, L.R. & Milgrom, P.R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* **14**, 71–100.
- [8] Jones, C., Kaul, G. & Lipson, M.L. (1994). Transactions, volume, and volatility, *Review of Financial Studies* **7**, 631–651.
- [9] Kyle, A.S. (1985). Continuous auctions and insider trading, *Econometrica* **53**, 1315–1335.
- [10] Madhavan, A. (2000). Market microstructure, *Journal of Financial Markets* **3**, 205–258.

Related Articles

Adverse Selection; Kyle Model; Limit Order Markets; Probability of Informed Trading.

LAWRENCE R. GLOSTEN

Kyle Model

The Kyle model [13] provides a tractable framework in which liquidity and the informativeness of prices can be measured. It played a key role in the proliferation of research papers on heterogeneous information that occurred in the past two decades.^a The model formalizes the explanation of liquidity given by Treynor [6]—reprinted as [14]—who observes that “the liquidity of a market . . . is inversely related to the average rate of flow of new information . . . and directly related to the volume of liquidity-motivated transactions.” These relationships are captured in “Kyle’s lambda”.

The Kyle model describes the price of a single risky asset over a time period $[0, 1]$. A risk-free asset is used as the numeraire, so the risk-free rate is zero. In the original model, there is a single trader with private information. An announcement is made at date 1 that reveals the information of this trader, and he/she is able to liquidate his/her position at date 1 at the revealed value. The “liquidity-motivated transactions” described by Bagehot [6] are modeled as random trades (by “liquidity traders” or “noise traders”) that are independent of the asset value. The trades of the informed and liquidity traders are “batched” and submitted to risk-neutral competitive market makers.^b The market makers trade on their own accounts to clear the market, and because of competition and risk neutrality, the clearing price is the expected value of the asset, conditional on the information in the order flow.

An important feature of the Kyle model is the strategic behavior of the informed trader, that is, he/she considers the impact his/her orders have on the market price. This assumption avoids the “schizophrenia” of traders who believe they have no impact on market prices yet understand that prices reflect their private information, which, as Hellwig [10] notes, prevails in some competitive rational expectations models.

Single-period Kyle Model

In the single-period model, trading occurs only at date 0. Before trading, the single informed trader observes a normally distributed random variable $\tilde{v}$ with mean $\bar{v}$ and variance σ_v^2 . The informed trader

chooses an order $\tilde{x}$ depending on $\tilde{v}$. Noise traders submit an order $\tilde{z}$ that is independent of $\tilde{v}$ and normally distributed with mean 0 and variance σ_z^2 . Market makers observe $\tilde{y} \equiv \tilde{x} + \tilde{z}$ and set the price equal to $\tilde{p} = E[\tilde{v}|\tilde{y}]$. At date 1, the value $\tilde{v}$ is revealed to the market, and the asset can be traded in arbitrarily large quantities at the price $\tilde{v}$. Thus, the informed trader’s profit is $\tilde{x}(\tilde{v} - \tilde{p})$, and his/her objective is to maximize the expected profit.

There is a unique “linear equilibrium”. Linearity means that $\tilde{p} = \mu + \lambda\tilde{y}$ and $\tilde{x} = \alpha + \beta\tilde{v}$, for constants μ , λ , α , and β . Equilibrium means the following:

- (A) Given α and β , pricing satisfies Bayes’ rule, that is, $\mu + \lambda\tilde{y} = E[\tilde{v}|\tilde{y}]$.
- (B) Given μ and λ , the insider’s strategy is optimal, that is,

$$\alpha + \beta\tilde{v} = \operatorname{argmax}_x E[x(\tilde{v} - \mu - \lambda(x + \tilde{z}))|\tilde{v}] \quad (1)$$

Note that optimality in (B) is among all possible demands x , including nonlinear functions of $\tilde{v}$. Linearity implies that $\tilde{v}$ and $\tilde{y}$ are jointly normal, so $\lambda = \operatorname{cov}(\tilde{v}, \tilde{y})/\operatorname{var}(\tilde{y})$. Using this fact and solving the optimization problem in (B), it is straightforward to calculate that the unique linear equilibrium is given by $\mu = \bar{v}$, $\alpha = -\beta\bar{v}$, $\beta = \sigma_z/\sigma_v$, and $\lambda = \sigma_v/(2\sigma_z)$. The unconditional expected profit of the informed trader is $\sigma_v\sigma_z/2$.

Kyle [13] defines the reciprocal of λ as the “depth” of the market. It measures the number of shares that can be traded, causing only a unit change in the price. The formula for $1/\lambda$ encapsulates Bagehot’s intuition: the depth of the market is proportional to the amount of noise trading as measured by σ_z and inversely proportional to the amount of private information as measured by σ_v .

Dynamic Kyle Model

A trader who recognizes that his/her trades impact the market price will typically want to execute a trade in small pieces, walking up or down the market supply curve. The single-period model gives the informed trader only one opportunity to trade, which is an unnatural constraint. Relaxing this constraint

enables one to examine how the dynamic optimization of the informed trader affects the evolution of liquidity over time. The general conclusion is that the informed trader spreads his/her trades over time in a manner that has the effect of smoothing liquidity intertemporally. We return to this issue in the final paragraph.

Kyle [13] analyzes a discrete-time N -period version of the model, assuming that the noise trades in each period are independent normally distributed variables with variance $\sigma_z^2 \Delta t$, where $\Delta t = 1/N$ is the length of each period. The variance of cumulative noise trades during the period $[0, 1]$ is again σ_z^2 . Kyle shows that the equilibria converge to the equilibrium of a continuous-time model in which the noise trades arrive as a Brownian motion with volatility σ_z .

The continuous-time model affords the informed trader the maximum flexibility in timing his/her trades. In the original continuous-time model, the single informed trader observes the normally distributed variable $\tilde{v}$ at date 0, and the asset trades at $\tilde{v}$ at date 1, as in the single-period model. The asset is traded continuously during $[0, 1]$. The number of shares Z held by noise traders is assumed to be a Brownian motion, independent of $\tilde{v}$, having volatility σ_z . Again, the variance of cumulative noise trades during $[0, 1]$ is σ_z^2 . At each date t , the informed trader chooses a number of shares X_t to hold. This choice depends on $\tilde{v}$ and the history of Z through date t .^c Market makers observe $Y \equiv X + Z$ and set the price P_t to be the expected value of $\tilde{v}$, conditional on the information in Y through date t .

The principal conclusions of the continuous-time model are as follows:

1. The price change at each instant is $dP_t = \lambda dY_t$, where $\lambda = \sigma_v/\sigma_z$. Thus, market depth is constant over time, and the depth of the market is only half of what it is when the informed trader is constrained to trade only once.
2. The price process is a Brownian motion with volatility σ_v . Constancy of the volatility means that information is incorporated into the price at a constant rate. The volatility does not depend on the level of noise trading σ_z .
3. All of the private information is incorporated into the price before the announcement date (the price converges to $\tilde{v}$ by date 1).
4. The expected profit of the informed trader is $\sigma_v\sigma_z$. Thus, the insider's expected profit is twice

what it is when he/she is constrained to trade only once. This implies that the expected losses of noise traders are also twice what they are in the single-period model.

5. The equilibrium informed trades are of order dt . This reflects the fact that the desired trades of a patient anonymous trader are optimally executed in small pieces.

To analyze the informed trader's optimization problem in this model, one must start with some guess about how P depends on Y . Two possible approaches are as follows: (i) conjecture $P_t = H(t, Y_t)$ for some function H and (ii) conjecture $dP_t = \mu_t dt + \lambda_t dY_t$ for nonrandom μ and λ . Either approach is feasible for the basic model. In some generalizations of the model, the first conjecture works, and in others the second works.^d Taking approach (ii), one can analyze the Hamilton–Jacobi–Bellman equation for the informed trader, using the price as the state variable, to deduce that μ must be zero and λ must be constant for the informed trader to have an optimum. When this is true, any continuous finite-variation strategy for the informed trader is optimal, provided the price is pushed to $\tilde{v}$ by the announcement date.

To deduce the value of λ , one can use the Kalman–Bucy filtering equation (*see Filtering*). Alternatively, in the basic model, one can reason as follows: because the price is always the conditional expectation of $\tilde{v}$, it is a martingale, given market makers' information. Because $dP = \lambda dY$, Y must also be a martingale relative to market makers' information. Given that dX is of order dt , it follows that the volatility of Y is the volatility of Z . Thus, Y is actually a Brownian motion with volatility σ_z , relative to the market makers' information. In particular, Y_1 is normally distributed with mean zero and variance σ_z^2 . Given that $\tilde{v} = P_0 + \lambda Y_1$, it follows that $P_0 = \tilde{v}$ and $\lambda = \sigma_v/\sigma_z$.

The informational advantage of the informed trader in this model can be summarized as follows. The price process appears to be a Brownian motion to market makers and the rest of the market. However, the informed trader knows its date 1 value $\tilde{v}$. Thus, to the informed trader, the price process is a Brownian bridge. The informed trader's equilibrium trading strategy is precisely the drift of a Brownian bridge:

$$dX_t = \frac{\sigma_z}{\sigma_v} \left(\frac{\tilde{v} - P_t}{1 - t} \right) dt \quad (2)$$

For general (i.e., nonnormal) continuous distributions of the asset value, the result of constant depth generalizes to λ being a martingale, given market makers' information. For example, if the asset value is lognormally distributed, then the equilibrium price process is a geometric Brownian motion: $dP = \lambda(P) dY = \phi P dY$ for constant ϕ [2]. This martingale property is robust to time-varying liquidity trading and a flow of information to the informed trader [5] but not to risk aversion [7, 12] or to the presence of multiple informed traders [4, 8, 11].

Acknowledgments

We are grateful to Pete Kyle for helpful comments.

End Notes

^aThe paper of Kyle [13] was the seventh most widely cited finance paper during the 1990s, according to Arnold *et al.* [1].

^bThe assumption that orders are batched is unimportant if liquidity trades are small and arrive frequently. This is established by Back and Baruch [3], who analyze a discrete-trade/Poisson model, which can be interpreted as a Glosten–Milgrom [9] model with explicit optimization by the informed trader. They show that the equilibria of the Glosten–Milgrom model converge to the equilibrium of a continuous-time Kyle model as the arrival rate of liquidity trades converges to infinity and the size of individual liquidity trades converges to zero.

^cThe informed trader can implicitly observe Z by observing the price and backing out the influence of X on the price.

^dA third approach is useful in the case of an informed trader with CARA utility: take $P_t = H(t, U_t)$ where $dU_t = \kappa_t dY_t$ for some function H and a nonrandom κ [7].

References

- [1] Arnold, T., Butler, A.W., Crack, T.F. & Altintig, A. (2003). Impact: what influences finance research? *Journal of Business* **76**, 343–361.
- [2] Back, K. (1992). Insider trading in continuous time, *Review of Financial Studies* **5**, 387–409.
- [3] Back, K. & Baruch, S. (2004). Information in securities markets: Kyle Meets Glosten and Milgrom, *Econometrica* **72**, 433–465.
- [4] Back, K., Cao, C.H. & Willard, G.A. (2000). Imperfect competition among informed traders, *Journal of Finance* **55**, 2117–2155.
- [5] Back, K. & Pedersen, H. (1998). Long-lived information and intraday patterns, *Journal of Financial Markets* **1**, 385–402.
- [6] Bagehot, W. (1971). The only game in town, *Financial Analysts' Journal* **22**, 12–14. (pseud.).
- [7] Baruch, S. (2002). Insider trading and risk aversion, *Journal of Financial Markets* **5**, 451–464.
- [8] Foster, F.D. & Viswanathan, S. (1996). Strategic trading when agents forecast the forecasts of others, *Journal of Finance* **51**, 1437–1478.
- [9] Glosten, L.R. & Milgrom, P.R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* **14**, 71–100.
- [10] Hellwig, M.F. (1980). On the aggregation of information in competitive markets, *Journal of Economic Theory* **22**, 477–498.
- [11] Holden, C.W. & Subrahmanyam, A. (1992). Long-lived private information and imperfect competition, *Journal of Finance* **47**, 247–270.
- [12] Holden, C.W. & Subrahmanyam, A. (1994). Risk aversion, imperfect competition, and long-lived information, *Economics Letters* **44**, 181–190.
- [13] Kyle, A.S. (1985). Continuous auctions and insider trading, *Econometrica* **53**, 1315–1336.
- [14] Treynor, J. (1995). The only game in town, *Financial Analysts' Journal* **51**, 81–83.

Related Articles

Adverse Selection; Filtering; Glosten–Milgrom Models; Liquidity; Liquidity Premium; Market Transparency.

KERRY BACK & SHMUEL BARUCH

Probability of Informed Trading

The extent of private information in markets is a critical determinant in the price formation process. By modeling the trading process and the market maker's learning process, we can use trade data to make inferences about this private information, or more precisely, about the probability of informed trading (PIN). The extent of private information affects bid and ask prices and, hence, the available liquidity in an asset. Perhaps more importantly, the effect of PIN transcends pure market microstructure issues, as evidence suggests that PIN affects the required rate of return on stocks.

The PIN Model

In a series of papers, Easley *et al.* model a market in which a competitive market maker trades a risky asset with informed and uninformed traders [4–8]. As a simple example, consider the sequential trade tree diagram shown in Figure 1. Trade occurs over $t = 1, \dots, T$ discrete trading days and, within each trading day, trade occurs in continuous time. Information events occur between trading days with probability α . When an information event occurs, it is either bad news, with probability δ , or good news with probability, $1 - \delta$. Good news at date t is that the asset is worth $\bar{V}_t$, and bad news is that it is worth V_t .

During any trading day, orders arrive at the market according to Poisson processes. The market maker sets prices to buy or sell throughout the day, and then executes orders as they arrive. Traders informed of bad news sell and those informed of good news buy. Orders from the informed traders follow a Poisson process, with daily arrival rate μ . Uninformed traders trade for liquidity reasons. Buy and sell orders from uninformed traders arrive at the market according to independent Poisson processes, with daily arrival rates ϵ_b for buy orders and ϵ_s for sell orders. If an order arrives at some time, the market maker observes the trade (either a buy or a sale), and he/she uses this information to update his/her beliefs. New prices are set, trades evolve, and the price process

moves in response to the market maker's changing beliefs.

We can use the model parameters to compute the probability that the opening trade on a day is information based, PIN, as

$$\text{PIN} = \frac{\alpha\mu}{\alpha\mu + \epsilon_b + \epsilon_s} \quad (1)$$

where $\alpha\mu + \epsilon_b + \epsilon_s$ is the arrival rate for all orders and $\alpha\mu$ is the arrival rate for information-based orders. Thus, the ratio is the fraction of orders that arise from informed traders or the probability that the opening trade is information based.

The model is quite flexible and allows for extensions with greater complexity in the trading and information processes. Easley *et al.* [3], for instance, present a dynamic extension with time variation in the traders' arrival rates.

Estimating the PIN Model

Suppose that we now view this problem from the perspective of an econometrician. If we, like the market maker, can observe the sequence of trades, we can use this to estimate the structural model *via* maximum likelihood. That is, we can determine the probability of information-based trading in a given stock.

The first step in the estimation is to note that the total number of buys and sells per day is a sufficient statistic for the trade data. This occurs because of the assumption that arrivals follow independent Poisson processes. The likelihood function induced by this model for the total number of buys and sells on a single trading day is

$$\begin{aligned} \mathcal{L}((B, S)|\theta) = & \alpha(1 - \delta)e^{-(\mu + \epsilon_b + \epsilon_s)} \frac{(\mu + \epsilon_b)^B (\epsilon_s)^S}{B!S!} \\ & + \alpha\delta e^{-(\mu + \epsilon_b + \epsilon_s)} \frac{(\mu + \epsilon_s)^S (\epsilon_b)^B}{B!S!} \\ & + (1 - \alpha)e^{-(\epsilon_b + \epsilon_s)} \frac{(\epsilon_b)^B (\epsilon_s)^S}{B!S!} \quad (2) \end{aligned}$$

where (B, S) is the total number of buys and sells for the day and $\theta = (\mu, \epsilon_b, \epsilon_s, \alpha, \delta)$ is the parameter vector. This likelihood function is a mixture of three Poisson probabilities, weighted by the probability of having a “good news day” $\alpha(1 - \delta)$, a “bad news day” $\alpha\delta$, and a “no news day” $(1 - \alpha)$.

The model assumes that each day the arrivals of an information event and trades, conditional on information events, are drawn from identical and independent distributions. Thus, the likelihood function for T days is a product of the above likelihood over days. The log likelihood function, after dropping a constant term and rearranging, can be written as

$$\begin{aligned} \mathcal{L}((B_t, S_t)_{t=1}^T | \theta) &= \sum_{t=1}^T [-\varepsilon_b - \varepsilon_s + M_t (\ln x_b + \ln x_s) \\ &\quad + B_t \ln(\mu + \varepsilon_b) + S_t \ln(\mu + \varepsilon_s)] \\ &\quad + \sum_{t=1}^T \ln \left[\alpha(1 - \delta) e^{-\mu} x_s^{S_t - M_t} x_b^{-M_t} \right. \\ &\quad \left. + \alpha \delta e^{-\mu} x_b^{B_t - M_t} x_s^{-M_t} + (1 - \alpha) x_s^{S_t - M_t} x_b^{B_t - M_t} \right] \end{aligned} \quad (3)$$

where $M_t = \min(B_t, S_t) + \max(B_t, S_t)/2$, $x_s = \frac{\varepsilon_s}{\mu + \varepsilon_s}$, and $x_b = \frac{\varepsilon_b}{\mu + \varepsilon_b}$. The factoring of $x_b^{M_t}$ and $x_s^{M_t}$ increases the computing efficiency and reduces the truncation error. This is particularly important for stocks with a large number of buys and sells as it avoids computation of small fractions (arrival rates) taken to large powers (numbers of trades).

The likelihood function (3) depends upon the number of buys and sells per day. In the United States, the most widely available high-frequency data sets do not contain information about the underlying orders, only the resulting transactions. Therefore, researchers rely on algorithms such as the Lee and Ready [10] algorithm, which mainly uses trade placement relative to the current bid and ask quotes to determine trade direction.

Easley *et al.* [4] obtain yearly PIN estimates for all New York Stock Exchange (NYSE)-listed stocks. They find that PIN is on average 0.19 across stocks and its cross-sectional standard deviation is 0.057. That is, for an average stock, 19% of trades are motivated by private information.

Applications

The estimates of the model's structural parameters can be used to construct the theoretical opening bid and ask prices. As is standard in microstructure models, the market maker sets trading prices such that his/her expected losses to informed traders just offset his/her expected gains from trading with uninformed traders. This balancing of gains and losses gives rise to the bid–ask spread. The opening spread is easiest to interpret if we express it explicitly in terms of PIN. In the case where the uninformed are equally likely to

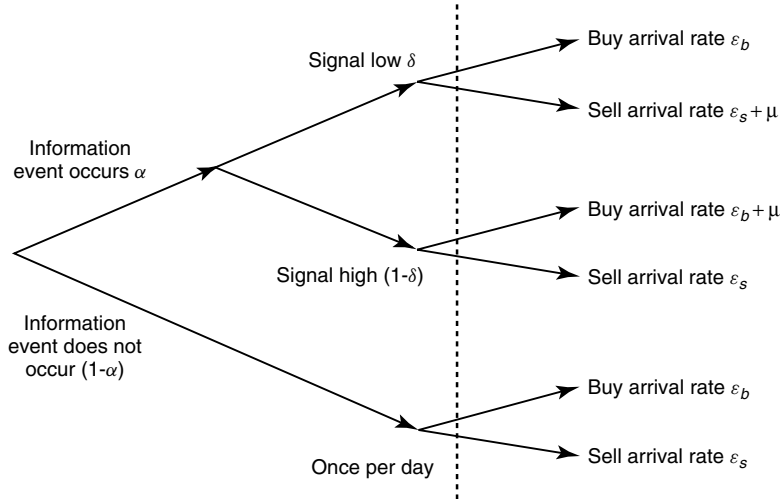


Figure 1 Tree diagram of the trading process. α is the probability of an information event, δ is the probability of a low signal, μ is the rate of informed trade arrival, ε_b is the arrival rate of uninformed buy orders, and ε_s is the arrival rate of uninformed sell orders. Nodes to the left-hand side of the dotted line occur once per day

buy and sell ($\varepsilon_b = \varepsilon_s = \varepsilon$) and news is equally likely to be good or bad ($\delta = 0.5$), the percentage opening spread is

$$\frac{\Sigma}{V_t^*} = \text{PIN} \frac{(\bar{V}_t - \bar{V}_t)}{V_t^*} \quad (4)$$

where Σ is the spread, ask minus bid price, and V_t^* is the unconditional expected value of the asset given by $V_t^* = \delta \bar{V}_t + (1 - \delta) \bar{V}_t$. The opening spread is therefore directly related to the probability of informed trading. Note that if PIN equals zero, because of the absence of either new information ($\alpha = 0$) or traders informed of it ($\mu = 0$), the spread is also zero. This reflects the fact that only asymmetric information affects spreads when market makers are risk neutral. Neither the estimated measure of information-based trading nor the predicted spread is related to market maker inventory because these factors do not enter into the model. Instead, these estimates represent a pure measure of the risk of private information.

Indeed, Easley *et al.* [8] and [4] find a strong relationship between PIN and the opening bid–ask spreads, consistent with the predictions of the model. In addition, PIN is negatively related cross-sectionally to variables such as firm size and the level trading activity and positively related to the level of insider and institutional holdings [1].

Informational asymmetry has implications for a firm's cost of capital. An increase in PIN represents a risk not only to the market maker but also to the uninformed investors who cannot adjust their portfolios to take the private information into account. Therefore, uninformed investors would need to be compensated to invest in high PIN stocks. Easley and O'Hara [9] formalize this notion in a rational expectations model and show that more private information and less public information lead to higher expected returns. Easley *et al.* [4] find that a difference in PIN of 10 percentage points (about 2 standard deviations) between two stocks leads to a difference in expected returns of 2.5% per year.

The PIN variable also represents a straightforward control or interaction variable for researchers interested in the impact of a third variable on, say, stock prices. As an example, Duarte *et al.* [2] find that the effect of regulation fair disclosure on a firm's cost of capital is negative related to PIN, which is consistent with the notion that high PIN firms benefit more from decreases in the informational asymmetry resulting from increased disclosure.

References

- [1] Aslan, H., Easley, D., Hvidkjaer, S. & O'Hara, M. (2007). *Firm Characteristics and Informed Trading: Implications for Asset Pricing*. Working paper, University of Houston, Cornell University and INSEAD.
- [2] Duarte, J., Han, X., Harford, J. & Young, L. (2008). Information asymmetry, information dissemination and the effect of regulation FD on the cost of capital. *Journal of Financial Economics* **87**, 24–44.
- [3] Easley, D., Engle, R., O'Hara, M. & Wu, L. (2008). Time-varying arrival rates of informed and uninformed trades. *Journal of Financial Econometrics* **6**(2), 171–207.
- [4] Easley, D., Hvidkjaer, S. & O'Hara, M. (2002). Is information risk a determinant of asset returns? *Journal of Finance* **57**, 2185–2221.
- [5] Easley, D., Kiefer, N.M. & O'Hara, M. (1996). Cream-skimming or profit-sharing? The curious role of purchased order flow. *Journal of Finance* **51**, 811–833.
- [6] Easley, D., Kiefer, N.M. & O'Hara, M. (1997). The information content of the trading process. *Journal of Empirical Finance* **4**, 159–186.
- [7] Easley, D., Kiefer, N.M. & O'Hara, M. (1997). One day in the life of a very common stock. *Review of Financial Studies* **10**, 805–835.
- [8] Easley, D., Kiefer, N.M., O'Hara, M. & Paperman, J.B. (1996). Liquidity, information, and infrequently traded stocks. *Journal of Finance* **51**, 1405–1436.
- [9] Easley, D. & O'Hara, M. (2004). Information and the cost of capital. *Journal of Finance* **59**(4), 1553–1583.
- [10] Lee, C.M.C. & Ready, M.J. (1991). Inferring trade direction from intraday data. *Journal of Finance* **46**, 733–746.

SOEREN HVIDKJAER

Roll Model

In his 1984 article entitled “A Simple Implicit Measure of the Effective Bid–Ask Spread in an Efficient Market”, Roll [4] developed a model that made it possible to measure an estimate of the bid–ask spread. This estimate depended only on the first-order serial covariance of price changes over time. As one of the first theoretical models to estimate the bid–ask spread, Roll’s model has served as a strong foundation upon which many future expansions and improvements have emerged to estimate the spread and its determinants.

Bid–Ask Bounce

One of the key concepts considered in the development of the Roll model is the idea of a “bid–ask bounce”. When a trade is undertaken by a market maker, it is generally executed at a price different from the actual market value of the asset. Instead a trader’s purchase from the market maker will be completed at the market maker’s ask price, while a sale to a market maker will be completed at the market maker’s bid price. The difference between the ask and the bid prices is referred to as the *spread*.

To estimate the bid–ask spread using the observed transaction prices, Roll assumed that the probability that the next transaction is a buy is equal to the probability of it being a sell. If the market maker’s previous transaction was a purchase at the bid price, then there is an equal probability that the next transaction will be executed at either the bid or the ask. The same logic holds if the market maker’s previous transaction were to be a sale to a customer at the ask price. For example, assume that the market maker’s trade at time $t - 1$ was executed at the ask price, then there is an equal probability that the next trade at time t will be at the bid or the ask. If at time t , the trade is again at the ask, then the change in price is zero. However, if it is at the bid, then the change in price is equal to the spread “ $-s$ ” (Figure 1), where s is the difference between the bid and the ask prices.

Figure 1 shows the possible sequence of trades given that a trade at time $t - 1$ takes place at the ask price. The fundamental value of the asset, P^* , is at the midpoint between the bid and the ask prices and

it follows a random walk. $P_t^* = P_{t-1}^* + \tilde{\eta}_t$, where $\tilde{\eta}_t$ is white noise such that

$$E[\eta_t] = 0, \text{Var}[\tilde{\eta}_t] = \sigma^2$$

$$\text{and } E[\tilde{\eta}_t \tilde{\eta}_{t-s}] = 0 \quad \forall t \neq s \quad (1)$$

Given Roll’s assumption that it is equally likely for the next trade to be at the bid or the ask, the probabilities associated with successive price changes are given in Table 1. The probabilities are based on the observation that there are a total of eight paths from $t - 1$ to $t + 1$, starting from either the ask or the bid prices. Therefore, each path has a probability of $1/8$. Each path is associated with joint price changes of ΔP_t (where $\Delta P_t = P_t - P_{t-1}$) and ΔP_{t+1} that is a combination of 0 , s , and $-s$. For example, there is only one path for $\Delta P_t = 0$ and $\Delta P_{t+1} = -s$ starting from the ask price at $t - 1$. Hence, the probability of this joint event is $1/8$. Similarly, we can find the probabilities for each combination of price changes.

On the basis of Table 1, the serial covariance of price changes $SCov(\Delta P_t, \Delta P_{t+1})$ is $-s^2/4$ and $s = 2\sqrt{-SCov(\Delta P_t, \Delta P_{t+1})}$. With this equation, an estimate of the spread can be obtained based on the estimate for the serial covariance of successive price changes. The standard serial covariance estimator used is the mean cross product: $\widehat{SCov} = \frac{1}{n} \sum_{t=1}^n \Delta P_t \Delta P_{t+1} - \overline{\Delta P}^2$, where $\overline{\Delta P}$ is the sample mean of P_t . Initially, Roll’s model was developed to estimate bid–ask spread from transaction prices. However, there are some drawbacks to its practical use. Some of the model’s assumptions are very specific and may not carry over to real-life application. For example, the Roll estimator is based on the assumption that the stock’s fundamental value follows a random walk and that there is an equal probability that the next trade will be a buy or sell. Hence, the probability of a future trade at the ask or the bid prices is independent of the prior trade direction. However, there is some evidence that the probability of a trade continuation might be different from a trade reversal. The model also assumes that the spread is entirely due to order processing costs and does not include adverse selection or inventory costs, implying that ΔP_t is independent of whether the transaction at time t is at the bid or the ask. Harris [2] provided a detailed analysis of the small sample properties of the Roll serial covariance estimator

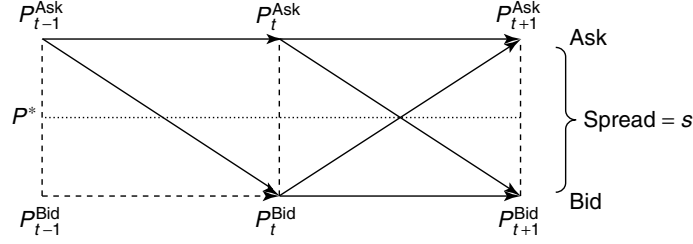


Figure 1 Possible sequence of trades given that the trade at time $t - 1$ takes place at the ask price (based on Roll [4])

Table 1 Joint price change distribution for ΔP_t and ΔP_{t+1} (based on Roll [4])

		ΔP_t		
		$+s$	0	$-s$
ΔP_{t+1}	$+s$	0	1/8	1/8
	0	1/8	1/4	1/8
	$-s$	1/8	1/8	0

for the bid–ask spread. He concluded that the serial covariance estimator is very noisy in daily data and is biased downward in small samples. Using weekly data would reduce the bias, but it will also lead to smaller sample sizes. He also found that the bias from Jensen’s inequality is very serious in daily and weekly data.

Bid–Ask Spread-induced Bias in Volatility Estimator

Stock return volatility estimates based on transaction prices will lead to an upward bias due to the bid–ask bounce. The transaction price, P_t , consists of the unobserved “true” price P_t^* , and the half-spread, $+\frac{s}{2}$ if the transaction was at the ask price and $-\frac{s}{2}$ if the transaction was at the bid price. Therefore, $\tilde{P}_t = P_t^* + \frac{s}{2} \tilde{D}_t$, where $\tilde{D}_t$ is the buy/sell indicator, which is equal to 1 with probability 0.5 (buyer-initiated trade) and equal to -1 with probability 0.5 (seller-initiated trade).

If we assume that P_t^* evolves as a random walk, such that $P_t^* = P_{t-1}^* + \tilde{\eta}_t$, then the variance of the price change based on transaction prices can be written as

$$\begin{aligned} \text{Var}(\Delta P_t) &= \text{Var}(\tilde{\eta}_t) + \text{Var}\left(\frac{s}{2} \Delta \tilde{D}_t\right) \\ &= \sigma^2 + \frac{s^2}{4} \text{Var}(\Delta \tilde{D}_t) = \sigma^2 + \frac{s^2}{2} \quad (2) \end{aligned}$$

where $\Delta P_t = P_t - P_{t-1}$ and $\Delta \tilde{D}_t = \tilde{D}_t - \tilde{D}_{t-1}$.

Since the variance calculated using transaction prices will include both the variance from the bid–ask bounce as well as the actual price variance, the calculated variance will overstate the fundamental volatility by $\frac{s^2}{2}$.

The magnitude of the bid–ask bounce-induced volatility remains the same whether we look at short-interval (high-frequency) or long-interval (low-frequency) returns. However, volatility due to the bid–ask bounce becomes a larger proportion of the fundamental return volatility with a decrease in the return horizon. French and Roll [1] provided a variance estimator that is unbiased but noisy.

A Generalized Roll Model with Adverse Selection

Since the publication of Roll’s paper in 1984, there have been many related papers that relax some of the assumptions made by Roll, and others have used Roll’s model as a foundation to develop general models to include asymmetric information, inventory costs, price discreteness, and other microstructure related effects. We can generalize the Roll model to include the adverse selection cost as follows.

The unobserved fundamental value P_t^* is given as

$$P_t^* = P_{t-1}^* + \tilde{\eta}_t + (1 - \pi) \left(\frac{s}{2} \tilde{D}_t \right) \quad (3)$$

where $(1 - \pi)$ is the fraction of the half-spread due to adverse selection and $\tilde{\eta}_t$ is a serially uncorrelated public information shock. If the fraction of the half-spread due to order processing is π , then the

transaction price is given as

$$P_t = P_t^* + \pi \left(\frac{s}{2} \tilde{D}_t \right) \quad (4)$$

If π is set to equal 1, we get back to the original Roll model. Under this model, it can be shown that $SCov(\Delta P_t, \Delta P_{t-1}) = -\pi \frac{s^2}{4}$. Hence, only when $\pi = 1$ (i.e., no adverse selection) can we obtain an unbiased estimator of “ s ” using the $SCov(\Delta P_t, \Delta P_{t-1})$ as in [4]. A more detailed description of a generalized Roll model can be found in [3].

Summary

Roll’s 1984 article developed a relatively simple model to estimate the bid–ask spread based on observable trade prices. This model has been very useful in situations where direct bid–ask data were not available. The Roll model has been a foundation upon which several extensions and other models have emerged to estimate the spread and its determinants.

References

- [1] French, K.R. & Roll, R. (1986). Stock return variances: the arrival of information and reaction of traders, *Journal of Financial Economics* **17**, 5–26.
- [2] Harris, L. (1990). Statistical properties of the roll serial covariance bid/ask spread estimator, *Journal of Finance* **XLV**(2), 579–590.
- [3] Hasbrouck, J. (2007). *Empirical Market Microstructure*, Oxford University Press.
- [4] Roll, R. (1984). A simple implicit measure of the effective bid–ask spread in an efficient market, *Journal of Finance* **39**, 1127–1139.

Related Articles

Adverse Selection; Bid–Ask Spreads; Glosten–Milgrom Models; High-frequency Data; Market Microstructure Effects.

MAHENDRARAJAH NIMALENDHAN

Call Auction Markets

Countless market structures exist for equity trading. Each is a variant of and/or a hybrid combination of four generics: (i) the continuous order-driven, agency market, (ii) the periodic order-driven call auction, (iii) the continuous quote-driven, dealer market, and (iv) the negotiated market (e.g., floor trading, upstairs block negotiation, and electronic negotiation). One of the common market mechanisms is the *call auction market*.

The distinguishing characteristic of a call auction is the batching of orders for simultaneous execution in a multilateral trade at a single clearing price, at a single point in time. This contrasts with continuous trading (either order or quote driven) where a trade can occur whenever a buy and a sell order match or cross in price.

At the specific point in time when a market is “called”, buy orders are sorted and the shares aggregated from the highest priced buy to the lowest, and sell orders are sorted and the shares aggregated from the lowest priced sell to the highest. The two arrays of aggregated shares are matched, and the clearing price is set at the value at which the largest number of shares will execute—where the aggregated number of shares to buy (which decreases with price) equals the aggregated number of shares to sell (which increases with price). Buy orders at the clearing price and higher execute, as do sell orders at this price and lower (further rules of order execution are used when the number of shares to buy does not match the number of shares to sell exactly). Because all orders execute at the common clearing price (not the price at which they have been placed), executable orders are generally price improved.

If the clearing price at the auction is determined as just described, the system is a *price discovery call*. If the clearing price is set elsewhere (exogenously), the system is a *crossing network*. The clearing price for an intraday crossing network is typically the midpoint of the bid and ask quotes posted by a major market center. The after-hours crosses typically use market closing prices. This article focuses on *price discovery* call auctions. For further discussion of the properties of call auction trading, see [4, 5] and [14], Chapter 4.

The Advent of Electronic Calls

Each of the four above-noted generic forms of trading can involve human-to-human interaction (either face-to-face on an exchange floor or via the telephone) and/or electronic technology. Call auction trading, however, predates electronic technology. One hundred and thirty eight years ago, the New York Stock Exchange was a call market. The Tel-Aviv Stock Exchange through the 1970s and the Paris Bourse, before it introduced its electronic, order-driven market in 1986, were also nonelectronic call auctions. Pressure from an increasing order flow and from competing continuous markets crowded out the non-electronic calls.

In the past two decades, call auction trading has made its re-entrance as an electronic marketplace. The electronic calls are being used, most notably, by Deutsche Börse; NYSE Euronext’s Paris, Amsterdam, and Brussels exchanges; the London Stock Exchange; SWX Swiss Exchange; and the Nasdaq Stock Market. The New York Stock Exchange opens and closes its markets with calls but, as of this writing, its calls are not fully electronic.

The electronic calls are not standalone systems but have been combined with continuous trading to create hybrid markets. With a hybrid structure, an investor can select among alternative trading venues depending on the size of his or her order, the liquidity of the stock being traded, and the investor’s own motive for trading.

Tempering Volatility at Market Openings and Closings

Because they consolidate orders and amass liquidity, call auctions have been used to sharpen price discovery. As such, they are generally run to open and to close a market, and to reopen a market that has been halted because of stressful market conditions. The opening is itself a time of stress because of the difficulty of translating overnight news into new share values. The closing is a time of stress because of trader impatience to “get the job done” before the trading session ends. The stress that characterizes openings and closings is apparent in the sharp accentuation of intraday volatility in the first and last minutes of a trading day [11].

Paris introduced electronic closing calls in 1996 and 1998 in response to derivative traders’ demand

for better closing prices. In 2004, Nasdaq introduced its electronic calls (which it refers to as its *Crosses*) for a similar reason. As reported in [11], “Nasdaq decided to introduce Closing Cross in Fall 2003 after a competing market, the American Stock Exchange, responding to a strongly expressed request from Standard & Poor’s for better closing prices, started planning a closing call of its own that would be used for Nasdaq stocks.”

The accentuation of volatility at market openings and closings has been well documented in the academic literature (relevant volatility studies include [1, 7–11, 16]). The ability of call auctions to contain this volatility and improve market quality has also been established [2, 11–13, 15]. Nevertheless, Ellul *et al.* [6] suggest that the call “suffers from a high failure rate to open and close trading especially when trading conditions are difficult.” The theoretical analysis of order submission to a call auction by Chakraborty *et al.* [3] also shows the difficulty of getting early order submission (a “bookbuilding” problem); these authors suggest that the effectiveness of a call depends on structural issues such as its transparency and whether or not it includes a market maker to “animate” bookbuilding.

Smaller Cap Stocks and Large Institutional Orders

The ecology of an order-driven market breaks down when it receives inadequate order flow. Call auction trading, because it amasses liquidity at fixed points in time, can help overcome this problem. Accordingly, call auctions are important for trading mid- and small-cap stocks (the Paris Bourse uses them for these issues).

Placing large, institutional-sized orders on a continuous, open book market is also problematic (disclosure of a large order before it executes fully can cause adverse price moves that raise execution costs). Order batching, by amassing liquidity and delivering price improvement, also mitigates this adverse effect.

Order Handling in Call versus Continuous Markets

Orders are handled differently in call auctions than in continuous trading:

- In continuous order-driven markets, limit orders set the prices at which market orders execute, and limit orders sitting on the book provide immediacy to market orders. In contrast, market orders in a call auction are nothing more than extremely aggressively priced limit orders. Specifically, a market order to buy has an effective price limit of infinity, and a market order to sell has an effective price limit of zero.
- Limit orders in continuous trading generally execute at the prices at which they have been placed. Because in call auction trading all orders execute at the common clearing price, the more aggressively priced limit orders are price improved. This encourages participants to price their orders more aggressively, which enhances market liquidity.
- All participants in a call auction wait until the next call for their orders to execute. Thus, market orders do not receive immediacy as they do in continuous trading.
- In continuous trading, limit orders supply liquidity and market orders dry up liquidity. This distinction does not apply in call auction trading. In a call auction, all participants supply liquidity to each other. However, if limit orders on the book are publicly displayed, those participants who place their orders early in the precall, order entry period are key contributors to liquidity creation (i.e., they animate bookbuilding).

Overview

Comprehensively viewed, imbedding call auctions in a continuous trading environment can have major positive benefits: most notably, lower volatility, lower execution costs, and sharper price discovery. The success of a call auction is not foreordained, however. To operate effectively, critical mass order flow must be received. Whether the requisite order flow is achieved depends critically on a call auction’s specific architectural design.

Acknowledgments

Parts of this article have been adapted, with kind permission of Springer Science and Business Media, from

The Encyclopedia of Finance, 2006, Alice C. Lee and C.F. Lee, Editors, Springer Science + Business Media, LLC, Chapter 35, pages 623–629, “Call Auction Trading,” Robert A. Schwartz and Reto Francioni.

References

- [1] Amihud, Y. & Mendelson, H. (1987). Trading mechanisms and stock returns: an empirical investigation, *Journal of Finance* **42**, 533–553.
- [2] Barclay, M.J., Hendershott, T. & Jones, C.M. (2005). *Order Consolidation, Price Efficiency, and Extreme Liquidity Shocks*, working paper, University of Rochester.
- [3] Chakraborty, A., Pagano, M.S. & Schwartz, R.A. (2009). *Order Revelation at Market Openings*, working paper, Baruch College.
- [4] Cohen, K.J. & Schwartz, R.A. (1989). An electronic call market: its design and desirability, in *The Challenge of Information Technology for the Securities Markets: Liquidity, Volatility, and Global Trading*, H.C. Luca & R.A. Schwartz, eds, Dow Jones-Irwin, Homewood, IL, pp. 15–58.
- [5] Economides, N. & Schwartz, R.A. (1995). Electronic call market trading, *Journal of Portfolio Management* **21**, 10–18.
- [6] Ellul, A., Shin, H.S. & Tonks, I. (2005). Opening and closing the market: evidence from the London stock exchange, *Journal of Financial and Quantitative Analysis* **40**, 779–801.
- [7] Fleming, M.J. & Remolina, E.M. (1999). Price formation and liquidity in the US treasury market: the response to public information, *Journal of Finance* **5**, 1901–1915.
- [8] Hasbrouck, J. (1993). Assessing the quality of a security market: a new approach to transaction-cost measurement, *The Review of Financial Studies* **6**, 191–212.
- [9] Hasbrouck, J. & Schwartz, R.A. (1988). Liquidity and execution costs in equity markets, *Journal of Portfolio Management* **14**, 10–16.
- [10] Ozenbas, D., Schwartz, R.A. & Wood, R.A. (2002). Volatility in US and European equity markets: an assessment of market quality, *International Finance* **5**, 437–461.
- [11] Pagano, M., Peng, L. & Schwartz, R.A. (2009). *Market Structure and Intra-day Price Volatility: An Event Study on Nasdaq's Crosses*, working paper, Villanova University.
- [12] Pagano, M.S. & Schwartz, R.A. (2003). A closing call's impact on market quality at Euronext Paris, *Journal of Financial Economics* **68**, 439–484.
- [13] Pagano, M.S. & Schwartz, R.A. (2005). Nasdaq's closing cross: has its new call auction given Nasdaq better closing prices? Early findings, *Journal of Portfolio Management* **31**, 100–111.
- [14] Schwartz, R.A., Francioni, R. & Weber, B.W. (2006). *The Equity Trader Course*, John Wiley & Sons.
- [15] Smith, J. (2006). Nasdaq's electronic closing cross: an empirical analysis, Nasdaq economic research, *Journal of Trading* **1**(3), 47–64.
- [16] Werner, I. & Kleidon, A. (1996). UK and US trading of British cross-listed stocks: an intra-day analysis of market integration, *Review of Financial Studies* **9**, 619–664.

ROBERT A. SCHWARTZ & PAUL L. DAVIS

Limit Order Markets

The organization of trading in financial markets has undergone tremendous changes at the end of the twentieth century. In stock markets, electronic limit order books have progressively replaced floor and dealership markets. For instance, the Paris Bourse closed its floor and became a fully electronic limit order market in 1986. Bond and currency markets follow a similar evolution with the emergence of trading platforms such as EBS (FX) or E.speed and MTS (bond markets). This evolution has been a major impetus for academic research on limit order markets. This article provides a brief review of this research.

Market and Limit Orders

Limit orders and market orders (*see* **Order Types**) are the core engine of limit order markets. Limit orders specify a price and a number of shares. For instance, a limit order to buy (respectively sell) 500 shares of stock XYZ at \$10 means that the order submitter is willing to buy up to 500 shares at \$10 or less (respectively more). As they have a prespecified price, limit orders cannot always be matched with contra-side orders upon arrival. In this case, they are stored in a limit order book. They exit the limit order book when they are hit by subsequent orders or cancelled by the order submitter.

A market order specifies a number of shares to buy or to sell with no price limit. It is matched upon arrival with contra-side limit orders (e.g., a buy market order is matched with sell limit orders). The latter execute in sequence according to price priority and other rules (e.g., first-in first-out) for orders tied at the same price. Consider, for instance, the arrival of a buy market order for 1000 shares when the limit order book features sell limit orders at (i) \$10 (best ask price) for 500 shares, (ii) \$10.5 for 700 shares, and (iii) \$10.7 for 200 shares. The market order executes for 500 shares at \$10 and for the remaining 500 shares at the next best offer, \$10.5. Execution of limit orders is “discriminatory” in the sense that each limit order is filled at its own price.

Copeland and Galai [12] point out that limit orders are similar to options. For instance, a buy (sell) limit order at \$10 is similar to a put (respectively

call) option with a \$10 strike price. Indeed, it gives the possibility to other market participants to sell the security at \$10 when its value falls below \$10. These options are free. Yet, they can be written at a profit because, in contrast to regular options, they are sometimes exercised out-of-the money by traders who need immediate execution. Bid and ask prices can then be viewed as the strike prices that balance losses when limit orders execute in-the-money (are “picked off”) with gains when they execute out-of-the money [12]. Foucault *et al.* [18] argue, and show empirically, that the limit order book contains information on future price volatility since limit orders are options.

Limit order traders can reduce the likelihood of being picked off by monitoring their orders. Foucault *et al.* [19] analyze this monitoring decision in a model similar to that in [12]. Liu [32] extends this model to study the interactions between monitoring decisions and cancellations in a limit order market. He provides evidence that traders submitting limit orders actively monitor their orders to reduce the risk of being picked off.

The choice between limit and market orders hinges upon a trade-off between the gain of a limit order (a price improvement relative to standing quotes) and its costs, namely, the cost of nonexecution, the cost of being picked off, the cost of trading with better informed investors, monitoring costs, waiting costs [14], and submission fees. Harris and Hasbrouck [24] study the performance of actual limit orders. They find that buy (respectively sell) limit orders execute more frequently in the case of a price decline (respectively increase). Moreover, they also find that the opportunity cost of unfilled limit orders behind the best quotes is significant. Handa and Schwartz [23] use hypothetical limit and market orders. They find that the performance of filled (respectively unfilled) limit orders is higher (respectively smaller) than that of market orders. Unconditionally, the performance of each order type is identical. Overall, these findings suggest that the cost of being picked off and the cost of nonexecution are significant.

The optimal order submission strategy depends on each trader’s objective (e.g., whether the trader is constrained to trade a specific number of shares or not), his/her expectation regarding future order submissions, and the current state of the book. In turn, traders’ orders affect the state of the book and future order submission choices. Biais *et al.* [6] offer the

first empirical description of the interactions between the state of the limit order book and the order flow in a pure limit order market (the Paris Bourse). They uncover a *diagonal effect*: the frequency of an order of a given type (e.g., a buy market order) conditional on an order of the same type is higher than the unconditional frequency of this order. This observation is confirmed by subsequent empirical analyses of other markets [22, 35]. Biais *et al.* [6] also show that order submission strategies depend on the state of the limit order book in systematic ways. These observations paved the road for research about the effect of the state of the limit order book (e.g., the size of the bid–ask spread, the depth at the best quotes, etc.) on order submission choices (see, for instance, Beber, A. & Caglio, C. (2004). *Order Submission Strategies and Information: Empirical Evidence from the NYSE*. Unpublished manuscript, University of Lausanne; Ellul, A., Holden, C.W., Pankaj, J. & Jennings, R.H. (2003). *Determinants of Order Choice on the New York Stock Exchange*. Unpublished manuscript, Indiana University [22, 35]).

Ahn *et al.* [1] study the interactions between order submissions and transitory volatility. They find that traders use limit orders more frequently after a rise in transitory volatility. One possibility is that this increase in volatility reflects an increase in the bid–ask spread due to transient liquidity pressures. This increase in the bid–ask spread attracts limit order traders as providing liquidity becomes more profitable. Consistent with this interpretation, Biais *et al.* [6] find reversals in the size of bid–ask spreads after the arrival of large market orders. The speed of this reversal determines the resiliency of a limit order market to liquidity shocks and constitutes one important dimension of market liquidity. Degryse *et al.* [13] and Large [30] provide empirical analyses of the resiliency of limit order markets.

Wyart *et al.* [41] uncover a high positive correlation between the bid–ask spread and the volatility per trade in limit order markets. They show that this relationship is expected if both liquidity providers and liquidity takers follow order submission strategies with positive expected profits.

Models of Price Formation in Limit Order Markets and Empirical Tests

The limit order book can be described by the mapping between quotes in the book and the cumulative

quantity available at these quotes (e.g., in the previous example, the cumulative quantity offered for sale is 500 shares at \$10, 1200 shares at \$10.5, and 1400 shares at \$10.7). Naturally, a strand of research on limit order markets focuses on this mapping as it determines the cost of execution for traders submitting market orders and thereby market liquidity. In the previous example, execution costs are smaller other things being equal if the limit order book features, say, a cumulative quantity of 1200 shares at \$10.5 instead of only, say, 600 shares.

Glosten [20] considers a model in which some traders submitting market orders can be better informed on the payoff of a risky security than traders submitting limit orders. He shows that the quotes associated with a given cumulative quantity are equal to limit order traders' expectation regarding the payoff of the security conditional on the size of the next market order being larger than this cumulative quantity. For instance, if V is the payoff of the risky security, $Q(A)$ is the cumulative quantity offered at the ask price A , and X is the size of the incoming market order, then, in equilibrium, $Q(A)$ satisfies

$$A = E(V | X > Q(A)) \quad (1)$$

Thus, ask prices are upper tail conditional expectations. By symmetry, bid prices are lower tail expectations. This feature is closely related to the discriminatory execution of limit orders and does not obtain in earlier models of trading with asymmetric information [29]. It implies that buy market orders execute at a mark-up compared to sell market orders, even if the size of these orders is infinitesimal. That is, there is “*small trade bid ask spread*.”

Glosten [20] assumes perfect competition among limit order traders. Thus, the restrictions on quotes and cumulative quantities (e.g., equation (1)) derive from a zero profit condition for limit orders. Biais *et al.* [7] provide a more general analysis allowing for imperfect competition (see also [5]). Seppi [38], Viswanathan and Wang [40], and Parlour and Seppi [34] study the mapping between quotes and cumulative quantities in limit order markets in models without asymmetric information.

These models are static. Hence, they are silent on the determinants of the choice between market and limit orders. Cohen *et al.* [11] offer the first theoretical analysis of this choice. They uncover the “*gravitational pull effect*.” That is, when the bid–ask spread becomes tight, the benefit of a

price improvement with a limit order becomes small compared to the cost of nonexecution. Thus, traders are “pulled” to use market orders. This effect provides another explanation for the existence of a small trade bid–ask spread in limit order markets.

Execution probabilities are exogenous in [11]. Foucault [16] (Foucault, T. (1995). *Price Formation and Order Placement Strategies in a Dynamic Order Driven Market*. Working paper, Universitat Pompeu Fabra.), Parlour [33], Foucault *et al.* [17], Rosu [36], and Large [31] introduce dynamic models of trading in limit order markets in which execution probabilities are consistent with order submission choices in equilibrium. These models explain variation in order submission choices by variations in the state of the book and traders’ valuations for the security. Traders with high valuations submit more aggressive buy orders and traders with low valuations submit more aggressive sell orders. Thus, traders with extreme valuations are more likely to submit market orders and traders with intermediate valuations are more likely to submit limit orders.

Parlour [33] studies a one tick market. Upon arrival, buyers (respectively sellers) choose to submit a limit order at the bid (respectively ask) price or a market order. In this model, limit orders face a risk of nonexecution. In equilibrium, traders’ order placement decisions depend on the state of the limit order book on *both* sides. For instance, the decision to submit a buy limit order depends on the depth on the ask side because this depth will affect future sellers’ order submission choices. Moreover, the equilibrium features systematic patterns on order submissions. For instance, a sell market order is more likely after a sell market order than after a buy market order as found empirically in [6].

Foucault [16] focuses on the effect of price volatility on order submission choices. In this model, traders’ valuations for the security are distributed around a consensus value that follows a random walk. Thus, changes in traders’ valuations are positively correlated with changes in the consensus value. Consequently, buy (respectively sell) limit orders are more likely to execute when the consensus value decreases (respectively increases) than when it increases (respectively decreases). For this reason, limit orders are exposed to the risk of being picked off. Under simplifying assumptions (e.g., limit orders expire after one period), Foucault [16] characterizes order submission strategies

in closed-form. In equilibrium, the fraction of limit orders in the order flow and the bid–ask spread increase with the volatility of the consensus value. In contrast, limit orders fill rate decreases with volatility since market orders are less frequent when the bid–ask spread is large. Van Achter extends this model by considering the case in which limit orders have longer time to expiration (Van Achter, M. (2006). *A Dynamic Limit Order Market with Diversity in Trading Horizons*. Working paper, available at <http://ssrn.com/abstract=967610>).

Foucault *et al.* [17], Rosu [36], and Large [31] study the role of waiting costs. In these models, traders care about time to execution for limit orders since it determines their expected waiting cost. Foucault *et al.* [17] study the determinants of expected time to execution and the speed at which the bid–ask spread reverts to competitive levels after liquidity shocks. Rosu [36] introduces the possibility of cancellations and provides an explanation for the high cancellation rate observed in some markets (See for instance Hasbrouck, J. and Saar, G. (2002). *Limit orders and volatility in a hybrid market: the Island ECN*, Working paper, NYU). Large [31] proposes a new approach that exploits the fact that there must be as many filled limit orders as submitted market orders.

All these models obtain analytical results at the cost of strong assumptions. Goettler *et al.* [21] relax many of these assumptions while retaining the idea that changes in traders’ valuations and the consensus value drive variations in order submission choices. They use techniques from computational economics to numerically calculate the equilibrium of the limit order market for specific parameter values. They find that the diagonal effect is an equilibrium property of the limit order market. It arises because (i) there is persistence in the state of the limit order book (hence, traders submit similar orders) and (ii) it takes time for traders to take advantage of stale limit orders after large changes in the consensus value (as initially suggested by Hollifield *et al.* [26]).

Do the trade-offs considered in theories of limit order markets explain actual order submissions? Sandås [37] tackles this question by testing restrictions implied by Glosten’s [20] model for quotes, cumulative quantities, and changes in transaction prices using data from the Stockholm Stock Exchange. He rejects the model because actual limit order books are not deep enough to be explained by the

model. One possible explanation is that competition among limit order traders is imperfect (a feature absent in [20]). Another, related, possibility is that it takes time for the limit order book to replenish after liquidity shocks. Thus, trades occur when profit opportunities remain in the limit order book.

Hollifield *et al.* [26] test a more general formulation of [16]. Their model makes predictions about which type of orders (e.g., a buy limit order one tick below the best ask price) should *not* be observed for each state of the limit order book because they are suboptimal for all traders (whatever be their valuation). They build a test of the model by tracking these suboptimal orders. They reject the model because the expected payoffs on limit orders with low execution probabilities are too small compared to the expected payoffs of limit orders with higher execution probabilities. Thus, these limit orders with low execution probabilities should not be observed according to the model.

New Research Directions

Despite their widespread use in financial markets, little is known on the allocative efficiency of limit order markets. Empiricists often use measures of bid–ask spreads as a proxy for efficiency, presuming that larger bid–ask spreads are associated with greater inefficiencies. However, Goettler *et al.* [21] show that bid–ask spreads are poor proxies of actual gains from trade in limit order markets. Hollifield *et al.* [27] develop a methodology to empirically estimate the fraction of the maximum gains from trade that are achieved in the limit order market. They implement this methodology with data from the Vancouver Stock Exchange and find that about 90% of the maximum gains from trade are achieved.

A related question is the performance of limit order markets relative to other market structures. Viswanathan and Wang [40] and Back and Baruch [3] compare equilibrium outcomes in a market with discriminatory execution of limit orders and a market with uniform execution of limit orders (i.e., all limit orders eligible for execution execute at the execution price of the marginal share). In a model related to Glosten [20], Back and Baruch [3] identify conditions under which the equilibrium of the two markets is identical. Some authors (e.g., Baruch [4] and Boehmer *et al.* [9]) also study theoretically or

empirically the effects of trading rules in limit order markets (transparency, priority rules, tick size, etc.) on their performance (e.g., informational efficiency and liquidity).

Models of limit order trading often assume that informed traders submit market orders. Kaniel and Liu [28] relax this assumption and establish conditions under which informed traders use limit orders. Bloomfield *et al.* [8] find experimental evidence that informed traders use both market and limit orders. These findings imply that limit order books may contain information about future price movements. Anand *et al.* [2], Harris and Panchapagesan [25], (Irvine, P., Benston, G. & Kandel, E. (2000). *Liquidity Beyond the Inside Spread: Measuring and Using Information in the Limit Order Book*. Working paper, Emory University.), and Kaniel and Liu [28] provide empirical evidence consistent with this implication.

Theories of limit order placement assume that traders have rational expectations and choose their orders optimally in any circumstances. Several papers [10, 15, 39] (Ladley, D. & Reiner Schenk-Hoppé, K. (2007). *Do Stylised Facts of Order Book need Strategic Behaviour*. Working paper no. 383. National Centre of Competence in Research Financial Valuation and Risk Management.) study models of limit order trading with zero-intelligence traders. These traders submit random orders satisfying some minimal conditions (e.g., they do not violate traders' budget constraint). Interestingly, some empirical regularities of limit order markets (e.g., the diagonal effect) emerge even with zero-intelligence traders. Thus, the trading mechanics of these markets can in itself explain some of the properties of limit order markets, independently of the degree of sophistication for the traders.

References

- [1] Ahn, H.J., Bae, K.H. & Chan, K. (2001). Limit orders, depth and volatility: evidence from the stock exchange of Hong Kong, *Journal of Finance* **56**, 767–788.
- [2] Anand, A., Chakravarty, Q. & Martell, T. (2005). Empirical evidence on the evolution of liquidity: choice of market versus limit orders by informed and uninformed traders, *Journal of Financial Markets* **8**, 288–308.
- [3] Back, K. & Baruch, S. (2007). Working orders in limit order markets and floor exchanges, *Journal of Finance* **62**, 1589–1621.
- [4] Baruch, S. (2005). Who Benefits from an Open Limit-Order Book? *Journal of Business* **78**, 1267–1306.

- [5] Bernhardt, D. & Hughson, E. (1997). Splitting orders, *Review of Financial Studies* **69**, 101.
- [6] Biais, B., Hillion, P. & Spatt, C. (1995). An empirical analysis of the limit order book and the order flow in the Paris bourse, *Journal of Finance* **50**, 1655–1689.
- [7] Biais, B., Martimort, D. & Rochet, J.C. (2000). Competing mechanisms in a common value environment, *Econometrica* **68**, 799–798.
- [8] Bloomfield, R., O'Hara, M. & Saar, G. (2005). The “make or take” decision in an electronic market: evidence on the evolution of liquidity, *Journal of Financial Economics* **75**, 165–199.
- [9] Boehmer, E., Saar, G. & Yu, L. (2005). Lifting the veil: an analysis of pre-trade transparency at the NYSE, *Journal of Finance* **60**, 783–815.
- [10] Bouchaud, J.P., Mézard, M. & Potters, M. (2002). Statistical properties of stock order books: empirical results and models, *Quantitative Finance* **2**, 251–256.
- [11] Cohen, K.J., Maier, S.F., Schwartz, R.A. & Whitcomb, D. (1981). Transaction costs, order placement strategy, and existence of the bid-ask spread, *Journal of Political Economy* **89**, 287–305.
- [12] Copeland, T. & Galai, D. (1983). Information effects on the bid-ask spread, *Journal of Finance* **38**, 1457–1469.
- [13] Degryse, H., de Jong, F., van Ravenswaaij, V. & Wuyts, G. (2005). Aggressive orders and resiliency of the limit order book, *Review of Finance* **9**, 201–242.
- [14] Demsetz, H. (1968). The costs of transacting, *Quarterly Journal of Economics* **82**, 33–53.
- [15] Farmer, J.D., Patelli, P. & Zovko, I.I. (2005). The predictive power of zero intelligence in financial markets, *Proceedings of the National Academy of Science* **102**, 2254–2259.
- [16] Foucault, T. (1999). Order flow composition and trading costs in a dynamic limit order market, *Journal of Financial Markets* **2**, 99–134.
- [17] Foucault, T., Kadan, O. & Kandel, E. (2005). Limit order book as a market for liquidity, *Review of Financial Studies* **18**, 1171–1217.
- [18] Foucault, T., Moinas, S. & Theissen, E. (2007). Does anonymity matter in electronic limit order markets? *Review of Financial Studies* **20**, 1707–1747.
- [19] Foucault, T., Roëll, A. & Sandås, P. (2003). Market making with costly monitoring: an analysis of SOES trading, *Review of Financial Studies* **16**, 345–384.
- [20] Glosten, L.R. (1994). Is the electronic open limit order book inevitable? *Journal of Finance* **49**, 1127–1161.
- [21] Goettler, R.L., Parlour, C.A. & Rajan, U. (2005). Equilibrium in a dynamic limit order market, *Journal of Finance* **60**, 2149–2192.
- [22] Griffiths, M., Smith, B., Turnbull, A. & White, R. (2000). The costs and determinants of order aggressiveness, *Journal of Financial Economics* **56**, 65–88.
- [23] Handa, P. & Schwartz, R.A. (1996). Limit order trading, *Journal of Finance* **51**, 1835–1861.
- [24] Harris, L. & Hasbrouck, J. (1996). Market vs. limit orders: the superdot evidence on order submission strategy, *Journal of Financial and Quantitative Analysis* **31**, 213–231.
- [25] Harris, L. & Panchapagesan, V. (2005). The Information content of the limit order book: evidence from NYSE specialist trading decisions, *Journal of Financial Markets* **7**, 25–67.
- [26] Hollifield, B., Miller, R.A. & Sandas, P. (2004). Empirical analysis of limit order markets, *Review of Economic Studies* **71**, 1027–1063.
- [27] Hollifield, B., Miller, R.A., Sandas, P. & Slive, J. (2006). Estimating the gains from trade in limit order markets, *Journal of Finance* **61**, 2753–2804.
- [28] Kaniel, R. & Liu, H. (2006). So what orders do informed traders use? *Journal of Business* **79**, 1867–1913.
- [29] Kyle, R. (1985). Continuous auctions and insider trading, *Econometrica* **53**, 1315–1335.
- [30] Large, J. (2007). Measuring the resiliency of an electronic limit order book, *Journal of Financial Markets* **10**, 1–25.
- [31] Large, J. (2009). A market clearing role for inefficiency on a limit order book, *Journal of Financial Economics* **91**, 102–117.
- [32] Liu, W. (2007). Monitoring and limit order submission risks, *Journal of Financial Markets* **12**, 107–141.
- [33] Parlour, C.A. (1998). Price dynamics in limit order markets, *Review of Financial Studies* **11**, 789–816.
- [34] Parlour, C.A. & Seppi, D.J. (2003). Liquidity-based competition for order flow, *Review of Financial Studies* **16**, 301–343.
- [35] Rinaldo, A. (2004). Order aggressiveness in limit order book markets, *Journal of Financial Markets* **7**, 53–74.
- [36] Rosu, I. (2009). A dynamic model of the limit order book, *Review of Financial Studies* April, 1–41.
- [37] Sandås, P. (2001). Adverse selection and competitive market making: empirical evidence from a limit order market, *Review of Financial Studies* **14**, 705–734.
- [38] Seppi, D.J. (1997). Liquidity provision with limit orders and a strategic specialist, *Review of Financial Studies* **10**, 103–150.
- [39] Smith, E., Farmer, D., Gillenot, L. & Krishnamurthy, S. (2003). Statistical theory of the continuous double auction, *Quantitative Finance* **3**, 481–514.
- [40] Viswanathan, S. & Wang, J. (2002). Market architecture: limit order books versus dealership markets, *Journal of Financial Markets* **5**, 127–168.
- [41] Wyart, M., Bouchaud, J.P., Kockelkoren, J., Potters, M. & Vettorazzo, M. (2008). Relation between bid-ask spread, impact and volatility in order-driven markets, *Quantitative Finance* **8**, 41–57.

Related Articles

Bid-Ask Spreads; Glosten-Milgrom Models; High-frequency Data; Market Microstructure Effects; Order Types.

THIERRY FOUCAULT

Specialist Markets

A specialist market is defined as a hybrid market structure that includes an auction component (e.g., a floor auction or a limit order book) together with one or more designated market makers (“specialists”) who trade as dealers for their own account. The designated market makers have some responsibility for the market; for example, specialists are usually obligated to continuously post quotes with reasonable depth. In return, they get a preferential treatment, be it access to order flow, information about the state of the market, discounted trading fees, or direct monetary compensation from the exchange or the listed firm.

The New York Stock Exchange (NYSE) is considered the quintessential specialist market, and one can trace several important elements in this definition to the historical development of that market. The NYSE started as an auction market, and, prior to World War II, the most important function of the specialist was as a brokers’ broker, coordinating the trading process and maintaining a limit order book [20]. In other words, the auction component of the hybrid specialist system (which enables investors to trade directly with other investors) has always been a feature of this market structure, and it makes specialist markets distinct from pure dealer markets.

The dealer function of the specialists on the NYSE received greater emphasis as brokers sought speedy execution of orders and the importance of the affirmative obligation of specialists to assist in maintaining a fair and orderly market was stressed [20, 21]. It is the specialist’s presence that makes this market structure different from pure auction markets (e.g., call auctions, floor markets, or limit order books; see **Call Auction Markets** and **Limit Order Markets**). While some specialist markets mandate only one specialist, others allow (and even encourage) multiple designated market makers. On the NYSE, for example, there were competing specialists in 466 stocks when the 1933 Securities Act was passed, but their numbers decreased over the years and, by 1967, no NYSE-listed securities were traded by multiple specialists (see [18], pp. 337–338).

As securities markets around the world face extensive competition, one of the most important questions in market microstructure is whether there is a “best” market structure for all securities that will eventually dominate the landscape. While the specialist market

system was successful in the past, an interesting trend over the past two decades has been the adoption of such a market structure by derivatives exchanges and stock markets around the world for the trading of less active securities. Hence, the viability of this market structure suggests that it fulfills certain important needs of investors. In the following sections, a brief review of some theoretical work on specialist markets is provided, recent empirical evidence on the impact of adopting a specialist system is surveyed, and the manner in which specialists are compensated is discussed.

Theory

Much of the early literature about specialists dealt with the provision of liquidity by dealers without specific attention to whether or not they coexist with an auction mechanism. The manner in which dealers set prices is affected by such considerations as inventory control (see **Inventory Effects**) and information asymmetry in the market (see **Adverse Selection**). The unique hybrid structure of the specialist system, however, gives rise to additional complexity due to the strategic interaction between investors who want to use the auction component of the market (e.g., limit order traders) and the specialists.

Conroy and Winkler [6] show in a simple setting that competition from public limit orders forces the specialist to offer a narrower spread. Rock [16] provides an insightful analysis of the strategic interaction between limit order traders and the specialist when some investors in the market have private information. He notes that a trader who submits a limit order to the book does not know *ex ante* whether his or her order will be executed by a small order arriving from an uninformed trader or as part of a large order coming from an informed trader. What makes the specialist on the NYSE different from other traders is the ability to condition his or her participation as a dealer and the prices in which he or she transacts on the size of the arriving market order (due to the specialist’s control over the execution of trades).

The presence of the specialist imposes adverse selection on the limit order traders because he or she takes the smaller market orders (most likely uninformed and hence profitable) and leaves the larger market orders (which are more likely to be motivated by private information) to transact

against the book. Without the specialist, (uninformed) small market orders transact at worse prices because the *ex ante* nature of prices in the limit order book pools them together with the (informed) large market orders. The specialist's ability to identify small market orders and offer them better prices helps small traders.

Parlour and Seppi [15] use this insight to show that a hybrid market comprised of a limit order book and a specialist can coexist and even dominate a pure limit order book. The interest in this comparison stems from Glosten's [10] claim that the electronic open limit order book architecture is somehow "inevitable". Crucial to Glosten's results, however, are the assumptions of order flow anonymity and parity in terms of costs among liquidity providers, both of which need not hold in the case of the specialist.

While the answer to whether hybrid specialist markets can compete effectively against electronic limit order books could depend on specific trading arrangements and structural features, the most recent theoretical insight is that these two markets are more similar than different. Back and Baruch [3] allow traders to "work" the orders by splitting their demand and trading it over time in smaller increments. When traders behave in this fashion, the *ex post* pricing power of the specialist assumed in [15, 16] and [19] no longer gives the specialist an advantage. The reason is that even if the specialist can condition on the size of the arriving market order, he does not know whether it is part of a much larger order of an informed trader that is worked over time (i.e., he cannot condition on the size of the demand underlying the order). The specialist market structure therefore does not offer any general advantage over the pure limit order book.^a

It is still possible that the presence of a specialist would be more beneficial for a certain subset of the securities in the market. Grossman and Miller [11] model risk averse liquidity provision and examine the demand for dealer services that is created by investors' desire for immediacy of execution. They note that the specialist is far more important for the less active stocks. Profitable dealing does not arise endogenously in those stocks even in the absence of adverse selection considerations because the potential for profit is small relative to the fixed cost of maintaining a market presence (after taking into account the risk of holding an

inventory of these stocks). Sabourin [17] models the interaction of a limit order book and a designated market maker and shows that this hybrid structure offers more attractive prices than a pure limit order book for stocks with high fundamental volatility.

Specialists and Market Quality

The recent introduction of designated market makers to derivatives exchanges and stock markets around the world affords us the opportunity to empirically investigate the benefits of adopting the specialist market structure. Anand and Weaver [2] show that spreads declined significantly when the Chicago Board of Options Exchange (CBOE) introduced designated market makers in 1999. On the Chicago Board of Trade, the introduction of a designated market maker to the 10-year interest rate futures contract in 2002 resulted in increased volume, reduced transaction costs, and improved price efficiency [22]. Similarly, volume increased substantially and transaction costs declined when the Tel Aviv Stock Exchange introduced designated market makers to its thinly traded shekel-euro options in 2004 [8].

Several European stock exchanges implemented programs whereby less active stocks are given a choice of adding a designated market maker to accompany the auction market (usually a limit order book), effectively creating a hybrid specialist market structure. These initiatives appear to have resulted in significant improvement in market quality. For example, volume increased and spreads decreased after designated market makers were introduced on the Italian Stock Exchange [14]. Liquidity improved and either volatility or liquidity risk declined when such programs were instituted by the Stockholm Stock Exchange and Euronext [1] and [13]. Market quality also improved for Paris Bourse stocks traded in call auctions when designated market makers were employed [23].

The aforementioned research documented significant benefits in terms of market quality to having a specialist market structure for less actively traded stocks and derivatives contracts. A natural question is, therefore, how can specialists be compensated to ensure the provision of their services in these securities?

The Compensation of Specialists

Specialists are expected to lose from trading with better-informed investors. They also incur costs for keeping a market presence, and due to risk aversion would require an appropriate compensation for deviating from their optimal positions in risky securities. How do specialists make money? Investors' liquidity shocks generate price reversals in the market and create profitable trading opportunities for specialists in their capacity as dealers [11]. Beyond this mechanism, other opportunities for specialists critically depend on the rules that govern market interaction and specific devices instituted to compensate them.

For example, a significant source of revenue for NYSE specialists has been the fees earned from their brokerage function [20], but this part of their revenue has declined over the years. In some markets (e.g., the NYSE), specialists have the ability to better discern uninformed trading (if due to repeated interaction with brokers on the floor or through *ex post* pricing that allows conditioning on order size).^b In addition, designated market makers in certain markets (e.g., CBOE, NYSE) enjoy privileged access to information on the content of the limit order book, which enhances their ability to trade profitably.

Another mechanism used (e.g., on CBOE) to ensure profitability of dealing by designated market makers is to give them *pro rata* share of the order flow for execution even if they only match, rather than establish, the best prices in the market. For less active securities, however, Grossman and Miller [11] suggest that it may be impossible for specialists to make sufficient money by dealing to cover their costs. Markets have offered designated market makers reduced fees (to give them a cost advantage over other traders) or direct payments (see [8]) in exchange for a commitment to make markets continuously and provide quality quotes that satisfy minimum depth and maximum spread requirements.

Furthermore, if liquidity is priced (see **Liquidity Premium**), then the benefit of greater liquidity of a firm's stock would ultimately accrue to the firm in terms of lower cost of capital. As such, the firm should be willing to pay for the services of a designated market maker. This has been the model that was recently adopted by several European stock exchanges. At the center of the market architecture of these exchanges is a transparent electronic limit order book, and therefore the designated

market makers do not have *ex post* pricing ability, guaranteed order flow, or an informational advantage as to the content of the book. Being on the same footing as other traders who do not have an obligation to maintain continuous presence subject to minimum depth and maximum spread rules, the designated market makers can hardly be expected to make money (at least in less active stocks). Therefore, compensation arrangements include direct payments from the firm to the designated market makers, as well as indirect payments in the form of future business prospects like secondary offerings, banking services, or investor relations services [1, 13], and [23].

Conclusion

The hybrid structure of specialist markets combines the ability of investors to trade directly with other investors in an auction setting together with the presence of designated market makers. Empirical research documents improvements in market quality when less active securities move from a pure limit order book structure to a hybrid specialist structure.

Bessembinder *et al.* [5] claim that a designated market maker subject to a maximum spread rule can increase the allocational efficiency of prices because more investors would choose to trade when transactions costs are lower. In addition, [5] and [7] note that the informational efficiency of prices could be enhanced by the presence of a designated market maker. The reason is that obligations of the specialists to maintain price continuity or narrow spreads increase the profitability of informed trading and therefore more investors optimally choose to invest in gathering and analyzing information. As yet, an unanswered question is whether the societal benefits to investors outweigh the costs of compensating the specialists.

End Notes

^aGlosten [9] and Leach and Madhavan [12] model the specialist as a monopolist dealer without analyzing the hybrid market interaction between the specialist and the limit order book. Both papers show that a monopolist specialist market may be robust to informational asymmetries in that it can remain open when a market populated by competitive dealers would shut down.

^b See [4] for a model of how repeated interactions with floor brokers enable the specialists to offer investors better terms of trade.

References

- [1] Anand, A., Tanggaard, C. & Weaver, D.G. (2005). *Paying for Market Quality*. Syracuse University working paper.
- [2] Anand, A. & Weaver, D.G. (2006). The value of the specialist: empirical evidence from CBOE, *Journal of Financial Markets* **9**, 100–118.
- [3] Back, K. & Baruch, S. (2007). Working orders in limit-order markets and floor exchanges, *Journal of Finance* **62**, 1589–1621.
- [4] Benveniste, L.M., Marcus, A.J. & Wilhelm, W.J. (1992). What's special about the specialist? *Journal of Financial Economics* **32**, 61–86.
- [5] Bessembinder, H., Hao, J. & Lemmon, M. (2007). *Why Designate Market Makers? Affirmative Obligation and Market Quality*. University of Utah working paper.
- [6] Conroy, R.M. & Winkler, R.L. (1986). Market structure: the specialist as dealer and broker, *Journal of Banking and Finance* **10**, 21–36.
- [7] Dutta, P.K. & Madhavan, A. (1995). Price continuity rules and insider trading, *Journal of Financial and Quantitative Analysis* **30**, 199–221.
- [8] Eldor, R., Hauser, S., Pilo, B. & Shurki, I. (2006). The contribution of market makers to liquidity and efficiency of options trading in electronic markets, *Journal of Banking and Finance* **30**, 2025–2040.
- [9] Glosten, L.R. (1989). Insider trading, liquidity, and the role of the monopolist specialist, *Journal of Business* **62**, 211–235.
- [10] Glosten, L.R. (1994). Is the electronic open limit order book inevitable? *Journal of Finance* **49**, 1127–1161.
- [11] Grossman, S.J. & Miller, M.H. (1988). Liquidity and market structure, *Journal of Finance* **43**, 617–633.
- [12] Leach, J.C. & Madhavan, A.N. (1993). Price experimentation and security market structure, *Review of Financial Studies* **6**, 375–404.
- [13] Menkveld, A.J. & Wang, T. (2008). *How do Designated Market Makers create value for Small-Caps?*. VU University of Amsterdam working paper.
- [14] Nimalendran, M. & Petrella, G. (2003). Do 'thinly-traded' stocks benefit from specialist intervention? *Journal of Banking and Finance* **27**, 1823–1854.
- [15] Parlour, C.A. & Seppi, D.J. (2003). Liquidity-based competition for order flow, *Review of Financial Studies* **16**, 301–343.
- [16] Rock, K. (1990). *The Specialist's Order Book and Price Anomalies*. Harvard University working paper.
- [17] Sabourin, D. (2006). *Are Designated Market Makers Necessary in Centralized Limit Order Markets?* University of Paris IX Dauphine working paper.
- [18] Seligman, J. (2003). *The Transformation of Wall Street*, 3rd Edition, Aspen Publishers, New York, NY.
- [19] Seppi, D.J. (1997). Liquidity provision with limit orders and a strategic specialist, *Review of Financial Studies* **10**, 103–150.
- [20] Stoll, H.R. (1985). The stock exchange specialist system: an economic analysis, in *Monograph Series in Finance and Economics*, A. Saunders, ed, New York University, New York, Vol. 2, pp. 1–51.
- [21] Stoll, H.R. (1998). Reconsidering the affirmative obligation of market makers, *Financial Analysts Journal* **54**, 72–82.
- [22] Tse, Y. & Zabolina, T. (2004). Do designated market makers improve liquidity in open-outcry futures markets? *Journal of Futures Markets* **24**, 479–502.
- [23] Venkataraman, K. & Waisburd, A.C. (2007). The value of the designated market maker, *Journal of Financial and Quantitative Analysis* **42**, 735–758.

Related Articles

Adverse Selection; Bid–Ask Spreads; Call Auction Markets; Inventory Effects; Limit Order Markets; Liquidity Premium; Order Types.

GIDEON SAAR

Order Types

Most modern financial markets are *limit order markets* and use a continuous double-auction mechanism [5]. On these markets, investors can choose between different *order types* when they wish to buy or sell stocks. Each of them have specific characteristics that make them appropriate for certain types of investors or in certain situations.

The number of different order types supported on the different exchanges is actually quite large and brokers might also add complexity and propose yet more sophisticated order types to their customers. This article only reviews the most frequently used order types. As a conclusion, some empirical facts on order placement and the performance of market *versus* limit orders are highlighted.

Market Orders

The simplest order type is perhaps the *market order*: the investor chooses to buy or sell at the current best available price. As long as there are willing buyers and sellers, the order is guaranteed to be executed. An important property of a market order is that the investor only specifies which quantity he/she wants to trade, not at which price. Indeed, on fast-moving markets, he/she may well obtain an execution price that differs from the best price quoted at the moment he/she decided to send the order. It may also happen that he/she receives different prices for parts of the order, in particular for large orders.

Limit Orders

Unlike a market order, a *limit order* is an order to buy or sell at a specific price. A buy (sell) limit order can only be executed at the limit price or lower (higher). Thus, the investor has some control over the execution price that he/she will obtain, but the order might not be filled. One can distinguish two types of limit orders: limit orders that are submitted at or above the market price and those that are not. The first ones, sometimes called *marketable limit orders*, are immediately executed. Those of the second kind are stored in a queue called the *limit order book*. They

are executed if they are hit by market orders from the opposite side.

Stop Orders

A *stop order* is an example of a conditional order: a market order is sent once the price of a stock reaches a specified level. This type of order is used by investors who want to avoid big losses or protect profits without having to monitor the stock performance for an extended period of time. For instance, if an investor has a short position, he/she might decide to buy it back once his/her losses reach a threshold, that is, when the stock price exceeds a certain level. In the same way, he/she can limit his/her profits by buying the stocks when the price drops below a specified price. Stop orders might also be limit orders. In that case, a limit order is sent when the price reaches a level.

Hidden Orders and Iceberg Orders

Investors who wish to submit a large limit order might be concerned when their order gets noticed. A possibility is then to submit the *hidden order*. Such an order is not visible in the order book, but has usually lower time priority, that is, regular limit orders at the same price that are submitted later will be executed first. An *iceberg* (or *reserve*) order is similar except that it is not fully hidden: a fraction of the order size, called *peak volume*, is publicly disclosed. When the visible volume is filled and a hidden volume is still available, a new peak volume enters the book. On exchanges where iceberg orders are allowed, they can be quite frequent. For example, according to studies of iceberg orders on Euronext [1], 30% of the book volume is hidden.

Immediate or Cancel Orders

If a market participant wants to profit from a trading opportunity that he/she expects to last only a short period of time, he/she can consider sending a *immediate or cancel* order. This order either (partially) gets filled or is else cancelled. Thus, it never goes into the book and leaves no visible trace if it cannot be executed. This order type is similar to the *all or nothing* (sometimes called *fill or kill*) order. Orders of the

latter type must be completely executed or not at all. In this way, the investor avoids revealing his intention to trade if the entire quantity was not available.

Day Orders and Good 'Til Canceled Orders

An investor might want his/her order to expire if it is not executed. He/she can then submit a *day order*. Such an order is canceled if it is not executed on the day the order is placed. On the other hand, if he/she submits a good 'til canceled order (GTC), the order will remain active until the investor cancels it. Typically, his/her broker will, however, set a maximum time limit of 30–90 days after which the broker cancels the order.

Market-on-close Orders

A market-on-close (MOC) order is a market order that is submitted to execute as close to the closing price as possible. If a double-sided closing auction takes place at the end of the day (such as on the New York Stock Exchange (NYSE) or the Tokyo Stock Exchange), the order will participate in the auction. Investors might want to trade near the close because they expect liquidity to be high. Another reason might be that closing prices are widely used as a reference price (for instance, for the valuation of mutual fund shares or the determination of the payoffs to cash-settled derivatives [5]). An MOC order is then a way to obtain an execution price close to the reference price. There also exist *limit on close* orders, that is, limit orders submitted at the close. Finally, there are limit order that become market orders at the close (in Japan called *Funari* orders).

Reg NMS Order Types

In reaction to recent National Market System regulations (Reg NMS) promulgated by the United States Securities and Exchange Commission (SEC), several American exchanges developed new order types. In fact, according to Reg NMS brokers should respect the so-called Order Protection Rule, which states that one cannot ignore the top-of-book best bid and offers. This rule came into effect during the year

2007. Suppose that the best ask price for a given security is \$123.45 on Nasdaq, \$123.46 on the NYSE, and \$123.46 or more on the other market centers where it is quoted. The \$123.45 quote on Nasdaq is then protected and cannot be traded through, that is, if one sends a buy market order to the NYSE, the exchange has in principle to route it to Nasdaq. A new order type that is Reg.NMS compliant is a *do not ship* order, an order that is to be executed at one exchange. If the order would route to another market center, it will be cancelled. Another example is a Intermarket Sweep Order (ISO). This is a limit order designated to be executed exclusively at one market center. The sender fulfills the Reg NMS obligations by sending other orders to any market center quoting better prices. For more examples, see [9].

Peg Orders

With the competition between liquidity pools in the US and in Europe, some recently created venues such as Chi-X or BATS have introduced *peg orders*. These are orders that “follow” quotes from other exchanges (either the primary exchange or the best bid or offers). The venues thus hope to attract trading volume. For market participants, a peg order is convenient because they do not need to monitor the quotes on other exchanges and replace their order themselves if the price moves away. Also, alternative trading systems might be cheaper (lower exchange fees) and faster (lower latency).

Empirical Studies of Order Placement

The question of how investors submit their orders determines how the price is formed at the microstructure level. It has recently received much interest, triggered by the availability of huge datasets of order book data.

The probability that a limit order is sent at a price p has been shown to decay with the distance from the best price. However, the decay is slow, that is, algebraic [7]. It can be fitted by a Student distribution [6]. The distribution is symmetric, centered around the best price and the same for buying and selling. The probability that an order will be cancelled also decreases with its distance from the best

price. These findings have been observed across different markets such as the London Stock Exchange, Euronext, and Instinet and are expected to be to some degree “universal”. The distribution of the time to fill for executed limit orders and the time to cancel for canceled ones was studied in [2].

Performance of Market versus Limit Orders

The performance of the different order types is an issue of interest for both the practitioners and academic researchers. Although the cost of a market order can be considered to be half the bid-ask spread, the cost of a limit order is not unambiguously defined. In case the order is not executed because of adverse price movement, one should in some way add the cost of the missed opportunity. Moreover, if the order does get filled, it might have been hit by an investor with more information about the future price, a phenomenon that is called *adverse selection*.

In an early study, Harris and Hasbrouck [4] suggest that limit orders are performing better, even taking into account some penalty for nonexecuted orders. According to Handa *et al.* [3], the order-driven market is an ecological system with a competition between market and limit orders. This idea has been made more quantitative in [8], where it was suggested that on average limit orders and market orders have equal costs. Imposing that strategies that participate in a infinitesimal fraction of all market or limit orders have at best a marginal profit, one obtains a linear relation between bid-ask spread and the instantaneous impact of market orders, which is empirically verified. Thus, an investor who would try to earn from the spread by putting limit orders in the book would actually not

make money due to the impact of incoming market orders.

References

- [1] D'Hondt, C., De Winne, R. & François-Heude, A. (2003). *Hidden Orders on Euronext: Nothing is Quite as it Seems*, working paper.
- [2] Eisler, Z., Kertesz, J., Lillo, F. & Mantegna, R.N. (2007). *Diffusive Behavior and the Modeling of Characteristic Times in Limit Order Executions*, Technical Report, arXiv:physics/0701335.
- [3] Handa, P., Schwartz, R.A. & Tiwari, A. (1998). The ecology of an order-driven market, *Journal of Portfolio Management* **24**(2), 47–55.
- [4] Harris, L. & Hasbrouck, J. (1996). Market vs. limit order: the SuperDOT evidence on order submission strategy, *The Journal of Financial and Quantitative Financial Analysis* **31**(2), 213–231.
- [5] Hasbrouck, J. (2007). *Empirical Market Microstructure*, Oxford University Press.
- [6] Mike, S. & Farmer, J.D. (2008). An empirical behavioral model of liquidity and volatility, *Journal of Economic Dynamics and Control* **32**, 200–234.
- [7] Potters, M. & Bouchaud, J.P. (2002). More statistical properties of order books and price impact, *Physica A* **324**(1–2), 133–140.
- [8] Wyart, M., Bouchaud, J.P., Kockelkoren, J., Potters, M. & Vettorazzo, M. (2008). Relation between bid-ask spread, impact and volatility in order-driven markets, *Quantitative Finance* **8**(1), 41–57.
- [9] www.nyse.com/pdfs/nyse_regnms_updates_final.pdf; www.nasdaqtrader.com/content/marketregulation/regnms/mopp.pdf (accessed Oct 2008).

Related Articles

Algorithmic Trading; Limit Order Markets; Order Flow.

JULIEN KOCKELKOREN

Automated Trading

Technological change has revolutionized the way financial assets are traded. Investors place orders *via* computer rather than speaking to a broker on the phone. Trading floors have largely been replaced by electronic trading platforms [15]. The nature of order execution has changed dramatically as well, as many market participants now employ algorithmic trading (AT), commonly defined as the use of computer algorithms to manage the trading process. Starting from near zero in the 1990s, by 2007 AT was responsible for roughly one-third of trading volume in the United States and is expected to account for perhaps half of trading volume by 2010 [8]. The intense activity generated by algorithms threatens to overwhelm exchanges and market data providers [9], forcing significant upgrades to their infrastructures [8].

Before algorithmic trading, a pension fund manager wanting to buy 100 000 shares of IBM might have hired a broker-dealer to search for a counterparty to execute the entire quantity at once in a block trade. Alternatively, that institutional investor might have hired a broker to quietly “work” the order (possibly on the floor of the New York Stock Exchange), using his/her judgment and discretion to buy a little bit here and there over the course of the trading day to keep from driving the IBM share price up too far (see [16] for evidence of large orders being broken up). As trading became more electronic, it became easier and cheaper to replicate the human trader with a computer program doing algorithmic trading. Now virtually every large broker-dealer offers a suite of algorithms to its institutional customers to help them execute orders in a single stock, in pairs of stocks, or in baskets of stocks. Algorithms typically determine the timing, price, and quantity of orders, dynamically monitoring market conditions across different securities and trading venues, reducing market impact by optimally (and possibly randomly) breaking large orders into smaller pieces, and closely tracking benchmarks such as the volume-weighted average price (VWAP) over the execution interval.

Algorithms for Executing Large Orders

Execution of trades over a trading horizon involves complex dynamic optimization problems to determine order size, order frequency, and order type.

Bertsimas and Lo [3] study optimal execution strategies in the presence of temporary price impacts. Conditional on the constraint of completing the entire transaction by a fixed date, orders are broken into pieces so as to minimize cost. Almgren and Chriss [1] extend this by considering the risk that arises from breaking up orders and slowly executing them. In the basic Almgren and Chriss setup, the authors first assume an initial holding of $x_0 = X$ units of a security that must be completely liquidated before time T . Dividing T into N intervals of length $\tau = T/N$ and defining the discrete times $t_k = k\tau$, for $k = 0, \dots, N$, they solve for the optimal holdings, x_k , at each time t_k . Equivalently, this can be formulated as the series of trades to accomplish this liquidation, with $n_k = x_k - x_{k-1}$ being the number of units sold between times t_{k-1} and t_k . To capture the risk involved in the trading process (see [10] for the relationship between execution risk and investment risk), a price process for the security must be specified:

$$S_k = S_{k-1} + \sigma \tau^{1/2} \xi_k - \tau g\left(\frac{n_k}{\tau}\right) \quad (1)$$

for $k = 0; \dots, N$. Here σ represents the volatility of the asset, the ξ_k are draws from independent random variables each with zero mean and unit variance, and the permanent price impact of the trade, $g(v)$, is a function of the average rate of trading $v = n_k/\tau$ during the interval t_{k-1} and t_k . In addition, there may be a temporary price impact function, $h(v)$, which captures the temporary drop in average price per share caused by trading at average rate v during a time interval. Thus, the actual price per share received on sale k is

$$\tilde{S}_k = S_{k-1} - h\left(\frac{n_k}{\tau}\right) \quad (2)$$

where $h(v)$ does not impact the subsequent market price S_k . $g(v)$ and $h(v)$ can be chosen to reflect the trader’s beliefs about market dynamics and market microstructure.

By combining these and summing across periods, the proceeds of the liquidation are obtained:

$$\begin{aligned} \sum_{k=1}^N n_k \tilde{S}_k &= X S_0 + \sum_{k=1}^N \left(\sigma \tau^{1/2} \xi_k - \tau g\left(\frac{n_k}{\tau}\right) \right) x_k \\ &\quad - \sum_{k=1}^N n_k h\left(\frac{n_k}{\tau}\right) \end{aligned} \quad (3)$$

with the first term being the initial market value, the middle term including both the volatility and the permanent price impacts, and the final term being the temporary price impacts. Calculating the expected cost of trading, $C(x)$, as the difference between the initial value and the proceeds of the sales, we get

$$C(x) = \sum_{k=1}^N x_k g\left(\frac{n_k}{\tau}\right) + n_k h\left(\frac{n_k}{\tau}\right) \quad (4)$$

The variance of $C(x)$ is

$$V(C(x)) = \sigma^2 \tau \sum_{k=1}^N x_k^2 \quad (5)$$

Using this setup, Almgren and Chriss calculate the mean-variance efficient frontier for trading/liquidation strategies. Furthermore, assuming the objective is to minimize the function $C(x) - \lambda V(C(x))$, they calculate the optimal strategies for various values of λ . Almgren and Chriss show that as long as the parameters are fixed and known in advance, the optimal trading strategy can be calculated in advance and does not need to be updated dynamically. The calculation of such strategies may still be complex depending on the functional form of the permanent and transitory price impact functions $g(v)$ and $h(v)$. Some typical impact functions include power functions, weighted sums of linear and nonlinear functions, and exponential decay functions. The significant computer power often required to find the optimum under complex price impact functions makes implementation *via* algorithms attractive.

Numerous features have been added to the above model that significantly increase the complexity of the calculations and algorithms needed to find and implement the optimal trading strategies. Some of the most common enhancements include portfolio considerations and modeling the information that motivates the trade. Obizhaeva and Wang [20] optimize trading strategies assuming that liquidity does not replenish immediately after it is taken but only gradually over time. Algorithms can be improved by modeling variation within and across days to account for changing market conditions and predictable price movements [2]. Kissell and Malamut [17] discuss how trading should be sped up or slowed down conditional on traders' beliefs about whether past prices changes will mean reverting or continue. Many brokers build models with these considerations into their AT products that they sell to their clients.

The above optimal execution approaches either abstract away from the market mechanism or assume liquidity is taken *via* market orders. Implementation strategies that allow for more passive trading strategies enable the algorithm/trader to choose the type and aggressiveness of each order/component of the larger transaction. Cohen *et al.* [6] and Harris [13] focus on the simplest choice: market order *versus* limit order. If a trader chooses a nonmarketable limit order, the aggressiveness of the order is determined by its limit price [12] and [21]. How aggressively a limit order should be priced depends on how price affects the time to execution [19]. For assets that trade on multiple venues, determining where to send orders adds further complexity.

Market-making Algorithms

Many observers think of algorithms from the standpoint of institutional buy-side investor, but there are other important users of algorithms. Some hedge funds and broker-dealers supply liquidity using algorithms, competing with designated market-makers and other liquidity suppliers.

Most models take the traditional view that one set of traders provides liquidity *via* quotes or limit orders and another set of traders initiates a trade to take that liquidity—for either informational or liquidity/hedging motivations. Liquidity supply involves posting firm commitments to trade. These standing orders provide free trading options to other traders. Using standard option pricing techniques, Copeland and Galai [7] value the cost of the option granted by liquidity suppliers. The arrival of public information renders existing orders stale and can put the free trading option into the money. Biais *et al.* [4] provide a general theoretical framework for a liquidity supply price schedule $q_t(p)$ at time t :

$$\max_{q_t(p)} EU(C_t + I_t v_t + (v_t - p)q_t(p) | H_t) \quad (6)$$

where $U(x)$ is the utility function, C_t is the cash, I_t is the inventory of the security, v_t is the expected fundamental value, and H_t is the information set available. The use of algorithms in managing this price schedule (quotes) lowers the cost of monitoring changes in the information set due to changes in information and prices regarding that security or any related security. Any time an announcement,

trade, or quote change occurs in any security, a liquidity provider may want to revise their prices and quantities. Foucault *et al.* [11] study the equilibrium level of effort that liquidity suppliers should expend in monitoring the market to avoid this picking off risk. Improvements in technology allow algorithms to inexpensively perform such monitoring. In this way, AT may be a way to implement Black's [5] idea of limit orders indexed to a market index.

The widespread adoption of algorithms by traders supplying and demanding liquidity has led some observers to liken the situation to an arms' race [18]. While evidence points to algorithmic trading improving liquidity [14], the growing dominance of algorithms in trading makes their impact worthy of continued study.

References

- [1] Almgren, R. & Chriss, N. (2000). Optimal execution of portfolio transactions, *Journal of Risk* **3**, 5–39.
- [2] Almgren, R. & Lorenz, J. (2006). Bayesian adaptive trading with a daily cycle, *Journal of Trading* **1**, 38–46.
- [3] Bertsimas, D. & Lo, A.W. (1998). Optimal control of execution costs, *Journal of Financial Markets* **1**, 1–50.
- [4] Biais, B., Glosten, L. & Spatt, C. (2005). Market microstructure: a survey of microfoundations, empirical results and policy implications, *Journal of Financial Markets* **8**, 217–264.
- [5] Black, F. (1995). Equilibrium exchanges, *Financial Analysts Journal* **51**, 23–29.
- [6] Cohen, K., Maier, S., Schwartz, R. & Whitcomb, D. (1981). Transaction costs, order placement strategy and existence of the bid-ask spread, *Journal of Political Economy* **89**, 287–305.
- [7] Copeland, T.E. & Galai, D. (1983). Information effects on the bid-ask spread, *Journal of Finance* **38**, 1457–1469.
- [8] Economist (2007). Ahead of the tape—algorithmic trading, *Economist* **383**, 85.
- [9] Economist (2007). Dodge tickers—stock exchanges, *Economist* **382**, 74.
- [10] Engle, R.F. & Ferstenberg, R. (2007). Execution risk, *Journal of Portfolio Management* **33**, 34–45.
- [11] Foucault, T., Roell, A. & Sandas, P. (2003). Market making with costly monitoring: an analysis of the SOES controversy, *The Review of Financial Studies* **16**, 345–384.
- [12] Griffiths, M., Smith, B., Turnbull, D.A. & White, R. (2000). The costs and determinants of order aggressiveness, *Journal of Financial Economics* **56**, 65–88.
- [13] Harris, L. (1998). Optimal dynamic order submission strategies in some stylized trading problems, *Financial Markets, Institutions, and Instruments* **7**, 1–76.
- [14] Hendershott, T., Jones, C. & Menkveld, A. (2008). *Does Algorithmic Trading Improve Liquidity?* Technical Report, Working Paper, UC Berkeley.
- [15] Jain, P.K. (2005). Financial market design and the equity premium: electronic versus floor trading, *Journal of Finance* **60**, 2955–2985.
- [16] Keim, D. & Madhavan, A. (1995). Anatomy of the trading process: empirical evidence on the behavior of institutional traders, *Journal of Financial Economics* **37**, 371–398.
- [17] Kissell, R. & Malamut, R. (2006). Algorithmic decision making framework, *Journal of Trading* **1**, 12–21.
- [18] Leinweber, D. (2007). Algo vs. Algo, *Institutional Investor's Alpha* **2**, 45–51.
- [19] Lo, W.A., MacKinlay, A.C. & Zhang, J. (2002). Econometric models of limit-order executions, *Journal of Financial Economics* **65**, 31–71.
- [20] Obizhaeva, A. & Wang, J. (2005). *Optimal Trading Strategy and Supply/Demand Dynamics*. MIT working paper.
- [21] Rinaldo, A. (2004). Order aggressiveness in limit order book markets, *Journal of Financial Markets* **7**, 53–74.

Related Articles

Algorithmic Trading; Bid-Ask Spreads; Execution Costs; Price Impact; Specialist Markets.

TERRENCE HENDERSHOTT

Market Transparency

Market transparency is the degree to which information is widely available to market participants; this overview of transparency will be confined to the transparency of the market trading process without considering issues related to disclosure about the fundamentals of the securities that are traded.^a Given the importance of fleeting informational advantages in determining trading profits and losses, it is not surprising that transparency is and has been one of the most contentious issues in the design and regulation of trading systems. Financial economists' research—including theoretical modeling, experimental work, and empirical research—has also yielded a wide range of insights and results that provide a fairly ambivalent view of whether imposing transparency is desirable, or even feasible. There are conflicting interests between market participants along a variety of dimensions: informed or uninformed traders, liquidity-seeking (“buy-side”) *versus* liquidity-supplying (“sell-side”) agents, large *versus* small traders, broker-dealers *versus* agency brokers, as well as trading venues competing for transactions and listings. These gains and losses shape the trading and regulatory landscape.

The degree of transparency varies widely across instruments and markets. The United States market for major listed stocks is traditionally the most transparent. Since the 1975 amendments to the Securities Exchange Act, consolidated market information has been made available through designated exclusive securities information processors (SIP) that aggregate both quote information and trade reports. Over the years, the amount of information made available has been upgraded considerably.

At the other extreme lie the markets in thinly traded securities like municipal bonds, which are typically traded over the counter (OTC) by broker-dealers who fill most of their customers' orders in a principal capacity, and who trade among themselves in an interdealer market to manage their inventories. No firm prices are quoted, and information regarding the prices of completed trades is only made available for the most frequently traded issues, and that too only with a 1-day lag.

Market participants' information can include *pretrade* or quote transparency—the accurate real-time display of available price quotes, especially for

large-size trades, and of buying and selling interest in general—as well *posttrade* transparency or trade publication—a publicly available real-time record of the recent history of trades, including their time, size, price, location, and perhaps even the identities of the counterparties involved. We start by discussing research on each facet of transparency before touching on policy issues.

Pretrade or Quote Transparency

Most theoretical models argue that pretrade transparency benefits the buy side. Several mechanisms have been suggested.

Biais [9] considers a setting where market makers differ in their inventories (and thus in their reservation value for the security) but are otherwise taken to be approximately risk neutral and in agreement on the fundamental value of the security. He compares two market structures: (i) a transparent market in which dealer quotes can be observed by all market participants and (ii) an opaque market structure in which competing dealers have no information on the quotes of their peers. Effectively, the former is an English auction and the latter is a sealed-bid auction, and market makers' bidding strategies differ in the two settings. The “revenue equivalence theorem” of auction theory applies to this setting (risk-neutral bidders with independent, identically distributed private values) so that, as Biais concludes, the expected spread will be the same in both market structures. Beyond this stylized setting, textbook auction theory states that revenue equivalence breaks down if the bidders (the market makers) are truly risk averse, their private valuations are correlated or asymmetrically distributed, or there is an imperfectly known common value leading to a winner's curse problem [14]. In particular, within Biais's setting, de Frutos and Manzano [19] treat market-maker risk aversion without approximations and conclude that the opaque market is more liquid.

Yin [46] extends the analysis of Biais [9] by incorporating the idea that quote transparency exerts competitive pressure by eliminating customers' search costs, building on Diamond's [20] classic insight that search costs, no matter how small, can effectively destroy price competition and lead to monopoly pricing. Yin [46] considers the case in which customers

must pay a search cost in the opaque market, finding that the average bid–ask spread is smaller in the transparent market.

Evidence from the municipal bond market [28, 30, 43] illustrates the scope for rent extraction in an opaque setting. Municipal bond trading costs are substantially higher than those for equities, and particularly high for retail-sized trades relative to institutional trades, presumably because the investors are less-sophisticated negotiators. The Securities and Exchange Commission (SEC) report (2004) shows that for retail transactions (around \$10 000), an over-3% difference in the prices charged by different dealers filling orders within the same day is quite common (over 9% of transactions), even though normally intraday fluctuations in the fundamental value of the securities are zero or minimal; meanwhile, for large and more sophisticated institutional transactions (sized around \$1 million) a difference of over 1% is rare (0.4% of transactions) and a difference of over 2% absent in the sample.^b And although it is possible to verify *ex post* the prices of other transactions in a bond on the day of an order, investors rarely avail themselves of this opportunity to check up on their broker’s diligence and probity. Only the professional competence, energy, and sense of duty of the broker, limits the extent of price gouging. The underlying problem is that best execution has no teeth in a setting where individual securities are so thinly traded that it makes no sense for any market maker to spend the time and effort needed to quote continuous prices at which he/she is committed to trade. As a result, there is no firmly quoted, easily identifiable going market price. For more thickly traded securities, market makers are willing to quote prices anyway, because they see enough trading volume to recoup occasional losses to traders who hit stale quotes.

Interestingly, Biais and Green [10] show that the bond markets—both corporate and municipal—have not always been as opaque as they have been in the last decades. Their research shows that when bonds were centrally traded on the New York Stock Exchange (NYSE) in the early decades of the twentieth century, transaction costs for retail investors were substantially lower than they were on the OTC markets to which trading migrated after the 1940s.

A second mechanism whereby pretrade transparency can improve liquidity is analyzed by Pagano and Röell [38], who argue that in a centralized market

where all trading interest is exposed before transactions take place, liquidity providers are better able to gauge the extent of informed trading. This reduces the indirect transfer of informational rents (in the form of transaction costs induced by adverse selection) from uninformed to informed traders.

In early 1997, the SEC instituted new order-handling rules that further enhanced transparency; in particular, the “Limit Order Display Rule” that requires dealers to execute or display customers’ limit orders that improve upon their own (or send them on to a competitor), ensuring that best-priced limit orders are displayed to all market participants; and the “quote rule” requiring dealers who place quotes in multiple trading systems to make their best quotes available to the public. Both led to an immediate and substantial reduction in trading costs [2].

Hendershott and Jones [32] consider the impact of a 2002 regulatory enforcement action, which prompted the Island electronic communications network to suspend the display of the limit order book in its three most actively traded exchange traded funds (ETFs) for over a year. They find that quote display reduces trading costs and retains price discovery on Island, and that both transparency and reduced fragmentation play important roles in enhancing liquidity.

Baruch [3] argues that the introduction of an open limit order book on a monopoly specialist market such as the NYSE has the further advantage of reducing the informational advantage of the NYSE specialist over other liquidity providers. Leveling the playing field in this manner reduces the specialist’s participation rate and profits, and ensures better prices for immediacy demanders at the expense of the sell side. The evidence provided by Barclay *et al.* [2] and Boehmer *et al.* [13] agrees with this analysis, but Madhavan *et al.* [36] document that an earlier similar reform in Toronto led to reduced liquidity, though it did reduce specialist profit as predicted. They argue that too much transparency increases the “free option” cost of posting limit orders [18], leading to order withdrawal and a reduction in market depth, as well as diverting institutional trading volume away from the central market due to concerns about front running. However, more recently, Eom *et al.* [23] do find that pretrade transparency enhances liquidity on Korea’s electronic order-driven market, arguing that the Toronto findings may have been driven by endogeneity bias.

While desirable in theory, the amount of pretrade transparency that is achievable is limited to the extent that market participants are unwilling to publicly display firm quotes or limit orders, so that enforced limit order exposure may lead to a less deep limit order book. Many traders avail themselves of the opportunity to hide their intentions by posting large orders as “hidden orders” that are not publicly displayed in full; the hidden depth only shows up onscreen as and when the visible limit orders are knocked out by trading activity. The popularity of this strategy suggests that it may well substantially increase the depth of the limit order book; see [8] for a description of the use of hidden orders on Euronext Paris. The proliferation of “dark pools” of liquidity also points to the wish to keep trading interest at least to some extent out of the public eye, mitigating problems associated with front running and free options.

Posttrade Transparency

Promptly published information regarding the price, size, and perhaps even the identity of the counterparties involved in recent transactions provides an invaluable guide to market sentiment. Extending such information to all market participants would enable price quoters and limit order placers to compete more effectively, because they would not need to worry as much about the threat of being undercut by peers who have better information regarding recent trades. However, many practitioners have argued that this is not the case and that large traders, in particular, can obtain better prices if their trades are not published. When an order can be observed exclusively by the market maker who fills it, he/she can use the information conveyed by the order to make a profit on subsequent trading, and therefore his/her price quotes for the initial trade will be keener. In this respect, market makers have been likened to the bookmaker who declared himself happy to lose money to a particularly successful punter: “He’s my most valuable client. I always shorten the odds when he bets, and it saves me a fortune.”^c The basic idea was first modeled by Kyle [33] and Röell [41], who argue that under competitive market making, the overall net impact of posttrade opaqueness involves more favorable prices for informed traders and higher transaction costs for uninformed traders overall. Moreover, market makers who are not subjected to last trade publication can outcompete any

who are, so that with competition between multiple exchanges or between on- and off-exchange dealers, order flow is likely to gravitate to the least posttrade transparent setting. Madhavan [35] further shows that, because information-driven trades tend to be on the large side, large uninformed traders whose transactions are similar in size may benefit from opaqueness alongside informed traders, with retail uninformed investors bearing the full burden.

Experimental work by Bloomfield and O’Hara [11, 12] is consistent with these predictions, showing that competing dealers are willing to narrow their price quotes at the start of a trading game in an opaque market in order to capture order flow (though surprisingly, they do not fully recoup the initial losses at a later stage). Moreover, dealers who are not obliged to reveal their trades outcompete those who do have to, and dealers prefer to be nontransparent if given the choice.

Empirical evidence regarding the impact of posttrade transparency is mixed. Gemmill [26] does not discern a significant impact on liquidity from three different trade publication regimes (no publication, 90-min delayed publication and 24-h delayed publication) on the London Stock Exchange. This is surprising in light of the fact that throughout the period studied, London market makers steadfastly and vociferously campaigned against prompt last trade publication on the grounds that it would curtail their ability to offer attractive terms, particularly for wholesale trades. As Gemmill points out, they may have been more concerned with a potential loss of business to off-exchange dealers than with the equilibrium impact on market pricing that is the focus of the study.

Recent work on the institutional market for corporate bonds finds that improved posttrade transparency has led to a dramatic decline in execution costs. Starting in 2002, last trade publication through the TRACE system was phased in for the US investment-grade corporate bond market; and the reporting lag has been reduced over the years from 75 to 15 min. Studies by Bessembinder *et al.* [7], Edwards *et al.* [22], and Goldstein *et al.* [27] all conclude that trade publication led to significant declines in transaction costs. As in the municipal bond market, transaction costs are largest for retail-size trades, and the benefits of transparency have been most significant at the retail end. Moreover, the greater transparency has mitigated the informational advantage of the larger

dealers, reducing their market share. The studies estimate that the reduction in transaction costs as a result of TRACE represents roughly a \$1bn per year reduction in corporate bond dealers' market-making revenue. Bessembinder and Maxwell [6] report that, not surprisingly, many bond traders, analysts, brokers, and salesmen have seen substantial reductions in their earnings and exited the market. Post-TRACE, dealers conduct less fundamental analysis and are less willing to hold inventories, so that locating bonds involves more search over multiple intermediaries. On the positive side, TRACE seems to have facilitated the development of an electronic limit order market (Market Axess) in corporate bonds, and possibly contributed to the NYSE's ongoing initiative to upgrade its trading platform for bonds of listed companies.

If prompt trade publication is desirable, is it enforceable, and does it have unintended consequences? If reported trades are published quickly, market participants will find it in their interest to delay the reporting. Franks and Schaefer [25] point out that trade reporting deadlines can be evaded by the use of "protected trades": market makers informally offer a price to a customer for a large deal (with the understanding that they will, in practice, honor the commitment), but the deal is not officially finalized and confirmed until they have had time to unwind the resulting inventory. This delays the trade report—and its publication—beyond the regulators' intention. Moreover, it reduces immediacy and imposes some execution risk on large traders. Porter and Weaver [39] provide evidence that Nasdaq market makers indeed use late trade reporting to manage the flow of information: trades reported "out of sequence" (thus evading the 90s reporting deadline) are disproportionately large in size and much too common to be attributable to legitimate causes (such as computer glitches). An additional concern is that trading will migrate to venues that do not impose as much posttrade transparency. Many observers of the migration of the wholesale markets in European blue chip equities to London in the early 1990s have argued that the lack of trade publication on SEAQ International was a primary determinant. Europe-wide negotiations on market regulation and transparency have typically pitted the looser, more freewheeling regime of the London wholesale markets against the more centralized transparent regime prevailing in the home markets.

Optimal Transparency

Most studies focus on the effects of market transparency on market liquidity, informational efficiency, and the distribution of trading profits among market participants. Unless the general public's motivation for trading is fully articulated, explicit analysis of Pareto gains is ruled out. Some models do postulate that trading arises as a portfolio rebalancing response of risk-averse agents to random endowments, permitting some measure of analysis of deadweight gains and losses.

Cespa and Foucault [15] study how posttrade transparency affects allocative efficiency. Specifically, they consider a dynamic rational expectations model in the manner of Hellwig [31] in which only a fraction of investors observe transaction prices in real time. They show that it is never Pareto optimal to distribute information on past transaction prices to all investors: in their setting, risk sharing is more efficient when the market is not fully transparent. The reason is that dissemination of information thwarts risk sharing through the so-called Hirshleifer effect as well as intensifying competition among investors.

Anonymity

One aspect of transparency that has attracted considerable attention is the degree to which traders' (or at least their brokers') trading motivations and/or identities are revealed, either pre- or posttrade. The information that is potentially available goes beyond the size, direction, and price of quotes and order flow. We group a discussion of the treatment of such items of information here under the heading of anonymity.

Pretrade Anonymity

Most papers on anonymity focus on pretrade transparency. Some market structures, such as telephone dealership markets or floor markets, allow buy-side investors or their representatives to share information on their trading motives. Naturally, these market structures are nonanonymous. In general, nonanonymous pretrade communication among traders reduces adverse selection and improves liquidity.

For instance, Admati and Pfleiderer [1] show that if customers can "sunshine trade", that is, publicly

preannounce their intention to carry out a large transaction and convincingly establish that their trade is *not* based on superior information, they can improve the price at which they trade. Röell [42] shows that even if such information is conveyed exclusively to the uninformed customer's broker-dealer, at least some of the benefits remain, with the broker-dealer absorbing ("internalizing") roughly half the customer's order flow on own account. In both cases, the customer benefits, though as a result of the withdrawal of a part of the liquidity-driven order flow noise, the market at large becomes less liquid for other traders, namely, both information-based traders as well as those liquidity traders who are unable to certify their lack of information. Easley *et al.* [21] present evidence consistent with the idea that purchased order flow similarly involves cream skimming of order flow that is thought to be noninformation-driven; however, Battalio *et al.* [4] find no adverse impact on the liquidity of the main market when opportunities for internalization are introduced on competing regional exchanges, possibly due to intensified competition as order flow is drawn away.

The converse situation arises when informed traders show their hand. Chung *et al.* [16] consider the impact of the special arrangements whereby Swiss companies can publicly repurchase their own shares on a separate, nonanonymous "second trading line" platform. Price movements around the repurchases are consistent with the notion that share repurchases convey positive information, and the mechanism seems to have a positive impact on the liquidity of the main market, which can arguably be attributed to reduced asymmetries of information.

Benveniste *et al.* [5] emphasize another benefit of nonanonymous pretrade communication: it enables market makers and their clients to enter into long-term implicit contracts. In this way, market makers can incentivize their clients to disclose whether their trade is informed or not and adjust their quotes accordingly. Benveniste *et al.* go on to show that price discrimination between informed and uninformed investors is Pareto-improving.

Recent work considers the impact of transparency regarding the identities of agents on the sell side rather than the buy side.

In electronic limit order markets, the identities of liquidity suppliers can either be suppressed or disclosed to market participants. The effects of providing information about liquidity suppliers' identities are

analyzed by Foucault *et al.* [24] and Rindi [40]. Rindi assumes that limit orders are placed both by informed and uninformed agents, for instance, as in [29]. Disclosure of liquidity suppliers' identities enables market participants to observe informed traders' limit orders. The resulting reduction in the profitability of trading on information discourages the acquisition of information. Paradoxically this means that transparency can sometimes reduce liquidity because when fewer competing traders obtain information, risk averse liquidity suppliers, facing more uncertainty about the asset's value, are less willing to bid aggressively. Foucault *et al.* [24] consider a model in which limit order traders are differentially informed about future volatility. Better informed limit order traders can better control their exposure to the risk of being picked off and therefore earn larger expected profits. These rents, however, are limited as uninformed limit order traders draw inferences about the risk of being picked off from the bids placed in the limit order book. These inferences are noisier in an anonymous environment as limit order traders cannot tell whether quotes are set by informed traders or uninformed traders, leading to different bidding strategies by limit order traders in the anonymous and the nonanonymous market. Consequently, disclosing information about liquidity suppliers' identities should affect market liquidity and the informativeness of the limit order book about future volatility. In particular, this identifies conditions under which the anonymous market is more liquid.

Foucault *et al.* [24] provide empirical evidence consistent with these predictions, using data from Euronext Paris, where the limit order book became anonymous in 2001. They show that the switch to anonymity is associated with an increase in liquidity and a decrease in the informativeness of the limit order book about future volatility. Comerton-Forde *et al.* [17] conduct a similar empirical analysis for a larger sample of stocks listed on the Paris Bourse, the Tokyo Stock Exchange, and the Korea Stock Exchange. They also find that liquidity is larger in the anonymous environment for the three markets considered in their study.

An alternative hypothesis consistent with the evidence that pretrade sell-side anonymity enhances liquidity is proposed by Simaan *et al.* [44], who argue that disclosing liquidity suppliers' identities facilitates collusion. They find that dealers post more

aggressive quotes on electronic communication networks (ECNs) than on Nasdaq, as predicted by the collusion hypothesis.

Posttrade Anonymity

Naik *et al.* [37] consider a model with risk-averse market makers who, after first trading with a member of the public whose trade may have some information content, go on to rebalance the inventory on an interdealer market. They show that as long as market makers do not learn anything beyond the order size from their customer, a regime with prompt publication of the initial order yields a better price for the customer than in an opaque regime because there are no asymmetries of information to thwart optimal risk sharing at the second stage. However, if market makers are able to learn more from their customer than what is contained in the order size alone (in particular, if they learn what proportion of an order is information-driven) they will rebate information-driven order flow to such an extent that customers on average might prefer an opaque market without trade publication, because it offers better insurance against price risk.

Lastly, the identities of the parties to a transaction can be disclosed after a transaction. Waisburd [45] analyzes the effect of providing this type of information, using data from the Paris bourse. He considers a sample of stocks trading in two different regimes: one in which brokers' identities are revealed posttrade and one in which these identities are concealed. He finds that liquidity is lower in the posttrade anonymous regime. Interestingly, these results go in the opposite direction to those obtained for pretrade anonymity by Foucault *et al.* [24] or Comerton-Forde *et al.* [17]; thus, pretrade and posttrade anonymity have distinct effects.

Concluding Comments

Our survey has focused on transparency of the trading process, both in terms of quote display for prospective trades and publication of information regarding completed transactions. It is clear from the current crisis that transparency regarding the solvency of potential trading counterparties plays an important role in the functioning of markets during periods of high volatility. The pros and cons of central clearing

systems in providing collective insurance against counterparty risk is a topical but underanalyzed issue. The question of the pricing of transparency, and the costs of collecting and disseminating data, has also received more attention in the policy debate than in formal academic research.

Lastly, it should be clear that decisions regarding how much transparency is to be imposed are political. The beneficiaries of opaqueness tend to be large players such as bond dealers and banks. Such players have resisted the advent of transparent, centralized electronic trading in the bond markets, and they generally take ownership stakes in the competing platforms as a means of exerting influence. Active regulatory intervention may be needed to promote the interests of market participants who are not organized and influential.

End Notes

^aThere is a large and interesting literature about corporate disclosure (surveyed by Leuz and Wysocki [34]) and the complexity of security design.

^bCNNMoney.com (9/1/2004) reported that the NASD cracked down on major dealers who did not meet their obligation to buy and sell at fair prices. The discrepancies were major: one bond was sold on behalf of a customer for less than half the price that it traded for later in the day; another client received 70¢ on the dollar for a bundle of bonds with a fair value of 97¢.

^cThe *Financial Times*, December 6, 1987.

References

- [1] Admati, A.R. & Pfleiderer, P. (1991). Sunshine trading and financial market equilibrium, *Review of Financial Studies* **4**, 443–481.
- [2] Barclay, M.J., Christie, W.G., Harris, J.H., Kandel, E. & Schultz, P.H. (1999). Effects of market reform on the trading costs and depths of Nasdaq stocks, *Journal of Finance* **54**, 1–34.
- [3] Baruch, S. (2005). Who benefits from an open limit-order book? *Journal of Business* **78**, 1267–1306.
- [4] Battalio, R., Greene, J. & Jennings, R. (1997). Do competing specialists and preferencing dealers affect market quality? *Review of Financial Studies* **10**, 969–993.
- [5] Benveniste, L.M., Marcus, A.J. & Wilhelm, W.J. (1992). What's special about the specialist? *Journal of Financial Economics* **32**, 61–86.
- [6] Bessembinder, H. & Maxwell, W. (2008). Transparency and the corporate bond market, *Journal of Economic Perspectives* **22**, 217–234.

- [7] Bessembinder, H., Maxwell, W. & Venkataraman, K. (2006). Market transparency, liquidity externalities, and institutional trading costs in corporate bonds, *Journal of Financial Economics* **82**, 251–288.
- [8] Bessembinder, H., Panayides, M.A. & Venkataraman, K. (2008). *Hidden Liquidity: An Analysis of Order Exposure Strategies in Electronic Stock Markets*. Available at SSRN: <http://ssrn.com/abstract=956476>.
- [9] Biais, B. (1993). Price formation and equilibrium liquidity in fragmented and centralized markets, *Journal of Finance* **48**, 157–185.
- [10] Biais, B. & Green, R. (2007). *The Microstructure of the Bond Market in the 20th Century*. Unpublished, Carnegie Mellon University.
- [11] Bloomfield, R. & O'Hara, M. (1999). Market transparency: who wins and who loses? *Review of Financial Studies* **12**, 5–35.
- [12] Bloomfield, R. & O'Hara, M. (2000). Can transparent markets survive? *Journal of Financial Economics* **55**, 425–459.
- [13] Boehmer, E., Saar, G. & Yu, L. (2005). Lifting the veil: an analysis of pre-trade transparency at the NYSE, *Journal of Finance* **60**, 783–815.
- [14] Bolton, P. & Dewatripont, M. (2005). *Contract Theory*. The MIT Press, Cambridge, MA.
- [15] Cespa, G. & Foucault, T. (2008). *Insiders-outsiders, Market Transparency and the Value of the Ticker*. CEPR WP **6794**, available at SSRN: <http://ssrn.com/abstract=1117581>.
- [16] Chung, D.Y., Isakov, D. & Pérignon, C. (2007). Repurchasing shares on a second trading line, *Review of Finance* **11**, 253–285.
- [17] Comerton-Forde, C., Frino, A. & Mollica, V. (2005). *The Impact of Limit Order Anonymity on Liquidity: Evidence from Paris, Tokyo and Korea*. Working paper, University of Sydney.
- [18] Copeland, T.E. & Galai, D. (1983). Information effects on the bid-ask spread, *Journal of Finance* **38**, 1457–1469.
- [19] De Futos, M.A. & Manzano, C. (2002). Risk aversion, transparency, and market performance, *Journal of Finance* **57**, 959–984.
- [20] Diamond, P. (1971). A model of price adjustment, *Journal of Economic Theory* **3**, 156–168.
- [21] Easley, D., Kiefer, N.M. & O'Hara, M. (1996). Cream-skimming or profit sharing? The curious role of purchased order flow, *Journal of Finance* **51**, 811–833.
- [22] Edwards, A.K., Harris, L.E. & Piwowar, M.S. (2007). Corporate bond market transaction costs and transparency, *Journal of Finance* **62**, 1421–1451.
- [23] Eom, K.S., Ok, J. & Park, J. (2007). Pre-trade transparency and market quality, *Journal of Financial Markets* **10**, 319–341.
- [24] Foucault, T., Moinas, S. & Theissen, E. (2007). Does anonymity matter in electronic limit order markets?, *Review of Financial Studies* **20**, 1707–1747.
- [25] Franks, J. & Schaefer, S. (1995). Equity market transparency on the London Stock Exchange, *Journal of Applied Corporate Finance* **8**(1), 70–77.
- [26] Gemmill, G. (1996). Transparency and liquidity: a study of block trades on the London Stock Exchange under different publication rules, *Journal of Finance* **51**, 1765–1790.
- [27] Goldstein, M.A., Hotchkiss, E.S. & Sirri, E.R. (2007). Transparency and liquidity: a controlled experiment on corporate bonds, *Review of Financial Studies* **20**, 235–273.
- [28] Green, R., Hollifield, B. & Schurhoff, N. (2007). Financial intermediation and the costs of trading in an opaque market, *Review of Financial Studies* **20**, 275–314.
- [29] Grossman, S.J. & Stiglitz, J.E. (1980). On the impossibility of informationally efficient markets, *American Economic Review* **70**, 393–408.
- [30] Harris, L.E. & Piwowar, M.S. (2006). Secondary trading costs in the municipal bond market, *Journal of Finance* **61**, 1361–1397.
- [31] Hellwig, M.F. (1980). On the aggregation of information in competitive markets, *Journal of Economic Theory*, **22**(3), 477–498, Elsevier.
- [32] Hendershott, T. & Jones, C.M. (2005). Island goes dark: transparency, fragmentation, and regulation, *Review of Financial Studies* **18**, 743–793.
- [33] Kyle, A.S. (1987). *Dealer Competition Against an Organized Exchange*. Unpublished draft.
- [34] Leuz, C. & Wysocki, P.D. (2008). *Economic Consequences of Financial Reporting and Disclosure Regulation: A Review and Suggestions for Future Research*. Available at SSRN: <http://ssrn.com/abstract=1105398>.
- [35] Madhavan, A. (1995). Consolidation, fragmentation, and the disclosure of trading information, *Review of Financial Studies* **8**, 579–603.
- [36] Madhavan, A., Porter, D. & Weaver, D. (2005). Should securities markets be transparent? *Journal of Financial Markets* **8**, 266–288.
- [37] Naik, N.Y., Neuberger, A. & Viswanathan, S. (1999). Trade disclosure regulation in markets with negotiated trades, *Review of Financial Studies* **12**, 873–900.
- [38] Pagano, M. & Röell, A.A. (1996). Transparency and liquidity: a comparison of auction and dealer markets with informed trading, *Journal of Finance* **51**, 579–611.
- [39] Porter, D.C. & Weaver, D.G. (1998). Post-trade transparency on Nasdaq's national market system, *Journal of Financial Economics* **50**, 231–252.
- [40] Rindi, B. (2008). Informed traders as liquidity providers. Anonymity, liquidity and price Formation, *Review of Finance* **12**, 497–532.
- [41] Röell, A.A. (1988). *Regulating Information Disclosure Among Stock Exchange Market Makers*. LSE Financial Markets Group Discussion Papers 51.
- [42] Röell, A.A. (1991). Dual-capacity trading and the quality of the market, *Journal of Financial Intermediation* **1**, 105–124.

- [43] Securities and Exchange Commission. (2004). *Report on Transactions in Municipal Securities*, Office of Economic Analysis.
- [44] Simaan, Y., Weaver, D. & Whitcomb, D. (2003). Market maker quotation behavior and pre-trade transparency, *Journal of Finance* **58**, 1247–1267.
- [45] Waisburd, A. (2003). *Anonymity and Liquidity: Evidence from the Paris Bourse*. Working paper, Texas Christian University.
- [46] Yin, X. (2005). A comparison of centralized and fragmented markets with costly search, *Journal of Finance* **60**, 1567–1590.

Related Articles

Adverse Selection; Intraday Price Efficiency; Liquidity.

THIERRY FOUCAULT, MARCO PAGANO &
AILSA RÖELL

Econometrics of Option Pricing

In 1997, Robert Merton and Myron Scholes were awarded the Nobel Memorial Prize in Economics “for a new method to determine the value of derivatives”. Of course, Fisher Black, who died two years earlier, had been associated with this huge contribution to financial engineering, now known as the *Black–Scholes* (BS) model. The purpose of econometrics of option pricing, as sketched later, is not really to check the empirical validity of this model. It has been widely documented that, by contrast with maintained assumptions of Black and Scholes geometric Brownian motion model, stock returns exhibit both stochastic volatility and jumps. Thus, the interesting issue is not the validity of the model itself. In this article, we focus on the assessment of the possible errors of the BS option pricing formula and on empirically successful strategies to alleviate them. We can get the right pricing and hedging formula with a wrong model. The widespread acceptance of the BS option pricing formula as a benchmark, among academics and investment professionals, is undoubtedly due to its usefulness for pricing and hedging options, irrespective of the unrealistic assumptions of the initial model. This is the reason the largest part of this article is focused on econometric modeling and inference about empirically valid extensions of the BS option pricing formula.

However, it is worth stressing the general econometric content of arbitrage pricing. As first emphasized by Cox *et al.* [20], the message of the BS approach goes beyond any particular specification of the underlying stochastic processes. Arbitrage-free pricing models generally allow to interpret derivative prices as expectations of discounted payoffs, when expectations are computed with respect to an equivalent martingale measure. In this respect, a nice correspondence between the theory of arbitrage pricing and econometrics is worth noting. While option contracts are useful to complete the markets and so to get a unique equivalent martingale measure. While only the historical probability distribution can be estimated from return data on the underlying asset, option prices data allow the econometrician to perform some statistical inference about the relevant martingale measure. Econometric approaches based

on implied volatility and also those used for option pricing that are unrelated to the lognormal benchmark of Black and Scholes, either because they are based on unrelated parametric models or because they introduce some nonparametric components, are discussed.

For sake of brevity, in this article, the only option contracts considered are European calls written on stocks. In the same manner, BS option pricing methodology has since been generalized to price many other derivative securities, and the econometric approaches sketched later can be extended accordingly.

The Information Content of Option Prices

Assume that all stochastic processes of interest are adapted in a filtered probability space $(\Omega, (\mathcal{F}_t), P)$. Under some regularity conditions, the absence of arbitrage is equivalent to the existence of an equivalent martingale measure Q (*see Equivalent Martingale Measures*). Without loss of generality, throughout we consider that the payoffs of options of interest are attainable [25]. Then, the arbitrage-free price of these options is defined, without ambiguity, as expectation under the probability measure Q of the discounted value of their payoff. Moreover, for a European call with maturity T , we characterize its arbitrage price at time $t < T$ as the discounted value at time t of its expectation under the time t forward measure $Q_{t,T}$ for time T . By Bayes rule, $Q_{t,T}$ is straightforwardly defined as equivalent to the restriction of Q on $\mathcal{F}_T$. The density function $dQ_{t,T}/dQ$ is $[B(t, T)]^{-1}(B_t/B_T)$, where B_t denotes the value at time t of a bank account, while $B(t, T)$ is the time t price of a pure discount bond (with unit face value) maturing at time T . If K and S_t denote, respectively, the strike price and the price at time t of the underlying stock, the option price C_t at time t is

$$C_t = B(t, T)E^{Q_{t,T}}\text{Max}[0, S_T - K] \quad (1)$$

Such a formula (1) provides a decomposition of the option price into two components:

$$C_t = S_t \Delta_{1t} - K \Delta_{2t} \quad (2)$$

where

$$\Delta_{2t} = B(t, T)Q_{t,T}[S_T \geq K] \quad (3)$$

and

$$\Delta_{1t} = \Delta_{2t}E^{Q_{t,T}}\left[\frac{S_T}{S_t} \mid S_T \geq K\right] \quad (4)$$

It immediately follows (see [34], pp. 140, 169) that

$$\Delta_{2t} = -\frac{\partial C_t}{\partial K} \quad (5)$$

In other words, a cross section at time t of European call option prices, all maturing at time T but with different strike prices K , provides information about the pricing probability measure $Q_{t,T}$. In the limit, a continuous observation of the function $K \longrightarrow C_t$ (or of its partial derivative $\partial C_t / \partial K$) would completely characterize the cumulative distribution function of the underlying asset return (S_T / S_t) under $Q_{t,T}$. Let us rather consider it through the probability distribution of the continuously compounded net return on the period $[t, T]$:

$$r_S(t, T) = \log \left[\frac{S_T B(t, T)}{S_t} \right] \quad (6)$$

With (log-forward) moneyness of the option measured by

$$x_t = \log \left[\frac{K B(t, T)}{S_t} \right] \quad (7)$$

the probability distribution under $Q_{t,T}$ of the net return on the stock $r_S(t, T)$ is characterized by its survival function deduced from equations (3) and (5) as

$$G_{t,T}(x_t) = -\exp(-x_t) \frac{\partial C_t}{\partial x_t} \quad (8)$$

where

$$\begin{aligned} C_t(x_t) &= \frac{C_t}{S_t} \\ &= E^{Q_{t,T}} \{ \text{Max}[0, \exp(r_S(t, T)) - \exp(x_t)] \} \end{aligned} \quad (9)$$

For the purpose of easy assessment of suitably normalized orders of magnitude, practitioners often prefer to plot the BS-implied volatility $\sigma_{t,T}^{\text{imp}}(x_t)$ rather than the option price $C_t(x_t)$ itself, as a function of moneyness x_t . The key reason that makes this sensible is that, in the BS model, the pricing distribution is indexed by a single volatility parameter σ . Under BS' assumptions, the probability distribution of the net return $r_S(t, T)$ under $Q_{t,T}$ is the normal one with mean $(-1/2)(T - t)\sigma^2$ and variance $(T - t)\sigma^2$. This distribution is denoted by $\aleph_{t,T}(\sigma)$.

Then, the BS-implied volatility $\sigma_{t,T}^{\text{imp}}(x_t)$ is defined as the value of the volatility parameter σ^2 that would generate the observed option price $C(x_t)$ as if the

distribution of net return under $Q_{t,T}$ was the normal $\aleph_{t,T}(\sigma)$. In other words, $\sigma_{t,T}^{\text{imp}}(x_t)$ is characterized as the solution of the following equation:

$$C_t(x_t) = BS_h[x_t, \sigma_h^{\text{imp}}(x_t)] \quad (10)$$

where $h = T - t$ and

$$BS_h[x, \sigma] = N[d_1(x, \sigma, h)] - \exp(x)N[d_2(x, \sigma, h)] \quad (11)$$

where N is the cumulative distribution function of the standardized normal distribution and

$$d_1(x, \sigma, h) = \frac{-x}{\sigma\sqrt{h}} + \frac{1}{2}h\sigma^2 \quad (12)$$

$$d_2(x, \sigma, h) = \frac{-x}{\sigma\sqrt{h}} - \frac{1}{2}h\sigma^2 \quad (13)$$

It is worth recalling that the common use of the BS-implied volatility $\sigma_h^{\text{imp}}(x_t)$ by no means implies that the BS model is thought as a well-specified one. By equation (10), $\sigma_h^{\text{imp}}(x_t)$ is nothing but a known strictly increasing function of the observed option price $C_t(x_t)$. When plotting the volatility smile as a function $x_t \longrightarrow \sigma_h^{\text{imp}}(x_t)$ rather than $x_t \longrightarrow C_t(x_t)$, a convenient rescaling of the characterization (8) of the pricing distribution is simply considered. However, this rescaling depends on x_t and, by definition, produces a flat volatility smile whenever the BS pricing formula is valid in the cross section (for all moneyness levels at a given maturity) for some specific value of the volatility parameter. Note that for this the validity of the BS model itself is only a sufficient but not a necessary condition.

How to Graph the Smile

When the volatility smile is not flat, its pattern obviously depends on whether implied volatility $\sigma_h^{\text{imp}}(x_t)$ is plotted against strike K , moneyness (K/S_t) , forward moneyness $(KB(t, T)/S_t) = \exp(x_t)$, log-forward moneyness x_t , and so on. The optimal choice of variable depends on what kind of information is expected to be revealed immediately on plotting implied volatilities. The common terminology *volatility smile* suggests that initially it was thought of as a kind of U-shaped pattern, whereas more recent words such as *smirk* or *frown* suggest that the focus is set on frequently observed asymmetries with respect to

a symmetric benchmark. Even more explicitly, since the volatility smile is supposed to reveal the underlying pricing probability measure $Q_{t,T}$, a common belief is that asymmetries observed in the volatility smile reveal a corresponding skewness in the distribution of (log) return under $Q_{t,T}$. Note, in particular, that, as mentioned above, a flat volatility smile at level σ characterizes a normal distribution with mean $(-\sigma^2/2)$ and variance σ^2 .

Beyond the flat case, the common belief of a close connection between smile asymmetries and risk-neutral skewness requires further qualification. First, the choice of variable must of course matter for discussion of smile asymmetries. We argue later that log-forward moneyness x_t is the right choice, that is, the fact that the smile asymmetry issue must be understood as a violation of the identity:

$$\sigma_h^{\text{imp}}(x_t) = \sigma_h^{\text{imp}}(-x_t) \quad (14)$$

This identity is actually necessary and sufficient to deduce from equation (10) that the general option pricing formula (9) fulfills the same kind of symmetry property as the BS one:

$$C_t(x) = 1 - \exp(x) + \exp(x)C_t(-x) \quad (15)$$

Although equation (15) is automatically satisfied when $C_t(x) = BS_h[x, \sigma]$ (by the symmetry property of the normal distribution: $N(-d) = 1 - N(d)$), it characterizes the symmetry property of the forward measure that corresponds to volatility smile symmetry. It actually mimics the symmetry property of the normal distribution with mean $(-\sigma^2/2)$ and variance σ^2 , which would prevail in the case of validity of the BS model. By differentiation of equation (15) and comparison with equation (8), the following can be easily checked:

The volatility smile is symmetric in the sense of equation (14) if and only if, when $f_{t,T}$ stands for the probability density function of the log-return $r_S(t, T)$ under the forward measure $Q_{t,T}$, $\exp(x/2) f_{t,T}(x)$ is an even function of x .

In conclusion, the relevant concept of symmetry amounts to considering pairs of moneynesses that are symmetric to each other in the following sense:

$$x_{1t} = \log \left[\frac{K_1 B(t, T)}{S_t} \right] = -x_{2t} = \log \left[\frac{S_t}{K_2 B(t, T)} \right] \quad (16)$$

In other words, the geometric mean of the two strike prices coincides with the current stock price:

$$\sqrt{K_1 B(t, T)} \sqrt{K_2 B(t, T)} = S_t \quad (17)$$

To conclude, it is worth noting that graphing the smile as a function of the log-moneyness x_t is even more relevant when one maintains the natural assumption that option prices are homogeneous functions of degree one with respect to the pair (S_t, K) . Merton [39] advocated this homogeneity property to preclude any “perverse local concavity” of the option price with respect to the stock price. It is obvious from equation (9) that a sufficient condition for homogeneity is that, as in the BS case, the pricing probability distribution $Q_{t,T}$ does not depend on the level S_t of the stock price. This is the reason, as discussed by Garcia and Renault [30], homogeneity holds with standard stochastic volatility option pricing models and does not hold for GARCH option pricing.

For our purpose, the big advantage of the homogeneity assumption is that it allows to compare volatility smiles (for a given time to maturity) at different dates since then the implied volatility $\sigma_h^{\text{imp}}(x_t)$ depends only on moneyness x_t and not directly on the level S_t of the underlying stock price. Moreover, from the Euler characterization of homogeneity,

$$C_t = S_t \frac{\partial C_t}{\partial S_t} + K \frac{\partial C_t}{\partial K} \quad (18)$$

we deduce (by comparing equations 2 and 5) that

$$\Delta_{1t} = \frac{\partial C_t}{\partial S_t} \quad (19)$$

is the standard delta-hedging ratio. Note that a common practice is to compute a proxy of Δ_{1t} by plugging $\sigma_h^{\text{imp}}(x_t)$ in the BS delta ratio. Unfortunately, according to some probability distribution of the volatility parameter, this approximation suffers from a convexity bias when the correct option price is a mixture of BS prices (see below). Renault [42] showed that the BS delta ratio (computed with $\sigma_h^{\text{imp}}(x_t)$) underestimates (respectively, overestimates) the correct ratio Δ_{1t} when the option is in the money (respectively, out of the money), that is, when $x_t < 0$ (respectively, $x_t > 0$).

Implied Risk Aversion

Under suitable assumptions for preferences and endowment shocks, it is well known that market completeness allows us to introduce a representative investor with a utility function U . Assuming that he/she can consume at date t and at the fixed future date T and that he/she receives one share of the stock as endowment at date t , the representative investor adjusts the dollar amount invested in the stock at each intermediary date in order to maximize the expected utility of his/her terminal consumption at time T (*see Expected Utility Maximization*). In equilibrium, the investor optimally invests all his/her wealth in the risky stock and then consumes the terminal value S_T of the stock. Thus, the Euler first-order condition for optimality imposes that the price π_t at time t of any contingent claim that delivers the dollar amount g_T at time t is such that

$$\pi_t = E^P \left[\beta^{T-t} \frac{U'(S_T)}{U'(S_t)} g_T \mid \mathcal{F}_t \right] \quad (20)$$

where $E^P [\cdot \mid \mathcal{F}_t]$ denotes the conditional expectation given $\mathcal{F}_t$ with respect to the historical probability measure P . β is the subjective discount factor. By identifying equation (20) with equation (1), we deduce that the density function given $\mathcal{F}_t$ of $Q_{t,T}$ with respect to P is proportional to $U'(S_T)$. Thus, if f_t^* (respectively, f_t) stands for the density function given $\mathcal{F}_t$ of the distribution function of S_T under the martingale measure Q (respectively under the historical measure P), we have (f_t^*/f_t) proportional to the marginal utility function and thus, by logarithmic derivatives,

$$\rho_t(S_T) = -S_T \frac{U''(S_T)}{U'(S_T)} = S_T \left(\frac{f_t'(S_T)}{f_t(S_T)} - \frac{f_t^{*'}(S_T)}{f_t^*(S_T)} \right) \quad (21)$$

where $\rho_t(S_T)$ is, by definition, the Arrow–Pratt measure of relative **risk aversion**. Note that the result would not have changed if we had used the forward probability measure $Q_{t,T}$ (*see Forward and Swap Measures*) instead of the risk-neutral one Q . Since as explained in the first section, the observation at time t of a cross section of option prices written on S_T allows us to infer these probability measures, it suggests that risk-aversion functions or preference parameters may be extracted from observed option

prices. Aït-Sahalia and Lo [4] and Jackwerth [35] proposed nonparametric approaches to recover risk-aversion functions across wealth states from observed stock and option prices. Bliss and Panigirtzoglou [10], Garcia *et al.* [29], and Rosenberg and Engle [45] have estimated preference parameters based on parametric asset pricing models with several specifications of the utility function.

These efforts to exploit prices of options to recover fundamental economic parameters have produced puzzling results. Aït-Sahalia and Lo [4] found that the nonparametrically implied function of relative risk aversion varies significantly across the range of S&P 500 index values, from 1 to 60, and is U-shaped. Jackwerth [35] also found that the implied absolute risk-aversion function is U-shaped around the current forward price and also that it can become negative. Parametric empirical estimates of the coefficient of relative risk aversion also show considerable variation. Rosenberg and Engle [45] reported values ranging from 2.36 to 12.55 for a power utility pricing kernel across time, while Bliss and Panigirtzoglou [10] estimated average values between 2.33 and 11.14 for the same S&P 500 index for several option maturities. Garcia *et al.* [29] estimated a consumption-based asset pricing model with regime-switching fundamentals and Epstein and Zin [23], the preferences. The estimated parameters for risk aversion and intertemporal substitution are reasonable, with average values of 0.6838 and 0.8532, respectively, over the period 1991–1995.

The general conclusion is that empirical risk-neutral distributions that are computed without a precise account of the state variables that enter into the investors' information sets cannot provide valuable insights on intertemporal preferences. For example, Chabi-Yo *et al.* [17] showed that in an economy with regime changes either in fundamentals or in preferences, an application of the nonparametric methodology used by Jackwerth [35] to recover risk aversion will lead to similar negative estimates of the risk-aversion function in some states of wealth even though the risk-aversion functions are consistent with economic theory within each regime.

Implied Volatility as a Calibrated Parameter

Formula (20) puts forward the so-called pricing kernel $m_{t,T} = \beta^{T-t} \frac{U'(S_T)}{U'(S_t)}$ that emphasizes the role of

risk aversion. In particular, in the case of constant relative risk aversion ($\rho_t(S_T) = -S_T \frac{U''(S_T)}{U'(S_T)} = a > 0$), the log-pricing kernel is perfectly correlated to the log-return on the stock:

$$\text{Log}(m_{t,T}) = -a \text{Log} \left[\frac{S_T}{S_t} \right] + (T - t) \text{Log}(\beta) \quad (22)$$

More generally, it is convenient to rewrite the call-pricing equation (1) in terms of the pricing kernel:

$$C_t = E^P [m_{t,T} \text{Max}[0, S_T - K] | \mathcal{F}_t] \quad (23)$$

where the general definition of the pricing kernel (*see also Pricing Kernels*) gives

$$\frac{dQ_{t,T}}{dP} = \frac{m_{t,T}}{B(t, T)} \quad (24)$$

It is easy to check that the general call-pricing formula (23) collapses into the BS one when the two following conditions are fulfilled:

1. The conditional distribution given $\mathcal{F}_t$ of the log-return $\text{Log} \left[\frac{S_T}{S_t} \right]$ is normal with constant variance σ .
2. The log-pricing kernel $\text{Log}(m_{t,T})$ is perfectly correlated to the log-return on the stock as in equation (22).

We get the first interesting generalization of the BS formula by now considering that the log-return $\text{Log} \left[\frac{S_T}{S_t} \right]$ and the log-pricing kernel $\text{Log}(m_{t,T})$ may be jointly normally distributed given $\mathcal{F}_t$, with conditional moments possibly depending on the conditional information at time t . Interestingly enough, it can be shown that the call price computed from formula (23) with this joint conditional lognormal distribution will explicitly depend on the conditional moments only through the conditional stock volatility:

$$(T - t) \sigma_{t,T}^2 = \text{Var} \left[\text{Log} \left[\frac{S_T}{S_t} \right] | \mathcal{F}_t \right] \quad (25)$$

More precisely, we get the following option pricing formula:

$$C_t = S_t B_{S_{T-t}}[x_t, \sigma_{t,T-t}] \quad (26)$$

Formula (26) is actually the generalization of the “risk-neutral valuation relationship” (RNVR) put forward by Brennan [12]. With joint log-normality

of return and pricing kernel, we find again the BS functional form due to the Cameron–Martin formula (a static version of Girsanov’s theorem), which tells us that when X and Y are jointly normal,

$$E \{ \exp(X) g(Y) \} = E[\exp(X)] E \{ g[Y + \text{cov}(X, Y)] \} \quad (27)$$

While the term $E[\exp(X)]$ will give $B(t, T)$ (with $X = \text{Log}(m_{t,T})$), the term $\text{cov}(X, Y)$ (with $Y = \text{Log} \left[\frac{S_T}{S_t} \right]$) will make the risk-neutralization since $E \{ \exp(X) \exp(Y) \} = E[\exp(X)] E \{ \exp[Y + \text{cov}(X, Y)] \}$ must be 1.

From an econometric viewpoint, the interest of equation (26), when compared to equation (10), is to deliver a flat volatility smile but with an implied volatility level that may be time varying and correspond to the conditional variance of the conditionally lognormal stock return. In other words, the time-varying volatility of the stock becomes observable as calibrated from option prices:

$$\sigma_{t,T} = \sigma_{T-t}^{\text{imp}}(x_t) \quad \forall x_t \quad (28)$$

The weakness of this approach is its lack of robustness with respect to temporal aggregation. In the GARCH-type literature, stock returns may be conditionally lognormal when they are considered on the elementary period of the discrete-time setting ($T = t + 1$) while implied time-aggregated dynamics are more complicated. This is the reason the GARCH option pricing literature [21, 33]) maintains formula (26) only for $T = t + 1$. Nonflat volatility smiles may be observed with longer times to maturity. Kallsen and Taqqu [37] provided a continuous-time interpretation of such GARCH option pricing.

Implied Volatility as an Expected Average Volatility

To account for excess kurtosis and skewness in stock log-returns, a fast empirical approach amounts to considering that the option price at time t is given by a weighted average:

$$\alpha_t S_t B_{S_h}[x_t, \sigma_{1t}] + (1 - \alpha_t) S_t B_{S_h}[x_t, \sigma_{2t}] \quad (29)$$

The rationale for equation (29) is to consider that a mixture of two normal distributions with standard

errors σ_{1t} and σ_{2t} and weights α_t and $(1 - \alpha_t)$, respectively, may account for both skewness and excess kurtosis in stock log-return. The problem with this naive approach is that it does not take into account any risk premium associated with the mixture component. More precisely, if we want to accommodate a mixture of normal distributions with a mixing variable $U_{t,T}$, we can rewrite equation (23) as

$$C_t = E^P \left\{ E^P [m_{t,T} \text{Max}[0, S_T - K] \mid \mathcal{F}_t, U_{t,T}] \mid \mathcal{F}_t \right\} \quad (30)$$

where, for each possible value $u_{t,T}$ of $U_{t,T}$, a BS formula such as equation (26) is valid for computing

$$E^P [m_{t,T} \text{Max}[0, S_T - K] \mid \mathcal{F}_t, U_{t,T} = u_{t,T}] \quad (31)$$

In other words, it is true that, as in equation (29), the conditional expectation operator (given $\mathcal{F}_t$) in equation (30) displays the option price as a weighted average of different BS prices with the weights corresponding to the probabilities of the possible values $u_{t,T}$ of the mixing variable $U_{t,T}$. However, the naive approach (29) is applied incorrectly if we do not take into account the fact that the additional conditioning information $U_{t,T}$ should lead to modify some key inputs in the BS option pricing formula. Suppose that investors are told that the mixing variable $U_{t,T}$ will take the value $u_{t,T}$, then the current stock price would no longer be

$$S_t = E^P [m_{t,T} S_T \mid \mathcal{F}_t] \quad (32)$$

but

$$S_t^*(u_{t,T}) = E^P [m_{t,T} S_T \mid \mathcal{F}_t, U_{t,T} = u_{t,T}] \quad (33)$$

For the same reason, the pure discount bond that delivers \$1 at time T will no longer be priced at time t as

$$B(t, T) = E^P [m_{t,T} \mid \mathcal{F}_t] \quad (34)$$

but as

$$B^*(t, T)(u_{t,T}) = E^P [m_{t,T} \mid \mathcal{F}_t, U_{t,T} = u_{t,T}] \quad (35)$$

In other words, various BS option prices that are averaged in a mixture approach such as equation (29) must be computed, not with actual values $B(t, T)$ and S_t of the current bond and stock prices, but with values $B^*(t, T)(u_{t,T})$ and $S_t^*(u_{t,T})$ not directly

observed but computed from equations (35) and (33). In particular, the key inputs, underlying stock price and interest rate, should be different in various applications of the BS formulas like $BS_h[x, \sigma_1]$ and $BS_h[x, \sigma_2]$ in equation (29). This remark is crucial for the conditional Monte Carlo approach, as developed, for instance, by Willard [47] in the context of option pricing with stochastic volatility. Revisiting a formula initially derived by Romano and Touzi [44], Willard [47] noted that the variance reduction technique, known as *conditional Monte Carlo*, can be applied even when the conditioning factor (the stochastic volatility process) is instantaneously correlated with the stock return as it is the case when leverage effect is present. He stressed that “by conditioning on the entire path of the noise element in the volatility (instead of just the average volatility), we can still write the option’s price as an expectation over Black–Scholes prices by appropriately adjusting the arguments to the Black–Scholes formula”. Willard’s [47] “appropriate adjustment” of the stock price is actually akin to equation (33). Moreover, he does not explicitly adjust the interest rate according to equation (35) and works with a fixed risk-neutral distribution. More generally, the continuous-time dynamics of a pricing kernel $m_{t,T}$ is considered, which is seen as the relative increment of a pricing kernel process M_t :

$$m_{t,T} = \frac{M_T}{M_t} \quad (36)$$

The key idea of the mixture model is then to define a conditioning variable $U_{t,T}$ such that the pricing kernel process and the stock price process jointly follow a bivariate geometric Brownian motion under the conditional probability distribution given $U_{t,T}$. The mixing variable $U_{t,T}$ will typically show up as a function of a state variable path $(X_\tau)_{t \leq \tau \leq T}$. More precisely, we specify the jump-diffusion model:

$$\begin{aligned} d(\log S_t) &= \mu(X_t) dt + \alpha(X_t) dW_{1t} \\ &\quad + \beta(X_t) dW_{2t} + \gamma_t dN_t \end{aligned} \quad (37)$$

$$\begin{aligned} d(\log M_t) &= h(X_t) dt + a(X_t) dW_{1t} \\ &\quad + b(X_t) dW_{2t} + c_t dN_t \end{aligned} \quad (38)$$

where (W_{1t}, W_{2t}) is a two-dimensional standard Brownian motion, N_t is a Poisson process with intensity $\lambda(X_t)$ depending on the state variable X_t . Given

the jump dates, the jump amplitudes at these dates are independent draws (γ_t, c_t) in a fixed bivariate normal distribution. These draws are also independent of the state variable process X . The Brownian motion W_1 is assumed to be part of the state variable vector X to capture the possible instantaneous correlation between ex-jump volatility of the stock (as measured by $V_t = \alpha^2(X_t) + \beta^2(X_t)$) and its Brownian innovation. More precisely, the correlation coefficient $\rho(X_t) = \frac{\alpha(X_t)}{\sqrt{V_t}}$ measures the so-called leverage effect.

The jump-diffusion model (37 and 38) is devised such that, given the state variables path $(X_\tau)_{t \leq \tau \leq T}$ as well as the number $(N_T - N_t)$ of jumps between times t and T , the joint normality of $(\log S_T, \log m_{t,T})$ is maintained. This allows us to derive a generalized Black and Scholes (GBS) option pricing formula by application of equations (30) and (26):

$$C_t = S_t E^P[\xi_{t,T} B S_{T-t}(x_t^*, \sigma_{t,T}) | \mathcal{F}_t] \quad (39)$$

where

$$\sigma_{t,T}^2 = \int_t^T [1 - \rho^2(X_\tau)] V_\tau d\tau + (N_T - N_t) \text{Var}(\gamma_t) \quad (40)$$

and

$$x_t^* = \log \left[\frac{K B^*(t, T)}{S_t \xi_{t,T}} \right] \quad (41)$$

where $S_t \xi_{t,T}$ and $B^*(t, T)$ correspond to $S_t^*(u_{t,T})$ and $B^*(t, T)(u_{t,T})$, respectively, which are defined in equations (33) and (35). General computations of these quantities in the context of a jump-diffusion model can be found in [27] and [48]. Let us exemplify these formulas when there is no jump. Then, we can define a short-term interest rate as

$$r(X_t) = -h(X_t) - \frac{1}{2}[a^2(X_t) + b^2(X_t)] \quad (42)$$

and then

$$\begin{aligned} B^*(t, T) = & \exp \left[- \int_t^T r(X_\tau) d\tau \right] \\ & \times \exp \left[\int_t^T a(X_\tau) dW_{1\tau} - \frac{1}{2} \int_t^T a^2(X_\tau) d\tau \right] \end{aligned} \quad (43)$$

and

$$\begin{aligned} \xi_{t,T} = & \exp \left[\int_t^T [a(X_\tau) + \alpha(X_\tau)] dW_{1\tau} \right. \\ & \left. - \frac{1}{2} \int_t^T [a(X_\tau) + \alpha(X_\tau)]^2 d\tau \right] \end{aligned} \quad (44)$$

In particular, it can be easily checked that

$$B(t, T) = E^P[B^*(t, T) | \mathcal{F}_t] \quad (45)$$

and

$$S_t = E^P[S_t \xi_{t,T} | \mathcal{F}_t] \quad (46)$$

For now, the difference between $S_t \xi_{t,T}$ and $B^*(t, T)$ and their respective expectations S_t and $B(t, T)$ are neglected. It is then clear that the GBS formula warrants the interpretation of the BS-implied volatility $\sigma_{T-t}^{\text{imp}}(x_t)$ as approximatively an expected average volatility. Neglecting convexity effects (non-linearity of the BS formula with respect to volatility), the GBS formula would actually give

$$[\sigma_{T-t}^{\text{imp}}(x_t)]^2 = E^P[\sigma_{t,T}^2 | \mathcal{F}_t] \quad (47)$$

The likely impact of the difference between $S_t \xi_{t,T}$ and $B^*(t, T)$ and their respective expectations S_t and $B(t, T)$ is twofold. First, a nonzero function $a(X_\tau)$ must be understood as a risk premium on the volatility risk. In other words, the above interpretation of $\sigma_{T-t}^{\text{imp}}(x_t)$ as approximatively an expected average volatility can be maintained by using risk-neutral expectations. Considering the BS-implied volatility as a predictor of volatility over the lifetime of the option is tantamount to neglecting the volatility risk premium. Beyond this risk premium effect, the leverage effect $\rho(X_t)$ will distort this interpretation through its joint impact on $\sigma_{t,T}^2$ and also on $\xi_{t,T}$ (through $\alpha(X_t) = \rho(X_t)\sqrt{V_t}$). Although Renault and Touzi [43] have shown that we will get a symmetric volatility smile in the case of zero-leverage, Renault [42] explained that, with nonzero leverage, the implied distortion of the stock price by the factor $\xi_{t,T}$ will produce asymmetric volatility smirks. More precisely, Yoon [48] characterized the cumulative impact of the two effects of leverage and showed that they compensate each other almost exactly for at the money options. Finally, Comte and Renault's [18] long-memory volatility model explains that, in spite of the time averaging in equations (40) and (47), the

volatility smile does not become flat even for long-term options.

The close connection (equations 40 and 47) between BS-implied volatility and the underlying volatility process ($\sqrt{V_t}$) has inspired a strand of literature on estimating volatility dynamics from option prices data. Pastorello *et al.* [41] consider directly $[\sigma_{T-t}^{\text{imp}}(x_t)]^2$ as a proxy for squared spot volatility V_t and correct the resulting approximation bias in estimating volatility dynamics by indirect inference. The “implied-states approach” (see [40] and references therein) more efficiently uses the exact relationship between $\sigma_{T-t}^{\text{imp}}(x_t)$ and V_t , as given by equations (40) and (47) for a given spot volatility model, to estimate the volatility parameters by maximum likelihood or generalized method of moments. Finally, Garcia *et al.* [28] simplified the procedure by taking advantage of high-frequency intraday return data for a direct measurement of the integrated volatility $\int_t^T V_\tau d\tau$.

Data-driven Approaches

In this section, we discuss methods that are, to various degrees, data-driven approaches, ranging from purely data-driven, sometimes called *nonparametric approaches* to semiparametric ones—where parametric and nonparametric estimators are combined. The estimation of option-implied densities can take various forms as there are many different ways to smooth a curve. Nadaraya–Watson kernel regression was used by Aït-Sahalia and Lo [3, 13, 14, 19], constrained splines by Bates [9], flexible parametric functional forms by Abadir and Rockinger [1], and neural networks by Aït-Sahalia and Duarte [2]; Bondarenko [11] and Garcia and Gençay [26], considered constrained least-square approaches. A more practitioner-oriented approach was pioneered by Jackwerth and Rubinstein [36] and Rubinstein [46] who constructed implied trees from listed option prices.

All these methods are data intensive and typically feature slow convergence rates. Only a few indices have enough liquid actively traded derivatives with a wide span of maturities and moneyness to warrant applications of these methods.

The extended method of moments (XMMs) estimator, introduced by Gagliardini *et al.* [31], extends the standard GMM to accommodate a more general set of moment restrictions. The standard GMM is based on uniform conditional moment restrictions,

such as equation (20), which are valid for any value of the conditioning variables. The XMM can handle not only the uniform moment restrictions but also the local moment restrictions that are valid only for a given value of the conditioning variables. This leads to a new field of application to derivative pricing, as the XMM can be used for reconstructing the pricing operator on a given day, by using the information in a cross section of observed traded derivative prices and a time series of underlying asset returns. To illustrate the principle of XMM, the S&P 500 index and its derivatives are considered. Suppose that an investor at date t_0 is interested in estimating the price $c_{t_0}(h, k)$ of a call option with time-to-maturity h and moneyness strike k that is currently not (actively) traded on the market. He/she has data on a time series of T daily returns of the S&P 500 index, as well as on a small cross section of current option prices $c_{t_0}(h_j, k_j)$, $j = 1, \dots, n$, of n highly traded derivatives. The XMM approach provides the estimated prices $\hat{c}_{t_0}(h, k)$ for different values of moneyness strike k and time-to-maturity h , which interpolate the observed prices of highly traded derivatives and satisfy the hypothesis of the absence of arbitrage opportunities. These estimated prices are consistent for a large number of dates T , but with a fixed, even small, number of observed derivative prices n .

A slightly less ambitious approach has been recently advocated by Eriksson *et al.* [24]. They suggest the use of the normal inverse Gaussian (NIG) family (see **Normal Inverse Gaussian Model**) to approximate an unknown distribution risk-neutral density. The appeal of the NIG family of distributions is that they are characterized by the first four moments: mean, variance, skewness, and kurtosis. These are the moments we care about in many applications, including derivative pricing. The unknown density function is approximated by matching the cumulants. The latter are obtained from the cross section of option prices using methods proposed by Bakshi *et al.* [5]. The strength of their approach is that they link the pricing of individual derivatives to the moments of the risk-neutral distribution, which has an intuitive appeal in terms of how volatility, skewness, and kurtosis of the risk-neutral distribution can explain the behavior of the derivative prices. Eriksson *et al.* [24] showed that the approximation errors are minor when compared to several option pricing models that have known densities.

Non-Gaussian Option Pricing

The affine jump-diffusion framework (*see Affine Models*) represents an important theoretical advance. Yet, it has been argued that it has its limitations, notably due to the exclusive use of compound Poisson processes to model jumps. These processes generate a finite number of jumps within a finite time interval and have accordingly been referred to as *finite activity jump processes*. The observation that asset prices actually display many small jumps on a fine timescale has led to the development of more general jump structures, which permit an infinite number of jumps to occur within any bounded time interval. Examples of infinite-activity jump models include the inverse Gaussian model of Barndorff-Nielsen [6, 7], the generalized hyperbolic class of Eberlein *et al.* [22], the variance gamma (VG) model (*see Variance-gamma Model*) of Madan and Milne [38], the generalization of VG by Carr *et al.* [15], and the finite moment log-stable model of Carr and Wu [16]. Empirical work by these authors is generally supportive of the use of infinite-activity processes as a way to model returns in a parsimonious way. The recognition that volatility is stochastic has led to further extensions of infinite-activity Lévy models (*see Barndorff-Nielsen and Shephard (BNS) Models; Time-changed Lévy Process*) by Barndorff-Nielsen and Shephard [8] and Carr and Wu [16]. However, these models often assume that changes in volatility are independent of asset returns, and consider the leverage effect only under special cases. Carr and Wu [16] used time-changed Lévy processes that generalize the affine Poisson jump diffusions by relaxing the affine structure and by allowing more general specifications of the jump structure. Since the pioneering work of Heston [32], the literature has used the characteristic function for deriving option prices. Accordingly, Carr and Wu focus on developing analytic expressions for the characteristic function of a time-changed Lévy process (*see Time-changed Lévy Process*).

References

- [1] Abadir, K.M. & Rockinger, M. (1998). *Density-embedding Functions*, University of York and Groupe HEC. Working Paper.
- [2] Aït-Sahalia, Y. & Duarte, J. (2003). Nonparametric option pricing under shape restrictions, *Journal of Econometrics* **116**, 9–47.
- [3] Aït-Sahalia, Y. & Lo, A.W. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices, *Journal of Finance* **53**, 499–547.
- [4] Ait-Sahalia Y. & Lo A. (2000). Nonparametric risk management and implied risk aversion, *Journal of Econometrics* **94**, 9–51.
- [5] Bakshi G., Kapadia N. & Madan D. (2003). Stock return characteristics, Skew laws, and differential pricing of individual equity options, *Review of Financial Studies* **16**, 101–143.
- [6] Barndorff-Nielsen, O.E. (1998). Processes of normal inverse Gaussian type, *Finance and Stochastics* **2**, 41–68.
- [7] Barndorff-Nielsen, O.E. (2001). Superposition of Ornstein-Uhlenbeck type processes, *Theory of Probability and Its Applications* **45**(2), 175–194.
- [8] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics (with discussion), *Journal of the Royal Statistical Society, Series B* **63**, 167–241.
- [9] Bates, D. (2000). Post-'87 crash fears in the S&P 500 futures option, *Journal of Econometrics* **94**, 181–238.
- [10] Bliss, R.R. & Panigirtzoglou, N. (2004). Option-implied risk aversion estimates, *Journal of Finance* **59**, 407–446.
- [11] Bondarenko, O. (2003). Estimation of risk-neutral densities using positive convolution approximation, *Journal of Econometrics* **116**, 85–112.
- [12] Brennan, M.J. (1979). The pricing of contingent claims in discrete-time models, *Journal of Finance* **34**, 53–68.
- [13] Broadie, M., Detemple, J., Ghysels, E. & Torrès, O. (2000). American options with stochastic dividends and volatility: a nonparametric investigation, *Journal of Econometrics* **94**, 53–92.
- [14] Broadie, M., Detemple, J., Ghysels, E. & Torrès, O. (2000). Nonparametric estimation of American options exercise boundaries and call prices, *Journal of Economic Dynamics and Control* **24**, 1829–1857.
- [15] Carr, P., Geman, H., Madan, D.B. & Yor, M. (2003). Stochastic volatility for Lévy processes, *Mathematical Finance* **13**, 345–382.
- [16] Carr, P. & Wu, L. (2004). Time-changed Lévy processes and option pricing, *Journal of Financial Economics* **71**, 113–141.
- [17] Chabi-Yo, F., Garcia, R. & Renault, E. (2008). State dependence can explain the risk aversion puzzle, *Review of Financial Studies* **21**, 973–1011.
- [18] Comte, F. & Renault E. (1998). Long-memory in continuous time stochastic volatility models, *Mathematical Finance* **8**, 291–323; Reprinted in Shephard, N. (ed.) (2005). *Stochastic Volatility: Selected Readings*, Oxford University Press.

- [19] Cont, R. & Da Fonseca, J. (2002). Dynamics of implied volatility surfaces, *Quantitative Finance* **2**(1), 45–60.
- [20] Cox, J.C., Ross, S.A. & Rubinstein, M. (1979). Option pricing: a simplified approach, *Journal of Financial Economics* **7**, 229–263.
- [21] Duan, J.C. (1995). The GARCH option pricing model, *Mathematical Finance* **5**, 13–32.
- [22] Eberlein, E., Keller, U. & Prause, K. (1998). New insights into smile, mispricing, and Value at Risk: the hyperbolic model, *Journal of Business* **71**, 371–405.
- [23] Epstein, L. & Zin, S. (1989). Substitution, risk aversion, and the temporal behavior of consumption and asset returns: a theoretical framework, *Econometrica* **57**, 937–969.
- [24] Eriksson, A., Ghysels, E. & Wang, F. (2009). The normal inverse Gaussian distribution and the pricing of derivatives, *Journal of Derivatives* **16**, 23–38.
- [25] Föllmer, H. & Schied, A.A. (2004). *Stochastic Finance: An Introduction in Discrete Time*, Walter de Gruyter.
- [26] Garcia, R. & Gençay, R. (2000). Pricing and hedging derivative securities with neural networks and a homogeneity hint, *Journal of Econometrics* **94**, 93–115.
- [27] Garcia, R., Ghysels, E. & Renault, E. (2009). Econometrics of option pricing models, in *Handbook of Financial Econometrics*, Y. Ait-Sahalia & L.P. Hansen, eds, North Holland, forthcoming.
- [28] Garcia, R., Lewis, M., Pastorello, S. & Renault, E. (2009). Estimation of objective and risk-neutral distributions based on moments of integrated volatility, *Journal of Econometrics*, forthcoming.
- [29] Garcia, R., Luger, R. & Renault, E. (2003). Empirical assessment of an intertemporal option pricing model with latent variables, *Journal of Econometrics* **116**, 49–83.
- [30] Garcia, R. & Renault, E. (1998). A note on GARCH option pricing and hedging, *Mathematical Finance* **8**(2), 153–161.
- [31] Gagliardini, P., Gouriéroux, C. & Renault, E. (2009). *Efficient Derivative Pricing by Extended Method of Moments*, Discussion Paper.
- [32] Heston, S. (1993). A closed form solution for options with stochastic volatility with applications to bond and currency options, *The Review of Financial Studies* **6**, 327–343.
- [33] Heston, S. & Nandi, S. (2000). A closed-form GARCH option valuation model, *Review of Financial Studies* **13**, 585–625.
- [34] Huang, C. & Litzenberger, R.H. (1998). *Foundations for Financial Economics*, North Holland.
- [35] Jackwerth, J. (2000). Recovering risk aversion from option prices and realized returns, *Review of Financial Studies* **13**, 433–451.
- [36] Jackwerth, J. & Rubinstein, M. (1996). Recovering probabilities distributions from option prices, *Journal of Finance* **51**, 1611–1631.
- [37] Kallsen, J. & Taqqu, M.S. (1998). Option pricing in ARCH-type models, *Mathematical Finance* **8**, 13–26.
- [38] Madan, D. & Milne, F. (1991). Option pricing with VG martingale components, *Mathematical Finance* **1**, 39–56.
- [39] Merton, R.C. (1973). Theory of rational option pricing, *Bell Journal of Economics and Management Science* **4**, 141–183.
- [40] Pastorello, S., Patilea, V. & Renault, E. (2003). Iterative and recursive estimation in structural non-adaptive models - invited lecture with discussion, *Journal of Business, Economics and Statistics* **21**, 449–509.
- [41] Pastorello, S., Renault, E. & Touzi, N. (2000). Statistical inference for random-variance option pricing, *Journal of Business and Economic Statistics* **18**, 358–367.
- [42] Renault, E. (1997). Econometric models of option pricing errors, in *Advances in Economics and Econometrics: Theory and Applications*, D. Kreps & K. Wallis, eds, Cambridge University Press, Vol. 3.
- [43] Renault, E. & Touzi, N. (1996). Option hedging and implied volatilities in a stochastic volatility model, *Mathematical Finance* **6**, 279–302.
- [44] Romano M. & Touzi N. (1997). Contingent claims an market completeness in a stochastic volatility model, *Mathematical Finance* **7**, 399–410.
- [45] Rosenberg, J.V. & Engle, R.F. (2002). Empirical pricing Kernels, *Journal of Financial Economics* **64**, 341–372.
- [46] Rubinstein, M. (1994). Implied binomial trees, *Journal of Finance* **44**, 771–818.
- [47] Willard, G.A. (1997). Calculating prices and sensitivities for path-independent derivative securities in multifactor models, *Journal of Derivatives* **5**, 45–61.
- [48] Yoon, J. (2008). *Option Pricing with Stochastic Volatility Models*. PhD Thesis, UNC at Chapel Hill.

Further Reading

- Bakshi, G. & Madan, D. (2000). Spanning and derivative-security valuation, *Journal of Financial Economics* **55**, 205–238.
- Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–659.
- Garcia, R., Luger, R. & Renault, E. (2005). Viewpoint: option prices, preferences and state variables, *Canadian Journal of Economics* **38**, 1–27.
- Ghysels, E. & Wang, F. (2009). *Some Useful Densities for Risk Management and their Properties*, Discussion Paper, UNC.

Related Articles

Affine Models; Arrow–Debreu Prices; Black–Scholes Formula; Equivalent Martingale Measures; Implied Volatility in Stochastic Volatility

Models; Implied Volatility Surface; Jump-diffusion Models; Jump Processes; Model Calibration; Option Pricing: General Principles; Pricing Kernels; Risk-neutral Pricing; Saddlepoint Approximation; Stochastic Volatility Models.

ERIC RENAULT

Long Range Dependence

From Long to Short-range Dependency in Discrete-time Models

The widely known class of ARMA (autoregressive moving average) processes used for modeling discrete-time processes (see [5, 6]) gives a first approach to short- and long-range dependence. For instance, the model that explains the value of a variable today, say X_t , mainly via its value yesterday, X_{t-1} , up to a centered error ε_t , leads under mild conditions to a zero mean stationary dependent process $(X_t)_{t \in \mathbb{Z}}$,

$$X_t = bX_{t-1} + \varepsilon_t \quad (1)$$

for an independent, identically distributed (i.i.d) noise sequence ε and $|b| < 1$. Such a process is, short-range dependent; this means that the process quickly forgets all about its past. A more rigorous criterion is that, under stationary assumption, the correlation function $\rho_X(h) = \text{corr}(X_t, X_{t+h})$ is exponentially decreasing, with

$$|\rho_X(h)| \sim_{h \rightarrow +\infty} e^{-\vartheta h} \quad (2)$$

for positive constant ϑ , a property satisfied by equation (1) with $\vartheta = -\log(|b|)$.

Now consider the lag operator L , $LX_t = X_{t-1}$ and rewrite model (1) as follows:

$$(1 - bL)X_t = \varepsilon_t \quad (3)$$

For the above behavior to occur, it is required that $|b| < 1$. If $|b| = 1$, the process is no longer stationary, and, for example, for $b = 1$ the so-called unit root problem appears. An intermediate case of high interest is that of a fractional unit root. More precisely, set

$$(1 - L)^\alpha X_t = \varepsilon_t \quad (4)$$

where $|\alpha| < 1/2$ is not an integer. The operator $(1 - L)^\alpha$ is then defined (see [14, 15, 22]) by its series expansion

$$(1 - L)^\alpha = 1 - \sum_{k=1}^{\infty} \frac{\alpha \Gamma(k - \alpha)}{\Gamma(k + 1) \Gamma(1 - \alpha)} L^k$$

where, for positive x , $\Gamma(x) = \int_0^{+\infty} u^{x-1} e^{-u} du$ and $L^k X_t = X_{t-k}$. In that case, we recover a stationary

process (X_t) , but for large values of h ,

$$|\rho_X(h)| \sim_{h \rightarrow +\infty} Ch^{2\alpha-1} \quad (5)$$

This means that the covariance function (which is a measure of dependence between the observations) decays slowly. This can also be characterized in the frequency domain by the behavior near zero of the discrete spectral density defined by

$$f_X(\lambda) = \sigma_0^2 \sum_{h \in \mathbb{Z}} \rho_X(h) e^{i\lambda h} \quad (6)$$

The spectral density is such that $f_X(0)$ is finite positive and well defined in case (1) under $|b| < 1$, whereas

$$f(\lambda) \sim_{\lambda \rightarrow 0} C\lambda^{-2\alpha} \quad (7)$$

in case (4). Here, we can see that in the range $-1/2 < \alpha < 1/2$, two different cases arise:

- the case $-1/2 < \alpha < 0$, where $f_X(0) = 0$, and the sum of the correlations is finite—also called *intermediate-memory case*;
- the case $0 < \alpha < 1/2$, where $f_X(0) = +\infty$, and the sum of the correlations is infinite—also called *long-memory case*.

It is worth mentioning that many strategies for statistical inference (see [3]) on α are related to the expansion $\log f_X(\lambda) = C - 2\alpha \log \lambda + \text{Remainder}$, for λ small enough (see [12, 20]). Note also that standard log-likelihood methods were studied in this framework, which was new in the 1980s in the sense that many crucial standard assumptions required to obtain “good” properties of the estimators were violated (see [10, 11]).

Such models have been applied to the term structure of interest rates [1] and to foreign exchange rates [8]. However, it is well known that linear models fail to adequately represent asset returns, which have been frequently modeled using GARCH (generalized autoregressive conditionally heteroskedastic) (see **GARCH Models**) models [13]. Fractional extensions of GARCH models have been proposed to capture long-range dependence properties in the volatility process [2, 4]. More precisely, a GARCH model represents $X_t = \ln(S_t/S_{t-1})$ where S_t is an asset price as $X_t = \sigma_t \varepsilon_t$ and ε_t is a centered white-noise process

with variance 1. Then the recurrence equations

$$\sigma_t^2 = a_0 + \sum_{j=1}^p a_j X_{t-j}^2 + \sum_{k=1}^q b_k \sigma_{t-k}^2 \quad (8)$$

can be written as ARMA representations with respect to the white noise $v_t = X_t^2 - \sigma_t^2$. It is then natural to study the effect of fractional unit root in such equations, and the long-range dependence properties. The point in GARCH models is that the existence of a unit root in this framework leading to processes called IGARCH (Integrated GARCH) models can be compatible with stationarity and is a mean of modeling persistence in variance of shocks. Thus, fractional models in this context turn out to be less essential. The theoretical study of the resulting processes can be also difficult: it is already the case without long-range dependence.

It is therefore natural to formulate such models in continuous time.

Long-range-dependent Continuous-time Models in Finance

Indeed, model (4) for X_t has properties that are very similar to fractional Brownian motion (*see Fractional Brownian Motion*) increments, introduced by Mandelbrot and Van Ness [18]. More precisely, a natural analog of $(1-L)^{-\alpha} \varepsilon_t$ is $\int_0^t (t-s)^{\alpha} dB(s)$ where B is a standard Brownian motion. The integral should start at $-\infty$ for stationarity purpose, but it would not have finite variance. This is why the truncated version above is often used for modeling purposes. Fractional Brownian motion, whether truncated or not, does not have the semimartingale property. As this property is crucially required for log asset prices in continuous-time arbitrage-free models, fractional models led to problems in this regard. This was confirmed by Rogers [21] who showed the existence of arbitrage in presence of fractional Brownian motion for the log-prices. Moreover, there is no compelling empirical evidence of long memory in asset returns. But Hull and White-type [17] or stochastic volatility models lead to the idea that there was some kind of persistence in the volatility process. This persistence can be modeled by the insertion of fractional Brownian motion in the volatility equation.

Consider first a stochastic volatility model (*see Stochastic Volatility Models*) defined as a bivariate

diffusion process:

$$\begin{cases} dS_t/S_t = \mu(t, S_t)dt + \sigma_t dB_t^{(1)}, \\ \sigma_t^2 = f(\eta_t), \\ d\eta_t = m(\eta_t)dt + g(\eta_t)dB_t^{(2)} \end{cases} \quad (9)$$

where $(B_t^{(1)}, B_t^{(2)})$ is a bidimensional standard Brownian motion. Let us consider the case where $B_t^{(1)}$ and $B_t^{(2)}$ are independent. To introduce some long-range dependence in the volatility behavior, the Brownian motion driving the volatility equation can be replaced by a fractional Brownian motion. Stochastic differential equations driven by fractional Brownian motions require advanced tools for proving existence and uniqueness of solutions [7, 16, 19]. A few simple examples can be studied with more “elementary” tools (see [9]). For instance, take

$$\begin{cases} dS_t/S_t = \mu(t, S_t)dt + \sigma_t dB_t^{(1)}, \\ \sigma_t^2 = \exp(2\eta_t), \\ d\eta_t = k(\theta - \eta_t)dt + g dB_t^{(\alpha)}, \quad k, \theta, g \text{ constants} \end{cases} \quad (10)$$

where $B_t^{(\alpha)} = \int_0^t (t-s)^{\alpha} dB_s^{(2)} / \Gamma(1+\alpha)$. For simplicity, we shall set $\mu \equiv 0$ in the following. Initial conditions for σ_t can be chosen so that the process is stationary. Then, it can be checked that the process σ_t is long-range dependent when considered as continuously observed, in the sense given in equations (5) and (7), where the spectral density is now the continuous-time one (i.e., $f_X^c(\lambda) = s_0^2 \int e^{i\lambda u} \rho_X(u) du$).

Of course, the process is, in fact, discretely observed. It is worth noting that the correlation property is stable from continuous to discrete-time observations. The folding formula that links the continuous-time formula of spectral density to the discrete-time one (namely, if Δ is the sample step, $f_X(\lambda) = (1/\Delta) \sum_{n \in \mathbb{Z}} f_X^c((\lambda + 2\pi n)/\Delta)$) shows, on the other hand, that the spectral long-memory behavior is preserved when going from continuous-time observation to discrete-time sample for $0 < \alpha < 1/2$.

The additional difficulty here is that the volatility σ_t^2 process is not observed. Therefore, for statistical purpose and estimation of the parameters of the model, approximations must be used. As the first equation is standard, the quadratic variation of the observed log price process is the integrated volatility,

and therefore, if $\{t_i\}_i$ is a partition of the interval $[t, T]$, then

$$\int_t^T \sigma_u^2 du = \lim_{\max_i |t_{i+1} - t_i| \rightarrow 0} \sum_i [\ln(S_{t_{i+1}}) - \ln(S_{t_i})]^2 \quad (11)$$

where the limit holds in probability. Thus, it is possible, if high-frequency data are available, to build pseudo-observations of the process $(1/h) \int_t^{t+h} \sigma_u^2 du \sim \sigma_t^2$, for small values of h . Note that $Z_t = \int_t^{t+1} \sigma^2(u) du$ can also be consistently estimated with high-frequency data; it is stationary and long memory, with the same memory parameter α as σ^2 . Statistical inference for the parameters can then be implemented, by using Whittle contrast or more classical log-likelihood maximization.

From the point of view of option pricing, the general formula of option prices for a deterministic interest rate $r(t)$ under the no-free-lunch assumption is still valid. The call option price C_t is given by $C_t = B(t, T) \mathbb{E}^{\mathbb{Q}}[\max(0, S_T - K) | \mathcal{F}_t]$ where $B(t, T) = \exp(-\int_t^T r(u) du)$ is the price at t of a zero coupon bond of maturity T and $\mathbb{Q}$ is the probability equivalent to $\mathbb{P}$ under which the discounted price process is martingale. Then we get

$$C_t = S_t \left\{ \mathbb{E}^{\mathbb{Q}} \left[\Phi \left(\frac{m_t}{U_{t,T}} + \frac{U_{t,T}}{2} \right) \right] - e^{-m_t} \mathbb{E}^{\mathbb{Q}} \left[\Phi \left(\frac{m_t}{U_{t,T}} - \frac{U_{t,T}}{2} \right) \right] \right\} \quad (12)$$

where $m_t = \ln(S_t / (K B(t, T)))$, $U_{t,T} = \sqrt{\int_t^T \sigma_u^2 du}$ and $\Phi(u) = \int_{-\infty}^u e^{-t^2/2} dt / \sqrt{2\pi}$. The point is that the bivariate process $(\ln(S_t), \sigma_t^2)$ is no longer Markovian so that a conditional expectation given $\mathcal{F}_t$, the information set generated by the past of the process until t , does not only depend on the value of the process at time t but on the whole past of σ^2 . This is the price to pay for incorporating long-range dependence of σ^2 in the model. Practitioners are used to computing the so-called Black–Scholes implied volatility by inversion of the Black–Scholes option pricing formula on the observed option prices. It can be checked that long-memory properties are transmitted to many approximations of implied volatilities that can be deduced from the above formula, under the assumption that there is no premium for volatility risk (so

that conditional expectations under $\mathbb{Q}$ are the same as under the current probability measure $\mathbb{P}$). Then statistical strategies can be devised by using option prices, at least to estimate α , or to recover integrated volatilities.

References

- [1] Backus, D.K. & Zin, S.E. (1993). Long-memory inflation uncertainty: evidence from the term structure of interest rates, *Journal of Money, Credit, and Banking* **3**, 681–700.
- [2] Baillie, R.T., Bollerslev, T. & Mikkelsen, H.O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **74**, 3–30.
- [3] Beran, J. (1994). Statistics for long memory processes, in *Monographs on Statistics and applied Probability*, Chapman and Hall, New York, Vol. 61.
- [4] Bollerslev, T. & Mikkelsen, H.O. (1996). Modelling and pricing long memory in stock market volatility, *Journal of Econometrics* **73**, 151–184.
- [5] Box, G. & Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*, Holden Day.
- [6] Brockwell, P.J. & Davis, R.A. (1991). *Time Series: Theory and Methods*, 2nd Edition, Springer.
- [7] Carmona, P., Coutin, L. & Montseny, G. (2003). Stochastic integration with respect to fractional Brownian motion, *Annales de l'Institut Henri Poincaré, Probability and Statistics* **39**, 27–68.
- [8] Cheung, Y.-W. (1993). Long memory in foreign exchange rates, *Journal of Business and Economic Statistics* **11**, 93–101.
- [9] Comte, F. & Renault, E. (1998). Long memory in continuous time stochastic volatility models, *Mathematical Finance* **8**, 291–323.
- [10] Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes, *Annals of Statistics* **17**, 1749–1766.
- [11] Fox, R. & Taqqu, M.S. (1986). Large sample properties of parameter estimates for strongly dependent time series, *Annals of Statistics* **14**, 517–532.
- [12] Geweke, J. & Porter-Hudak, S. (1983). The estimation and application of long memory time series models, *Journal of Time Series Analysis* **4**, 221–238.
- [13] Giraitis, L., Leipus, R. & Surgailis, D. (2007). Recent advances in ARCH modelling, in *Long Memory in Economics*, eds G. Teyssière & A.P. Kirman, eds, Springer, Berlin, pp. 3–38.
- [14] Granger, C.W.J. & Joyeux, R. (1980). An introduction to long memory time series models and fractional differencing, *Journal of Time Series Analysis* **1**, 15–29.
- [15] Hosking, J.M.R. (1981). Fractional differencing, *Biometrika* **68**, 165–176.
- [16] Hu, Y., Oksendal, B. & Sulem, A. (2003). Optimal consumption and portfolio in a Black–Scholes market

- driven by fractional Brownian motion, *Infinite Dimensional Analysis Quantum Probability and Related Topics* **6**(4), 519–536.
- [17] Hull, J. & White, A. (1987). The pricing of options on assets with stochastic volatilities, *The Journal of Finance* **3**, 281–300.
 - [18] Mandelbrot, B.B. & Van Ness, J.W. (1968). Fractional Brownian motions, fractional noises and applications, *SIAM Review* **10**, 422–437.
 - [19] Nualart, D. (2006). Fractional Brownian motion: stochastic calculus and applications, in *Proceedings of the International Congress of Mathematicians (ICM), Madrid, Spain*, M. Sanz-Solè, J. Soria, J.L. Varona & J. Verdera, eds, European Mathematical Society, Zürich, Vol. III, Invited lectures, pp. 1541–1562.
 - [20] Robinson, P.M. (1995). Log periodogram regression of time series with long range dependence, *Annals of Statistics* **23**, 1630–1660.
 - [21] Rogers, L.C.G. (1997). Arbitrage with fractional Brownian motion, *Mathematical Finance* **7**, 95–105.
 - [22] Sowell, F. (1990). The fractional unit root distribution, *Econometrica* **58**, 495–505.

Further Reading

- Black, F & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **3**, 637–654.

Related Articles

Fractional Brownian Motion; Multifractals.

FABIENNE COMTE

Factor Models

Factor models of security returns decompose the random return on each of a cross section of assets into factor-related and asset-specific returns. Letting r denote the vector of random returns on n assets, and assuming k factors, a factor decomposition has the form

$$r = a + Bf + \varepsilon \quad (1)$$

where B is an $n \times k$ -matrix of factor betas, f is a random k -vector of factor returns, and ε is an n -vector of asset-specific returns. The n -vector of coefficients a is set so that $E[\varepsilon] = 0$. By defining B as the least squares projection $B = \text{cov}(r, f)C_f^{-1}$, it follows that $\text{cov}(f, \varepsilon) = 0^{k \times n}$.

The factor decomposition (1) puts no empirical restrictions on returns beyond requiring that the means and variances of r and f exist. So in this sense it is empty of empirical content. To add empirical structure, it is commonly assumed that the asset-specific returns ε are cross-sectionally uncorrelated, $E[\varepsilon\varepsilon'] = D$ where D is a diagonal matrix. This implies that the covariance matrix of returns can be written as the sum of a matrix of rank k and a diagonal matrix:

$$\text{cov}(r, r') = B\text{cov}(f, f')B' + D \quad (2)$$

This is called a *strict factor model*. Without loss of generality, one can assume that $\text{cov}(f, f')$ has rank k , since otherwise one of the factors can be removed (giving a $k - 1$ factor model) without affecting the fit of the model.

An important (and often troublesome) feature of factor models is their rotational indeterminacy. Let L denote any nonsingular $k \times k$ -matrix and consider the set of factors $f^* = Lf$ and factor betas $B^* = L^{-1}B$. Note that f^*, B^* can be used in place of f, B since only their matrix product affects returns and the linear “rotation” L disappears from this product. This means that factors f and associated factor betas B are only defined up to a $k \times k$ linear transformation. In order to empirically identify the factor model, one can set the covariance matrix of the factors equal to an identity matrix, $E[ff'] = I_k$, without loss of generality.

Approximate Factor Models

Security market returns have strong comovements, but the assumption that returns obey a strict factor model is easily rejected. In practice, for any reasonable value of k , there will be at least some discernible positive correlations between the asset-specific returns of at least some assets. An *approximate factor model* (originally developed by Chamberlain and Rothschild [7]) weakens the strict factor model of exactly zero correlations between all asset-specific returns. Instead, it assumes that there is a large number, n , of assets and the proportion of the correlations that are nonnegligibly different from zero is close to zero. This condition is formalized as a bound on the eigenvalues of the asset-specific return covariance matrix:

$$\lim_{n \rightarrow \infty} \max \text{eigval}[\text{cov}(\varepsilon, \varepsilon')] < c \quad (3)$$

for some fixed $c < \infty$. Crucially, this condition implies that asset-specific returns are *diversifiable risk* in the sense that any well-spread portfolio w will have asset-specific variance near zero:

$$\begin{aligned} \lim_{n \rightarrow \infty} w' \text{cov}(\varepsilon, \varepsilon') w &= 0 \text{ for any } w \\ \text{such that } \lim_{n \rightarrow \infty} w' w &= 0 \end{aligned} \quad (4)$$

Note that an approximate factor model uses a “large n ” modeling approach; the restrictions on the covariance matrix need only hold approximately as the number of assets n grows large.

Letting $V = \text{cov}(\varepsilon, \varepsilon')$ which is no longer diagonal, and choosing the rotation so that $\text{cov}(f, f') = I$, we can write the covariance matrix of returns as

$$\text{cov}(r, r') = BB' + V \quad (5)$$

In addition to equation (4), it is appropriate to impose the condition that $\lim_{n \rightarrow \infty} \min \text{eigval}[BB'] = \infty$. This ensures that each of the k factors represents a pervasive source of risk in the cross section of returns.

Statistical Factor Models

Financial researchers differentiate between characteristic-based, macroeconomic, and statistical factor models [2]). In a characteristic-based model, the

factor betas of asset are tied to observable characteristics of the securities, such as company size or the book-to-price ratio, or the industry categories to which each security belongs. In macroeconomic factor models, the factors are linked to the innovations in observable economic time series such as inflation and unemployment. In a statistical factor model, neither factors nor betas are tied to any external data sources and the model is identified from the covariances of asset returns alone.

Recall the convenient rotation $E[ff'] = I$ that allows us to write the strict factor model (1) as

$$\text{cov}(r, r') = BB' + D \quad (6)$$

Assuming that the cross section of return is multivariate normal and independent and identically distributed (i.i.d.) through time, the sample covariance matrix $\widehat{\text{cov}}(r, r')$ has a Wishart distribution. Imposing the strict factor model assumption (6) on the true covariance matrix, it is possible to estimate the set of parameters B, D by maximum likelihood. This maximum likelihood problem requires high-dimensional nonlinear maximization: there are $nK + n$ parameters to estimate in B, D . There is also an inequality constraint on the maximization problem: the diagonal elements of D must be nonnegative, since they represent variances. The solution to the maximum likelihood problem yields estimates of B and D , which correspond to the systematic and unsystematic risk measures. It is often the case that estimates of the time series of factors, f , are of interest. These are called *factor scores* in the statistical literature and can be obtained through cross-sectional generalized least squares (GLS) regressions of r on $\widehat{B}$

$$\widehat{f}_t = (\widehat{B}\widehat{D}^{-1}\widehat{B})^{-1}\widehat{B}\widehat{D}^{-1}r_t \quad (7)$$

See [5] for a review of the various iterative algorithms, which can be used to numerically solve the maximum likelihood factor analysis problem and estimate the factor scores. Roll and Ross [25] provide an early empirical application to equity returns data.

The first k eigenvectors of the return covariance matrix scaled by the square roots of their respective eigenvalues are called the k *principal components* of the covariance matrix. A restrictive version of the strict factor model is the *scalar factor model*, given by equation (2) plus the scalar matrix condition $D = \sigma_\varepsilon^2 I$. Under the assumption of a scalar factor model, the maximum likelihood

problem simplifies, and the principal components are the maximum likelihood estimates of the factor beta matrix B (the arbitrary choice of rotation is slightly different in this case). This provides a quick and simple alternative to maximum likelihood factor analysis, under the restrictive assumption $D = \sigma_\varepsilon^2 I$.

Asymptotic Principal Components

The maximum likelihood method of factor model estimation relies on a strict factor model assumption and a time-series sample which is large relative to the number of assets in the cross section. Standard principal components require the even stronger condition of a scalar factor model. Neither method is well suited for asset returns where the cross section tends to be very large. Connor and Korajczyk [11] develop an alternative method called *asymptotic principal components*, building on the approximate factor model theory of Chamberlain and Rothschild [7]. Connor and Korajczyk analyze the eigenvector decomposition of the $T \times T$ cross product matrix of returns rather than that of the $n \times n$ covariance matrix of returns. They show that given a large cross section, the first k eigenvectors of this cross-product matrix provide consistent estimates of the $k \times T$ matrix of factor returns. Stock and Watson [31] extend the theory to allow both large time series and large cross-sectional samples, time-varying factor betas, and provide a quasi-maximum likelihood interpretation of the technique. Bai [3] analyzes the large-sample distributions of the factor returns and factor beta matrix estimates in a generalized version of this approach.

Macroeconomic Factor Models

The rotational indeterminacy in statistical factor models makes these unsuitable for the application of factor models to many research problems. Statistical factor models do not allow the analyst to assign meaningful labels to the factors and betas; one can identify the k pervasive risks in the cross section of returns, but not what these risks represent in terms of economic and financial theory.

One approach to making the factor decomposition more interpretable is to rotate the statistical factors so that the rotated factors are maximally correlated with prespecified macroeconomic factors. If f_t is a

k -vector of statistical factors and m_t is a k -vector of macroeconomic innovations, we can regress the macroeconomic factors on the statistical factors:

$$m_t = \Pi f_t + \eta_t \quad (8)$$

As long as Π has rank k , the span of the rotated factors, Πf_t , is the span of the original statistical factors, f_t . However, the rotated factors can now be interpreted as the return factors that are correlated with the specified macroeconomic series. With this rotation the new factors are no longer orthogonal, in general. This approach is described in [12].

Alternatively, one can work with the prespecified macroeconomic series directly. Chan *et al.* [8] and Chen *et al.* [9] develop a macroeconomic factor model in which the factor innovations f are observed directly (using innovations in economic time series) and the factor betas are estimated via time-series regression of each asset's return on the time series of factors. They begin with the standard description of the current price of each asset, p_{it} , as the present discounted value of its expected cash flows:

$$p_{it} = \sum_{s=1}^{\infty} \frac{E[c_{it}]}{(1 + \rho_{st})^s} \quad (9)$$

where ρ_{st} is the discount rate at time t for expected cash flows at time $t + s$. Chen, Roll, and Ross note that the common factors in returns must be variables which cause pervasive shocks to expected cash flows $E[c_{it}]$ and/or risk-adjusted discount rates ρ_{st} . They propose inflation-, interest-rate-, and business-cycle-related variates to capture these common factors. Shanken and Weinstein [29] find that empirically the model lacks robustness in that small changes in the included factors or the sample period have large effects on the estimates. Connor [10] argues that although macroeconomic factor models are theoretically attractive since they provide a deeper explanation for return comovement than statistical factor models, their empirical fit is substantially weaker than statistical and characteristic-based models. Vassalou [33] argues, on the other hand, that the ability of the Fama–French model (see below) to explain the cross section of mean returns can be attributed to the fact that Fama–French factors provide good proxies for macroeconomic factors.

Characteristic-based Factor Models

A surprisingly powerful method for factor modeling of security returns is the characteristic-based factor model. Rosenberg [26] was the first to suggest that suitably scaled versions of standard accounting ratios (book-to-price ratio, market value of equity) could serve as factor betas. Using these predefined betas, he estimates the factor realizations f_t by cross-sectional regression of time- t asset returns on the predefined matrix of betas.

In a series of very influential papers, Fama and French [16, 17, 18] propose a two-stage method for estimating characteristic-based factor models. In the first stage, they sort assets into fractile portfolios based on book-to-price and market value characteristics. They use the differences between returns on the top and bottom fractile portfolios as proxies for the factor returns. They also include a market factor proxied by the return on a capitalization-weighted market index. In the second stage, the factor betas of portfolios and/or assets are estimated by time-series regression of asset returns on the derived factors. Carhart [6] and Jegadeesh and Titman [22, 23] show that the addition of a momentum factor (proxied by high 12-month return minus low 12-month return) adds explanatory power to the Fama–French three-factor model, both in terms of explaining comovements and mean returns. Goyal and Santa-Clara [19] and Ang *et al.* [1, 2] also find evidence for an own-volatility-related factor, for explaining both return comovements and mean returns.

Industry–Country Components Models

One of the most empirically powerful factor decompositions for equity returns is an error-components model using industry affiliations. This involves setting the factor beta matrix equal to zero/one dummies, with row i containing a one in the j th column if and only if firm i belongs to industry j . This is the simplest type of characteristic-based factor model of equity returns.

The first statistical factor is dominant in equity returns, accounting for 80–90% of the explanatory power in a multifactor model. The standard specification of an error-components model does not isolate the “first” factor since its influence is spread across the factors. Heston and Rouwenhorst [20] describe an alternative specification in which this factor is

separated from the k industry factors. They add a constant to the model, so that the expanded set of factors $k + 1$ is not directly identified (this lack of identification is sometimes called the dummy variable trap, referring to a model that includes a full set of zero-one dummies plus a constant). Heston and Rouwenhorst, impose an adding-up restriction on the estimated $k + 1$ factors: the set of industry factors must sum to zero. This adding-up restriction on the factors restores statistical identification to the model, requiring constrained least squared in place of standard least squares estimation. It also provides a useful interpretation of the estimated factors: the factor associated with the constant term is the “marketwide” or “first” factor, and the factors associated with the industry dummies are the extra-market industry factors. This specification has been widely adopted in the literature.

Heston and Rouwenhorst’s adding-up condition is particularly useful in a multicountry context. It allows one to include an overlapping set of country and industry dummies without encountering the problem of the dummy variable trap. Including a constant, an international industry–country factor model must impose adding-up conditions both on the estimated industry factors and on the estimated country factors. This type of country–industry specification is useful, for example, in measuring the relative contribution of cross-border and national influences to return comovements; see, for example, [21].

Determining the Number of Factors

In the case of maximum likelihood estimation of a strict factor model, it is possible to test for the correct number of factors by comparing the likelihood of the model with k factors to that of a $k + 1$ factor model. Under the standard assumptions, for large time-series samples, the ratio of the log likelihoods has an approximate chi-squared distribution. There are drawbacks to this test in the context of asset returns data and its performance has been problematic; see, for example, [14, 15], and [28], in which the test is found to be unreliable. The test may rely too strongly on the assumption of a strict factor model (an exactly diagonal covariance matrix of asset-specific returns), which is at best a convenient fiction in the case of asset returns. Also, it tests the hypothesis that the correct number of factors has been prespecified,

rather than optimally determining the best number of factors to use.

Connor and Korajczyk [13] derive a test for the number of factors, which is robust to having an approximate, rather than strict factor model. It is based on the decline in average idiosyncratic variance as additional factors are added.

Bai and Ng [4] take a different approach to the factor-number decision. They view the choice of the number of factors as a model selection problem, and build on the Akaike and Bayesian information criteria (BIC) information criteria-based tests for model selection. Bai and Ng compute the average asset-specific variance (both across time and across securities) in a model with k factors:

$$\sigma_\varepsilon^2(k) = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \hat{\varepsilon}_{it}^2 \quad (10)$$

The Bai–Ng procedure involves choosing k to minimize a degrees-of-freedom adjusted variant of average asset-specific variance:

$$k^* = \arg \min \sigma_\varepsilon^2 + kg(n, T) \quad (11)$$

where the penalty function $g(n, T)$ serves to compensate for the lost degrees of freedom when estimating the model with more factors.

Factor Beta Pricing Theory

A central concern in asset pricing theory is the determination of asset risk premia and their connection to sources of pervasive risk. Factor models have been central to this research area. In a factor beta pricing model, the expected return of each asset equals the risk-free return plus a linear combination of the factor betas of the assets, with the linear weights (factor risk premia) constant across securities:

$$E[r] = 1^n r_0 + B\pi \quad (12)$$

where r_0 is the risk-free return and π is a k –vector of factor risk premia. Important special cases include the capital asset pricing model [30] (Treynor, J.L. (1961). *Toward a Theory of Market Value of Risky Assets*. (unpublished working paper).) [32], the arbitrage pricing theory [27], the Fama–French model [17, 18], and [24] intertemporal capital asset pricing model.

Acknowledgments

Author Korajczyk would like to acknowledge financial support from the Zell Center for Risk Research and the Jerome Kenney Fund.

References

- [1] Ang, A., Hodrick, R.J., Xing, Y. & Zhang, X. (2006). The cross-section of volatility and expected returns, *Journal of Finance* **61**, 259–299.
- [2] Ang, A., Hodrick, R.J., Xing, Y. & Zhang, X. (2009). High idiosyncratic risk and low returns: international and further U.S. evidence, *Journal of Financial Economics* **91**, 1–23.
- [3] Bai, J. (2003). Inferential theory for factor models of large dimension, *Econometrica* **71**, 135–171.
- [4] Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models, *Econometrica* **70**, 191–221.
- [5] Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods*, Wiley, New York.
- [6] Carhart, M.M. (1997). On persistence in mutual fund performance, *Journal of Finance* **52**, 57–82.
- [7] Chamberlain, G. & Rothschild, M. (1983). Arbitrage, factor structure and mean–variance analysis in large asset markets, *Econometrica* **51**, 1305–1324.
- [8] Chan, K.C., Chen, N.-F. & Hsieh, D.A. (1985). An exploratory investigation of the firm size effect, *Journal of Financial Economics* **14**, 451–471.
- [9] Chen, N.-F., Roll, R. & Ross, S.A. (1986). Economic forces and the stock market, *Journal of Business* **59**, 383–403.
- [10] Connor, G. (1995). The three types of factor models: a comparison of their explanatory power, *Financial Analysts Journal* **51**, 42–46.
- [11] Connor, G. & Korajczyk, R.A. (1986). Performance measurement with the arbitrage pricing theory: a new framework for analysis, *Journal of Financial Economics* **15**, 373–394.
- [12] Connor, G. & Korajczyk, R.A. (1991). The attributes, behavior, and performance of U.S. mutual funds, *Review of Quantitative Finance and Accounting* **1**, 5–26.
- [13] Connor, G. & Korajczyk, R.A. (1993). A test for the number of factors in an approximate factor model, *Journal of Finance* **48**, 1263–1291.
- [14] Conway, D.A. & Reinganum, M.R. (1988). Stable factors in security returns: identification using cross-validation, *Journal of Business & Economic Statistics* **6**, 1–15.
- [15] Dhrymes, P.J., Friend, I. & Gultekin, N.B. (1984). A critical reexamination of the empirical evidence on the arbitrage pricing theory, *Journal of Finance* **39**, 323–246.
- [16] Fama, E.F. & French, K.R. (1992). The cross-section of expected stock returns, *Journal of Finance* **47**, 427–465.
- [17] Fama, E.F. & French, K.R. (1993). Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* **33**, 3–56.
- [18] Fama, E.F. & French, K.R. (1996). Multifactor explanations of asset pricing anomalies, *Journal of Finance* **51**, 55–84.
- [19] Goyal, A. & Santa-Clara, P. (2003). Idiosyncratic risk matters, *Journal of Finance* **58**, 975–1007.
- [20] Heston, S.L. & Rouwenhorst, K.G. (1994). Does industrial structure explain the benefits of international diversification? *Journal of Financial Economics* **36**, 3–27.
- [21] Hopkins, P.J.B. & Miller, C.H. (2001). *Country, Sector and Company Factors in Global Equity Models*, The Research Foundation of AIMR and the Blackwell Series in Finance, Charlottesville, VA.
- [22] Jegadeesh, N. & Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency, *Journal of Finance* **48**, 65–91.
- [23] Jegadeesh, N. & Titman, S. (2001). Profitability of momentum strategies: an evaluation of alternative explanations, *Journal of Finance* **56**, 699–718.
- [24] Merton, R.C. (1973). An intertemporal capital asset pricing model, *Econometrica* **41**, 867–887.
- [25] Roll, R. & Ross, S. (1980). An empirical investigation of the arbitrage pricing theory, *Journal of Finance* **35**, 1073–1103.
- [26] Rosenberg, B. (1974). Extra-market components of covariance in security returns, *Journal of Financial and Quantitative Analysis* **9**, 263–274.
- [27] Ross, S. (1976). The arbitrage theory of capital asset pricing, *Journal of Economic Theory* **13**, 341–360.
- [28] Shanken, J. (1987). Nonsynchronous data and the covariance-factor structure of returns, *Journal of Finance* **42**, 221–231.
- [29] Shanken, J. & Weinstein, M.I. (2006). Economic forces and the stock market revisited, *Journal of Empirical Finance* **13**, 129–144.
- [30] Sharpe, W.F. (1964). Capital asset prices: a theory of market equilibrium under conditions of risk, *Journal of Finance* **19**, 425–442.
- [31] Stock, J.H. & Watson, M.W. (2002). Macroeconomic forecasting using diffusion indexes, *Journal of Business and Economic Statistics* **20**, 147–162.
- [32] Treynor, J.L. (1999). Towards a theory of market value of risky assets, in *Asset Pricing and Portfolio Performance: Models, Strategy, and Performance Metrics*, R.A. Korajczyk, ed, Risk Publications, London.
- [33] Vassalou, M. (2003). News related to future gdp growth as a risk factor in equity returns, *Journal of Financial Economics* **68**, 47–73.

Related Articles

Capital Asset Pricing Model; Style Analysis.

GREGORY CONNOR & ROBERT KORAJCZYK

Generalized Method of Moments (GMM)

The generalized method of moments (GMM) provides a computationally convenient method of obtaining consistent and asymptotically normally distributed estimators of the parameters of statistical models. The method has been applied in many areas of economics but has arguably been most frequently applied in finance. In fact, it was in empirical finance that the power of the method was first illustrated: for while Hansen [7] introduced GMM and presented its fundamental statistical theory, Hansen and Hodrick [9] and Hansen and Singleton [10] showed the potential of the GMM approach to estimation through their empirical analyses of, respectively, foreign exchange markets and asset pricing. This article briefly describes the GMM framework for estimation and inference in the case where time-series data are used. The reader is referred to [6] for a comprehensive treatment of GMM that presents both a rigorous statistical analysis of the method and empirical illustrations in finance. For applications of GMM to panel data, see [15].

The popularity of GMM can be understood by comparing the requirements for the method to those for maximum likelihood (ML). While ML is the best available estimator within the paradigm of classical statistics, its optimality stems from its basis on the joint probability distribution of the data. However, in many scenarios of interest in finance, this dependence on the probability distribution can become a weakness for two main reasons. First, in some models, the joint probability distribution is only implicitly specified and is not available in a convenient closed form with the result that ML is computationally infeasible, for example, in stochastic volatility models [11]. Secondly, in other cases, the underlying theoretical model places restrictions on the distribution of the data but does not completely specify its form, with the result that ML is infeasible unless the researcher imposes an arbitrary assumption about the distribution. The latter is an unattractive strategy because if the assumed distribution is incorrect, then the optimality of ML is lost, and the resulting estimator may even be inconsistent; for example, in nonlinear Euler equation models [10]. In contrast, GMM estimation is based on population moment conditions. It turns

out that in many cases where the financial-theoretic model does not specify the complete distribution, it does specify population moment conditions. Therefore, in these settings, GMM can be preferred to ML because it offers a way to estimate the parameters solely on the basis of information deduced from the underlying financial model.

The cornerstone of GMM estimation is the population moment condition:

Definition 1 Population Moment Condition.

Let θ_0 be a vector of unknown parameters that are to be estimated, $\mathbf{v}_t$ be a vector of random variables, and $\mathbf{f}(\cdot)$ a vector of functions. Then a population moment condition takes the form

$$E[\mathbf{f}(\mathbf{v}_t, \theta_0)] = 0 \quad (1)$$

for all t .

To illustrate the types of population moment conditions that can arise in finance models, we consider the specific examples in the seminal papers [9] and [10].

Example 1 The efficiency of foreign exchange markets

Consider the following regression model for the relationship between the forward and spot exchange rates:

$$s_{t+k} = \beta_{0,1} + \beta_{0,2}f_{t,k} + u_t \quad (2)$$

where $f_{t,k}$ is the log of the k -period forward exchange rate at time t , s_{t+k} is the log of the spot exchange rate in period $t+k$, $\beta_{0,1}$ and $\beta_{0,2}$ are unknown parameters, and u_t is an error term. Hansen and Hodrick [9] demonstrate that under the simple efficient-markets hypothesis, the following population moment condition holds

$$E[f_{t,k}(s_{t+k} - \beta_{0,1} - \beta_{0,2}f_{t,k})] = 0 \quad (3)$$

where $(\beta_{0,1}, \beta_{0,2}) = (0, 1)$.

Example 2 The consumption-based asset-pricing model

The consumption-based asset-pricing model involves a representative agent who makes investment and consumption decisions at time t to maximize his/her discounted lifetime utility subject to a budget constraint. Assuming that the agent's utility function is of the constant relative risk aversion form and that the

only asset available as a possible investment yields a pay-off in the following period, Hansen and Singleton [10] show that the model implies the following population moment condition:

$$E \left[\mathbf{z}_t \left\{ \delta_0 \left(\frac{c_{t+1}}{c_t} \right)^{\gamma_0-1} \left(\frac{g_{t+1}}{p_t} \right) - 1 \right\} \right] = 0 \quad (4)$$

where c_t is consumption in period t , g_{t+1} is the gain (cash flow) on the asset in period $t+1$, p_t is the price of the asset in period t , $\mathbf{z}_t$ is a vector of variables in the agent's information set at time t , and the unknown parameters $1 - \gamma_0$ and δ_0 are, respectively, the agent's coefficient of relative risk aversion and discount factor.

The population moment condition states that $E[\mathbf{f}(\mathbf{v}_t, \boldsymbol{\theta})]$ equals zero when evaluated at $\boldsymbol{\theta}_0$. For the GMM estimation to be successful, this must be a unique property of $\boldsymbol{\theta}_0$, that is, $E[\mathbf{f}(\mathbf{v}_t, \boldsymbol{\theta})]$ is not equal to zero when evaluated at any other value of $\boldsymbol{\theta}$. If the latter holds, then $\boldsymbol{\theta}_0$ is said to be *identified* by $E[\mathbf{f}(\mathbf{v}_t, \boldsymbol{\theta}_0)] = 0$. A first-order condition for identification (often referred to as a *local condition*) is that $\text{rank}\{\mathbf{G}(\boldsymbol{\theta}_0)\} = p$, where $\mathbf{G}(\boldsymbol{\theta}) = E[\partial \mathbf{f}(\mathbf{v}_t, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}']$, and this plays a crucial role in standard asymptotic distribution theory for GMM. By definition, the moment condition involves q pieces of information about p unknowns; therefore identification can only hold if $q \geq p$. For reasons that emerge below, it is convenient to split this scenario into two parts: $q = p$, in which case $\boldsymbol{\theta}_0$ is said to be *just identified*, and $q > p$, in which case $\boldsymbol{\theta}_0$ is said to be *overidentified*.

Let $\mathbf{g}_T(\boldsymbol{\theta})$ be the analogous sample moment to $E[\mathbf{f}(\mathbf{v}_t, \boldsymbol{\theta})]$, that is, $\mathbf{g}_T(\boldsymbol{\theta}) = T^{-1} \sum_{t=1}^T \mathbf{f}(\mathbf{v}_t, \boldsymbol{\theta})$ where T is the sample size. The GMM estimator is then as follows.

Definition 2 Generalized Method of Moments Estimator.

The generalized method of moments estimator based on equation (1) is the value of $\boldsymbol{\theta}$ that minimizes

$$Q_T(\boldsymbol{\theta}) = \mathbf{g}_T(\boldsymbol{\theta})' \mathbf{W}_T \mathbf{g}_T(\boldsymbol{\theta}) \quad (5)$$

where $\mathbf{W}_T$ is known as the weighting matrix and is restricted to be a positive semidefinite matrix that converges in probability to $\mathbf{W}$, some positive definite matrix of constants.

The restrictions on the weighting matrix are required to ensure that $Q_T(\boldsymbol{\theta})$ is a meaningful measure of the distance the sample moment is from zero at different values of $\boldsymbol{\theta}$. The choice of weighting matrix is discussed further below.

The first-order conditions for this minimization are

$$\mathbf{G}_T(\hat{\boldsymbol{\theta}}_T)' \mathbf{W}_T \mathbf{g}_T(\hat{\boldsymbol{\theta}}_T) = 0 \quad (6)$$

where $\mathbf{G}_T(\boldsymbol{\theta}) = T^{-1} \sum_{t=1}^T \partial \mathbf{f}(\mathbf{v}_t, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$. The structure of these conditions reveals some interesting insights into GMM estimation. Since $\mathbf{G}_T(\boldsymbol{\theta})$ is $q \times p$, it follows that equation (6) involves calculating $\hat{\boldsymbol{\theta}}_T$ as the value of $\boldsymbol{\theta}$ that sets the p linear combinations of $\mathbf{g}_T(\cdot)$ to zero. Therefore, if $p = q$ (and $\mathbf{G}_T(\hat{\boldsymbol{\theta}}_T)' \mathbf{W}_T$ is nonsingular), then $\hat{\boldsymbol{\theta}}_T$ satisfies the analogous sample moment condition to equation (1), $\mathbf{g}_T(\hat{\boldsymbol{\theta}}_T) = 0$, and is, thus, the method of moments estimator based on the original moment condition. However, if $q > p$, then the first-order conditions are not equivalent to solving the sample moment condition. Instead, $\hat{\boldsymbol{\theta}}_T$ is equivalent to the method of moments estimator based on

$$\mathbf{G}(\boldsymbol{\theta}_0)' \mathbf{W} E[\mathbf{f}(\mathbf{v}_t, \boldsymbol{\theta}_0)] = 0 \quad (7)$$

Although equation (1) implies equation (7), the reverse does not hold because $q > p$; therefore, in this case, the estimation is actually based on only part of the original information. As a result, if $q > p$, then GMM can be viewed as decomposing the original moment condition into two parts, the *identifying restrictions*, which contain the information actually used in the estimation, and the *overidentifying restrictions*, which represent a remainder. Furthermore, GMM estimation produces two fundamental statistics each of which is associated with a particular component: the estimator $\hat{\boldsymbol{\theta}}_T$ is a function of the information in the identifying restrictions, and the estimated sample moment, $\mathbf{g}_T(\hat{\boldsymbol{\theta}}_T)$, is a function of the information in the overidentifying restrictions. While unused in estimation, the overidentifying restrictions play a crucial role in inference about the validity of the model as discussed below.

To establish the consistency and asymptotic normality of the GMM estimator, it is necessary to place restrictions on $\mathbf{v}_t$ and $\mathbf{f}(\cdot)$. In the standard asymptotic framework (see [7]), it is assumed that $\mathbf{v}_t$ is stationary and ergodic, $\mathbf{f}(\cdot)$ is continuous and differentiable and that its derivative is continuous in $\boldsymbol{\theta}$. Various expectations of $\mathbf{f}(\cdot)$ and its derivative must also exist.

These conditions permit the applications of laws of large numbers and central limit theorems to the requisite sample averages. Chief among the restrictions on expectations of $\mathbf{f}(\cdot)$ are that the population moment condition is valid and identifies θ_0 . Under the above conditions, it can be shown that the $\hat{\theta}_T \xrightarrow{p} \theta_0$ and

$$T^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, V) \quad (8)$$

where

$$V = [\mathbf{G}(\theta_0)' \mathbf{W} \mathbf{G}(\theta_0)]^{-1} \mathbf{G}(\theta_0)' \mathbf{W} \mathbf{S}(\theta_0) \mathbf{W} \mathbf{G}(\theta_0) \\ \times [\mathbf{G}(\theta_0)' \mathbf{W} \mathbf{G}(\theta_0)]^{-1}$$

and $\mathbf{S}(\theta) = \lim_{T \rightarrow \infty} \text{Var}[T^{1/2} \mathbf{g}_T(\theta)]$.

At this point, we return to the issue of the choice of weighting matrix. In the discussion of the first-order conditions above, it is noted that if $p = q$, then the GMM estimator can be found by solving $\mathbf{g}_T(\hat{\theta}_T)$. As a result, the estimator does not depend on $\mathbf{W}_T$. The asymptotic properties must also be invariant to $\mathbf{W}_T$ and it can be shown that $\mathbf{V}$ reduces to $\{\mathbf{G}(\theta_0)' \mathbf{S}(\theta_0)^{-1} \mathbf{G}(\theta_0)\}^{-1}$ in this case. However, if $q > p$, then the first-order conditions depend on $\mathbf{W}_T$ and therefore so does $\hat{\theta}_T$, in general. This dependence is unattractive because it raises the possibility that subsequent inferences can be affected by the choice of weighting matrix. However, in terms of asymptotic properties, the choice of weighting matrix only manifests itself in $\mathbf{V}$, the asymptotic variance of the estimator. Since this is the case, it is natural to choose $\mathbf{W}_T$ such that $\mathbf{V}$ is minimized in a matrix sense. Hansen [7] shows that this can be achieved by setting $\mathbf{W}_T = \hat{\mathbf{S}}_T^{-1}$ where $\hat{\mathbf{S}}_T$ is a consistent estimator of $\mathbf{S}(\theta_0)$. The resulting asymptotic variance is $\mathbf{V} = \mathbf{V}^0 = \{\mathbf{G}(\theta_0)' \mathbf{S}(\theta_0)^{-1} \mathbf{G}(\theta_0)\}^{-1}$; Chamberlain [4] shows $\mathbf{V}^0$ is a semiparametric efficiency bound for estimation of θ_0 based on equation (1).

In practical terms, two issues arise in the implementation of GMM with this choice of weighting matrix: (i) how to construct $\hat{\mathbf{S}}_T$ so that it is a consistent estimator of $\mathbf{S}(\theta_0)$; (ii) how to handle the dependence of $\hat{\mathbf{S}}_T$ on $\hat{\theta}_T$. We treat each in turn.

(i) Estimation of $\mathbf{S}(\theta_0)$.

Under stationarity and ergodicity and certain other technical restrictions, it can be shown that $\mathbf{S}(\theta_0) = \Gamma_0(\theta_0) + \sum_{i=1}^{\infty} \{\Gamma_i(\theta_0) + \Gamma_i(\theta_0)'\}$ where $\Gamma_i(\theta_0) = \text{Cov}[\mathbf{f}(\mathbf{v}_t, \theta_0), \mathbf{f}(\mathbf{v}_{t-i}, \theta_0)]$ is known as the i -lag autocovariance matrix of $\mathbf{f}(\mathbf{v}_t, \theta_0)$ [2]. In some cases, the

structure of the model implies $\Gamma_i(\theta_0) = 0$ for all $i > k$ for some k , and this simplifies the estimation problem ([6], Section 3.5). In the absence of such a restriction on the autocovariance matrices, the long-run variance can be estimated by a member of the class of *heteroscedasticity autocorrelation covariance (HAC)* estimators defined by

$$\hat{\mathbf{S}}_{\text{HAC}} = \hat{\mathbf{r}}_0 + \sum_{i=1}^{T-1} \omega(i; b_T) (\hat{\mathbf{r}}_i + \hat{\mathbf{r}}_i') \quad (9)$$

where $\hat{\mathbf{r}}_j = T^{-1} \sum_{t=j+1}^T \hat{\mathbf{r}}_t \hat{\mathbf{r}}_{t-j}'$, $\hat{\mathbf{r}}_t = \mathbf{f}(\mathbf{v}_t, \hat{\theta}_T)$, $\omega(\cdot)$ is known as the *kernel*, and b_T is known as the *bandwidth*. The kernel and bandwidth must satisfy certain restrictions to ensure $\hat{\mathbf{S}}_{\text{HAC}}$ is both consistent and positive semidefinite. As an illustration, Newey and West [13] propose the use of the kernel $\omega(i, b_T) = \{1 - i/(b_T + 1)\} \mathcal{I}\{i \leq b_T\}$ where $\mathcal{I}\{i \leq b_T\}$ is an indicator variable that takes the value 1 if $i \leq b_T$ and 0, otherwise. This choice is an example of a truncated kernel estimator because the number of included autocovariances is determined by b_T . For consistency, we require $b_T \rightarrow \infty$ with $T \rightarrow \infty$ but $b_T = o(T^{1/2})$. Various choices of kernel have been proposed and their properties analyzed: while theoretical rankings are possible, the evidence suggests that the choice of b_T is a far more important determinant of finite sample performance. Andrews [2] and Newey and West [14] propose data-based methods for the selection of b_T . See ([6], Section 3.5) for further discussion.

(ii) Dependence of $\hat{\mathbf{S}}_T$ on $\hat{\theta}_T$.

As is apparent from the above discussion, the calculation of a consistent estimator for $\mathbf{S}(\theta_0)$ requires knowledge of a (consistent) estimator of θ_0 . Therefore, to calculate a GMM estimator that attains the efficiency bound, a multistep procedure is used. In the first step, GMM is performed with an arbitrary weighting matrix; this preliminary estimator is then used in the calculation of $\hat{\mathbf{S}}_T$. In the second step, GMM estimation is performed with $\mathbf{W}_T = \hat{\mathbf{S}}_T^{-1}$. For obvious reasons, the resulting estimator is commonly referred to as the *two-step GMM estimator*. Instead of stopping after just two steps, the procedure can be continued so that on the i th step the GMM estimation is performed using $\mathbf{W}_T = \hat{\mathbf{S}}_T^{-1}$, where $\hat{\mathbf{S}}_T$ is based on the estimator from the $(i - 1)$

th step. This yields the so-called *iterated GMM estimator*.

While two steps are sufficient to attain the efficiency bound, simulation evidence suggests that there are often considerable gains to iteration in the sense of improvements in the quality of asymptotic theory as an approximation to finite sample behavior ([6], Chapter 6). Notwithstanding these improvements, there is evidence that the quality of the asymptotic approximation may be poor in some circumstances of interest and this motivated the development of alternative methods for attaining the efficiency bound. One such method is the *continuous-updating generalized method of moments (CUGMM) estimator* proposed by Hansen et al. [8]. The CUGMM estimator is the value of θ that minimizes $\mathbf{g}_T(\theta)' \mathbf{S}_T(\theta)^{-1} \mathbf{g}_T(\theta)$ where $\mathbf{S}_T(\theta)$ is a (matrix) function of θ such that $\mathbf{S}_T(\theta) \xrightarrow{p} \mathbf{S}(\theta)$ uniformly in θ . While the CUGMM estimator has the same limiting distribution as the iterated GMM estimator, Newey and Smith [12] and Anatolyev [1] demonstrate analytically that the continuous-updating estimator can be expected to exhibit lower finite sample bias than its two-step counterpart; see also [3].

The foregoing results are predicated on the assumption that equation (1) represents valid information about θ_0 , and, in this sense, that the model is correctly specified. In practice, it is desirable to assess whether the data appear consistent with the restriction implied by the population moment condition. As noted above, if $p = q$, then the first-order conditions force $\mathbf{g}_T(\hat{\theta}_T) = 0$ irrespective of whether equation (1) holds, and so the latter cannot be tested directly using the estimated sample moment, $\mathbf{g}_T(\hat{\theta}_T)$. However, if $q > p$, then $\mathbf{g}_T(\hat{\theta}_T) \neq 0$ because GMM estimation only imposes the identifying restrictions and ignores the overidentifying restrictions. The latter represent $q - p$ restrictions which are true if equation (1) is itself true. To motivate the most commonly applied test statistic, it is useful to recall two aspects of our discussion above: (i) the GMM minimand measures the distance of $\mathbf{g}_T(\theta)$ from zero; and (ii) the estimated sample moment contains information about overidentifying restrictions. Combining (i) and (ii), it can be shown that GMM minimand evaluated at $\hat{\theta}_T$ is a measure of how far the sample is from satisfying the overidentifying restrictions. This leads to the *overidentifying restrictions test*

statistic,

$$J_T = T \mathbf{g}_T(\hat{\theta}_T)' \hat{\mathbf{S}}_T^{-1} \mathbf{g}_T(\hat{\theta}_T) \quad (10)$$

which converges to a χ_{q-p}^2 distribution under the null hypothesis that equation (1) holds. Notice that J_T is conveniently calculated as the sample size times the two-step (or iterated) GMM minimand evaluated at the associated estimator. The overidentifying restrictions test is the standard model diagnostic within the GMM framework. Nevertheless, J_T does not have power against all alternatives of interest: Ghysels and Hall [5] show that J_T has no (local) power against parameter variation and so it is prudent to complement the overidentifying restrictions test with tests of structural stability ([6], Section 5.4).

As noted above, there is evidence that in some cases of interest, the aforementioned (standard) asymptotic theory can provide a poor approximation to the finite sample performance of the GMM estimator and its associated inference techniques. This has prompted researchers to develop ways of ameliorating these deficiencies. The proposed methods fall into three basic categories: (i) methods for *moment selection*, which are designed to determine the “best” set of moments to use with a view to then basing inference on the standard asymptotic theory described above; (ii) resampling methods such as the *bootstrap* that are valid under similar assumptions to the standard theory but aim to provide more accurate approximations to finite sample behavior; and (iii) *alternative asymptotic theories* that are based on different assumptions to the standard framework and are designed to provide more accurate approximations in certain scenarios of interest, such as *weak identification*. For further discussion of (i)–(iii), see ([6], Chapters 6, 7, and 8).

References

- [1] Anatolyev, S. (2005). GMM, GEL, Serial correlation, and asymptotic bias, *Econometrica* **73**, 983–1002.
- [2] Andrews, D. (1991). Heteroscedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica* **59**, 817–858.
- [3] Antoine, B., Bonnal, H. & Renault, E. (2007). On the efficient use of the informational content of estimating equations: implied probabilities and Euclidean Empirical Likelihood, *Journal of Econometrics* **138**, 461–487.
- [4] Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions, *Journal of Econometrics* **34**, 305–334.

-
- [5] Ghysels, E. & Hall, A. (1990). Are consumption based intertemporal asset pricing models structural? *Journal of Econometrics* **45**, 121–139.
- [6] Hall, A. (2005). *Generalized Method of Moments*, Oxford University Press, Oxford, UK.
- [7] Hansen, L. (1982). Large sample properties of Generalized Method of Moments estimators, *Econometrica* **50**, 1029–1054.
- [8] Hansen, L., Heaton, J. & Yaron, A. (1996). Finite sample properties of some alternative GMM estimators obtained from financial market data, *Journal of Business and Economic Statistics* **14**, 262–280.
- [9] Hansen, L. & Hodrick, R. (1980). Forward exchange rates as optimal predictors of future spot rates, *Journal of Political Economy* **87**, 829–853.
- [10] Hansen, L. & Singleton, K. (1982). Generalized instrumental variables estimation of nonlinear rational expectations models, *Econometrica* **50**, 1269–1286.
- [11] Melino, A. & Turnbull, S. (1990). Pricing foreign currency options with stochastic volatility, *Journal of Econometrics* **45**, 239–266.
- [12] Newey, W. & Smith, R. (2004). Higher order properties of GMM and generalized empirical likelihood estimators, *Econometrica* **72**, 219–255.
- [13] Newey, W. & West, K. (1987). A simple positive semi-definite heteroscedasticity and autocorrelation consistent covariance matrix, *Econometrica* **55**, 703–708.
- [14] Newey, W. & West, K. (1994). Automatic lag selection in covariance matrix estimation, *Review of Economic Studies* **61**, 631–653.
- [15] Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

ALASTAIR R. HALL

Stochastic Volatility Models

The Black–Scholes model (*see* **Black–Scholes Formula**) rests upon a number of assumptions that are, to some extent, “counterfactual”. Among these are continuity in time of the asset price process (they do not jump), the ability to hedge continuously without transaction costs, independent Gaussian returns, and constant volatility. Stochastic volatility models relax the last assumption by allowing volatility to be randomly varying; this is motivated by the following reason: the dependence of Black–Scholes implied volatilities computed from market prices of call and put options, known as *the smile curve* (*see* **Implied Volatility Surface**), is accounted for by stochastic volatility models. In other words, this modification of the Black–Scholes theory succeeds in one area where the classical model fails. In fact, it has been shown [18, 23] that any randomness in volatility produces a smile of implied volatilities. Empirical evidence for the smile curve of implied volatility appears in many studies, including [20], using data from before the 1987 crash, and in [16], where the postcrash skew is documented (*see* **Implied Volatility Surface**). Early work on fitting an implied volatility surface, includes [1], and [4, 6, 21]. An extensive empirical study of the stability of the fitted surfaces can be found in [5].

In fact, modeling volatility as a stochastic process is motivated *a priori* by empirical studies of stock price returns in which estimated volatility is observed to exhibit “random” characteristics. Additionally, the effects of transaction costs show up, under many models, as uncertainty in the volatility; fat-tailed returns distributions can be simulated by stochastic volatility; market “jump” phenomena are often best modeled as volatility jump processes. Stochastic volatility modeling is therefore more than a simple fix to one particular Black–Scholes assumption, but rather a powerful modification that describes a much more complex market.

Pricing derivatives in a market with stochastic volatility is an incomplete markets problem, a distinction explained below, and one which has far-reaching consequences, particularly for the hedging problem and the problem of parameter estimation. It is the

latter *inverse* problem that is the biggest mathematical and practical challenge introduced by such models.

Early surveys can be found in [9, 11], or [14], and books on stochastic volatility are also available [7, 17], and [10].

In most stochastic volatility models, the asset price $(X_t)_{t \geq 0}$ satisfies the stochastic differential equation

$$dX_t = \mu X_t dt + \sigma_t X_t dW_t \quad (1)$$

where $(\sigma_t)_{t \geq 0}$ is called the volatility process. It must satisfy some regularity conditions for the model to be well defined, but it does not have to be an Itô process: it can be a jump process, a Markov chain, and so on. In order for it to be a volatility, it should be positive. Unlike the implied deterministic volatility models (*see* **Local Volatility Model**), the volatility process is *not* perfectly correlated with the Brownian motion (W_t) . Therefore, volatility is modeled to have an independent random component of its own. This leads to an incomplete market, that is, derivatives cannot, in general, be perfectly hedged by trading the underlying asset and money market only. In fact, there is no *unique* equivalent martingale measure.

Mean-reverting Stochastic Volatility Models

Typically, volatility is taken to be an Itô process satisfying a stochastic differential equation driven by a second Brownian motion. This is the easiest way to incorporate correlation with stock price changes. One desirable feature is *mean reversion*. The term *mean reversion* refers to the characteristic (typical) time it takes for a process to get back to the mean level of its invariant distribution (the long-run distribution of the process) and “forget” about its past. From a financial modeling perspective, mean reverting refers to a linear pull-back term in the drift of the volatility process itself, or in the drift of some (underlying) process of which volatility is a function. Let us denote $\sigma_t = f(Y_t)$ where f is some positive function. Then mean-reverting stochastic volatility means that the stochastic differential equation for (Y_t) has the following form:

$$dY_t = \alpha(m - Y_t)dt + \cdots d\hat{Z}_t \quad (2)$$

where $(\hat{Z}_t)_{t \geq 0}$ is a Brownian motion correlated with (W_t) : $d\langle W, \hat{Z} \rangle_t = \rho dt$, with $|\rho| \leq 1$. It is often found

from financial data that $\rho < 0$, and there are economic arguments for a negative correlation or *leverage effect* between stock price and volatility shocks. From common experience and empirical studies, when volatility goes up, asset prices tend to go down. Some common driving processes (Y_t) are

LN Lognormal (non-mean-reverting):

$$dY_t = c_1 Y_t dt + c_2 Y_t d\hat{Z}_t \quad (3)$$

OU Ornstein–Uhlenbeck:

$$dY_t = \alpha(m - Y_t)dt + \beta d\hat{Z}_t \quad (4)$$

CIR Feller or Cox–Ingersoll–Ross:

$$dY_t = \kappa(m' - Y_t)dt + v\sqrt{Y_t}d\hat{Z}_t \quad (5)$$

Some models studied in the literature are listed in Table 1.

Classical references investigating specific stochastic volatility models are [2, 13, 15, 22, 24], and [25]. These particular models are chosen for their nice properties (e.g., positivity and mean reversion) and analytical tractability, rather than from deeper financial reasons. The Heston model is very popular for its computational tractability.

Probability density functions of the stock price obtained from simulations exhibit fatter tails because of the random volatility. In particular, the negative correlation causes the tails to be asymmetric: the left tail is fatter.

Option Pricing

Suppose first that there is an equivalent martingale measure $\mathbb{P}^*$ under which the discounted stock price

Table 1 Stochastic volatility models

Authors	Correlation	$f(y)$	Y Process
Hull–White	$\rho = 0$	$f(y) = \sqrt{y}$	Lognormal
Scott	$\rho = 0$	$f(y) = e^y$	Mean-reverting OU
Stein–Stein	$\rho = 0$	$f(y) = y $	Mean-reverting OU
Ball–Roma	$\rho = 0$	$f(y) = \sqrt{y}$	CIR
Heston	$\rho \neq 0$	$f(y) = \sqrt{y}$	CIR

$e^{-rt}X_t$ is a martingale. Then, if we price any (and all) derivatives with maturity time T and payoff H using the formula $V_t = \mathbb{E}^*\{e^{-r(T-t)}H|\mathcal{F}_t\}$, for all $t \leq T$ where $(\mathcal{F}_t)$ is the filtration generated by the two Brownian motions, then there is no arbitrage opportunity. Thus V_t is a possible price for the claim. Equivalent martingale measures are constructed by rewriting $\hat{Z}_t = \rho W_t + \sqrt{1 - \rho^2}Z_t$ where W and Z are independent standard Brownian motions under the physical measure $\mathbb{P}$, and by setting

$$W_t^* = W_t + \int_0^t \frac{(\mu - r)}{f(Y_s)} ds, \quad Z_t^* = Z_t + \int_0^t \gamma_s ds \quad (6)$$

The shift of the second independent Brownian motion does not affect the drift of X_t . By Girsanov’s theorem, (W^*) and (Z^*) are independent standard Brownian motions under a measure $\mathbb{P}^{*(\gamma)}$ defined by

$$\begin{aligned} \frac{d\mathbb{P}^{*(\gamma)}}{d\mathbb{P}} &= \exp\left(-\frac{1}{2} \int_0^T ((\theta_s^{(1)})^2 + (\theta_s^{(2)})^2) ds \right. \\ &\quad \left. - \int_0^T \theta_s^{(1)} dW_s - \int_0^T \theta_s^{(2)} dZ_s \right), \\ \theta_t^{(1)} &= \frac{\mu - r}{f(Y_t)}, \quad \theta_t^{(2)} = \gamma_t \end{aligned} \quad (7)$$

Here (γ_t) is *any* adapted (and suitably regular) process. Technically, one needs to make an assumption on the pair $(\frac{\mu - r}{f(Y_t)}, \gamma_t)$ so that $\mathbb{P}^{*(\gamma)}$ is well defined as a probability measure. In particular, this is the case if μ may depend on Y_t in such a way that the Sharpe ratio $\frac{\mu(Y_t) - r}{f(Y_t)}$ and γ_t are bounded. Then, under $\mathbb{P}^{*(\gamma)}$, the stochastic differential equations (1, 2) become

$$dX_t = rX_t dt + f(Y_t)X_t dW_t^* \quad (8)$$

$$dY_t = [\alpha(m - Y_t) - \beta\Lambda] dt + \beta d\hat{Z}_t^* \quad (9)$$

where $\hat{Z}_t^* = \rho W_t^* + \sqrt{1 - \rho^2}Z_t^*$, and $\Lambda = \rho \frac{(\mu - r)}{f(Y_t)} + \gamma_t \sqrt{1 - \rho^2}$. Any allowable choice of γ leads to an equivalent martingale measure $\mathbb{P}^{*(\gamma)}$ and to the *no-arbitrage* derivative prices

$$V_t = \mathbb{E}^{*(\gamma)}\{e^{-r(T-t)}H|\mathcal{F}_t\} \quad (10)$$

where the conditional expectation is taken with respect to the filtration generated by the two Brownian motions. In the Markovian case, $V_t = P(t, X_t, Y_t)$ where $P(t, x, y)$ is the solution of the pricing partial differential equation (PDE) obtained by the *Feynman–Kac formula*.

Special Case: Uncorrelated Volatility

It turns out that the $\rho = 0$ case is much easier to handle mathematically. In equity markets, it is widely believed $\rho < 0$, but some studies suggest ρ is close to zero in foreign-exchange data. We now assume $\rho = 0$ and (γ_t) is independent of the Brownian motion (W_t) driving the stock price so that under the risk-neutral probability $IP^{*(\gamma)}$ the volatility remains uncorrelated with (W_t) . The pricing formula (10) can be simplified under these assumptions. We can condition on the path of the volatility process and by iterated expectations, the price of a call option, for example, is then given by

$$C(t, X_t, Y_t) = \mathbb{E}^{*(\gamma)} \left\{ \mathbb{E}^{*(\gamma)} \{ e^{-r(T-t)} (X_T - K)^+ \times [\mathcal{F}_t, \sigma_s, t \leq s \leq T] | \mathcal{F}_t \} \right\} \quad (11)$$

The inner expectation is just the Black–Scholes computation with a time-dependent volatility. The answer is the Black–Scholes formula with appropriately averaged volatility, leading to the *Hull–White formula*:

$$C(t, x, y) = \mathbb{E}^{*(\gamma)} \left\{ C_{BS}(t, x; K, T; \sqrt{\overline{\sigma^2}}) | Y_t = y \right\} \quad (12)$$

where

$$\overline{\sigma^2} = \frac{1}{T-t} \int_t^T f(Y_s)^2 ds \quad (13)$$

and, for simplicity, we have assumed (Y_t) is a Markov process. Thus $\sqrt{\overline{\sigma^2}}$ is the root-mean-square time average of $\sigma(\cdot)$ over the remaining trajectory of each realization, and the call option price is the average over all possible volatility paths. This formula admits a generalization in the correlated case obtained simultaneously in [19] and [26].

Summary

The positive aspects of the stochastic volatility approach are as follows:

- It directly models the observed random behavior of market volatility.
- It allows to reproduce more realistic returns distributions, in particular fatter-than-lognormal tails.
- The asymmetry of the distribution is easily incorporated by correlating the noise sources.
- The smile/skew effect in option prices is exhibited in stochastic volatility models, where the correlation controls the skew.
- Stochastic volatility has also been successfully used in fixed income, foreign exchange, and credit markets.

However, new difficulties are associated with this approach:

- Volatility is not directly observed. As a consequence, estimation of the parameters of a specific model and the current level of volatility is not straightforward.
- There is no canonical stochastic volatility model that is generally accepted and the relevance of explicit formulas for particular models is not obvious.
- We have to deal with an incomplete market, which means that derivatives cannot be perfectly hedged with just the underlying. In addition, a volatility risk premium has to be estimated from option prices.
- Besides the choice of a particular model and its calibration, pricing with stochastic volatility models is a real challenge requiring sophisticated numerical methods. Various approximation techniques have also been proposed such as regular and singular perturbation methods in [7] and [8], WKB expansion for SABR models in [12], or small-time asymptotics in [3].

References

- [1] Avellaneda, M., Friedman, C., Holmes, R. & Samperi, D. (1997). Calibrating volatility surfaces via relative-entropy minimization, *Applied Mathematical Finance* 4(1), 37–64.

- [2] Ball, C. & Roma, A. (1994). Stochastic volatility option pricing, *Journal of Financial and Quantitative Analysis* **29**(4), 589–607.
- [3] Berestycki, H., Busca, J. & Florent, I. (2004). Computing the implied volatility in stochastic volatility models, *Communications on Pure and Applied Mathematics* **57**(10), 1352–1373.
- [4] Derman, E. & Kani, I. (1994). Riding on a smile, *Risk* **7**, 32–39.
- [5] Dumas, B., Fleming, J. & Whaley, R. (1998). Implied volatility functions: empirical tests, *Journal of Finance* **53**(6), 2059–2106.
- [6] Dupire, B. (1994). Pricing with a smile, *Risk* **7**, 18–20.
- [7] Fouque, J.-P., Papanicolaou, G. & Sircar, K. (2000). *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press.
- [8] Fouque, J.-P., Papanicolaou, G., Sircar, K. & Solna, K. (2003). Multiscale stochastic volatility asymptotics, *SIAM Journal Multiscale Modeling and Simulation* **2**(1), 22–42.
- [9] Frey, R. (1996). Derivative asset analysis in models with level-dependent and stochastic volatility, *CWI Quarterly* **10**(1), 1–34.
- [10] Gatheral, J. (2006). *The Volatility Surface: A Practitioner's Guide*, Wiley Finance.
- [11] Ghysels, E., Harvey, A. & Renault, E. (1996). Stochastic volatility, in *Statistical Methods in Finance*, G. Maddala & C. Rao, eds, Handbook of Statistics, North Holland, Amsterdam, Vol. 14, chapter 5, pp. 119–191.
- [12] Hagan, P., Kumar, D., Lesniewski, A. & Woodward, D. (2002). *Managing Smile Risk*, Willmott Magazine, pp. 84–108.
- [13] Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options, *Review of Financial Studies* **6**(2), 327–343.
- [14] Hobson, D. (1996). *Stochastic Volatility*. Technical report, School of Mathematical Sciences, University of Bath.
- [15] Hull, J. & White, A. (1987). The pricing of options on assets with stochastic volatilities, *Journal of Finance* **XLII**(2), 281–300.
- [16] Jackwerth, J. & Rubinstein, M. (1996). Recovering probability distributions from contemporaneous security prices, *Journal of Finance* **51**(5), 1611–1631.
- [17] Lewis, A. (2000). *Option Valuation under Stochastic Volatility*, Finance Press.
- [18] Renault, E. & Touzi, N. (1996). Option hedging and implied volatilities in a stochastic volatility model, *Mathematical Finance* **6**(3), 279–302.
- [19] Romano, M. & Touzi, N. (1997). Contingent claims and market completeness in a stochastic volatility model, *Mathematical Finance* **7**, 399–410.
- [20] Rubinstein, M. (1985). Nonparametric tests of alternative option pricing models, *Journal of Finance* **XL**(2), 455–480.
- [21] Rubinstein, M. (1994). Implied binomial trees, *Journal of Finance* **LXIX**(3), 771–818.
- [22] Scott, L. (1987). Option pricing when the variance changes randomly: theory estimation, and an application, *Journal of Financial and Quantitative Analysis* **22**(4), 419–438.
- [23] Sircar, K.R. & Papanicolaou, G.C. (1999). Stochastic volatility, smile and asymptotics, *Applied Mathematical Finance* **6**(2), 107–145.
- [24] Stein, E. & Stein, J. (1991). Stock price distributions with stochastic volatility: an analytic approach, *Review of Financial Studies* **4**(4), 727–752.
- [25] Wiggins, J. (1987). Option values under stochastic volatility, *Journal of Financial Economics* **19**(2), 351–372.
- [26] Willard, G. (1997). Calculating prices and sensitivities for path-independent derivative securities in multifactor models, *Journal of Derivatives* **5**, 45–61.

Related Articles

Barndorff-Nielsen and Shephard (BNS) Models; Bates Model; Heston Model; Implied Volatility in Stochastic Volatility Models; Implied Volatility: Market Models; Regime-switching Models; Time Change; Uncertain Volatility Model.

JEAN-PIERRE FOUQUE

Duration Models

Duration models appear when studying the moment in time that certain events occur. Next to general applications in economics and medical sciences, their financial applications are in the more specific field of market microstructure (transaction times of assets or derivatives in a given market), and also in corporate governance (tenure of management).

Denote the unknown times at which the events of interest occur as T_i , $i = 1, \dots, n$. Almost invariably in the literature, it is assumed that the times T_i form an *increasing* sequence of *stopping times*. The random times are increasing if $T_{i+1} \geq T_i$ (almost surely (a.s.)). In that case, the *duration* D_i is defined by $D_i = T_i - T_{i-1} \geq 0$, with the usual convention $T_0 = 0$, which explains the name *duration models*. The stopping time property is with respect to a given underlying filtration $(\mathcal{F}_t)_{t \geq 0}$. This filtration is supposed to model the information available to either the investigator (in order to perform inference) or an agent in a financial market (in order to take, e.g., investment decisions). Care is to be taken when extrapolating results on inference to actual financial market decisions if these information sets are different (where, often, the investigator's information set is smallest). Mathematically, the information structure is generally assumed to satisfy the so-called *usual conditions* [5].

The stopping-time property of T_i formally states that, for each $t \geq 0$, the event $\{T_i \leq t\}$ belongs to $\mathcal{F}_t$. As the filtration is, by definition, increasing ($\mathcal{F}_s \subset \mathcal{F}_t$ for $s \leq t$), we also have $\{T_i \leq s\} \in \mathcal{F}_t$ for $s \leq t$. Thus, informally, if T_i is a stopping time denoting the occurrence of an event then, at each time $t \geq 0$, we know whether the event has occurred already. Also, in case the event has occurred by time t , we know at which time $s \leq t$ it occurred.

Scale Models

A popular way to model the times T_i or, equivalently, the durations D_i is *via* a scale model. This method became popular in the financial econometrics literature because of its resemblance to volatility modeling (see, for instance, the *autoregressive conditional duration* (ACD) specification of [2]). In

these models, we write the i th duration as the product of two positive random variables, $D_i = \psi_i \varepsilon_i$. Here $(\varepsilon_i)_{i=1}^n$ is some *innovation sequence* and the properties of the model depend on the assumptions made on this innovation sequence. The scale factor ψ_i is assumed to be predetermined, meaning that ψ_i is measurable with respect to $\mathcal{F}_{T_{i-1}}$, the information available at time T_{i-1} . This information set allows the formalization of, for example, the expectation given the information at T_{i-1} and, in particular, leads to the *optional stopping theorem* [5]. Most of the literature identifies the scale of the innovation sequence by $E\{\varepsilon_i | \mathcal{F}_{T_{i-1}}\} = 1$. Such an assumption is innocuous insofar as the modeling is concerned, but may change the interpretation of parameters specifying ψ_i and thus inferential aspects as (semiparametric) adaptivity [1]. In the ACD specification, it is assumed that $\mathcal{F}_{T_i}$ is observed by the investigator, so that ψ_i is observed as well (up to unknown parameters). An alternative is given by the stochastic volatility duration (SVD) model of [3], where the expected duration ψ_i is modeled using an unobserved factor allowing for unobserved heterogeneity.

Point or Counting Process Models

In the probabilistic literature, point processes refer to stochastic processes that generate random points in a set [4]. Applied to the positive half line, this leads to models for random times/durations. As describing the location of certain points is equivalent to count the number of points in general sets, these processes are also called *counting processes*. The distributional properties of these processes are characterized by, for instance, the (conditional) *intensity measure*, which essentially gives the (conditionally) expected number of events in a small (time) interval.

Passage Time Models

A structural approach to duration models can be obtained, using as a duration the time it takes for a stochastic process to pass a certain threshold. In particular, if the underlying process is a Brownian motion this leads to Brownian hitting times, which are well studied. In this simplest setting, these hitting times (durations) are distributed according to the

inverse Gaussian distribution, which also explains the name for this distribution.

Proportional Hazard and Survival Models

Survival analysis is used in (bio)statistics to model uncertain lifetimes and can be adopted to model other (financial) durations. This literature uses, in particular, the *hazard* rate to model the duration of an event. The hazard rate is the (instantaneous) probability that the event will occur in the next time instance, given that it has not occurred so far. Mathematically, the hazard rate equals the density of the duration distribution, divided by its survival function (one minus the cumulative distribution function). In particular, the proportional hazard model has been used a lot. In such a model, predetermined variables multiply the baseline hazard, to give a conditional hazard.

Multivariate Models

In financial applications, there is much interest in jointly modeling several durations, for example, the transaction times for multiple assets. Owing to the nonsynchronous occurrence of events, this is a non-trivial exercise. In general, the point process approach is most suited in these circumstances, as the multivariate intensities can be modeled. Inference, though, is fairly involved.

Similarly, one is often interested in the joint behavior of durations and other quantities as volume or volatility. Such multivariate extensions can be complicated by possible subtle causality relations (both instantaneous and Granger) between the processes. Moment-based inference that takes these effects into account has been proposed in [6].

References

- [1] Drost, F.C. & Werker, B.J.M. (2004). Semiparametric duration models, *Journal of Business and Economic Statistics* **22**, 40–50.
- [2] Engle, R.F. & Russell, J.R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data, *Econometrica* **66**, 1127–1162.
- [3] Ghysels, E., Gouriéroux, C. & Jasiak, J. (2004). Stochastic volatility duration models, *Journal of Econometrics* **119**, 413–433.
- [4] Last, G. & Brandt, A. (1995). *Marked Point Processes on the Real Line: The Dynamic Approach*, Springer-Verlag.
- [5] Protter, P.E. (1990). *Stochastic Integration and Differential Equations : A New Approach*, Springer-Verlag.
- [6] Renault, E. & Werker, B.J.M. (2009). Causality effects in return volatility measures with random times, *Journal of Econometrics* Forthcoming.

Related Articles

High-frequency Data; Order Flow; Point Processes; Poisson Process.

BAS J.M. WERKER

Mixed Data Sampling

Regression models that involve data sampled at different frequencies, or the so-called MIXED DATA Sampling (MIDAS) regression models, were introduced in both filtering and regression context in a number of recent papers, including [15, 17, 19], among others. These regressions deal with a situation often encountered in practice. For example, macroeconomic data are typically sampled at monthly or quarterly frequencies, whereas financial time series are sampled at almost any arbitrarily high frequency. Although most economic time series are not sampled at the same frequency, the typical practice of estimating econometric models involves aggregating all variables to the same (low) frequency using an equal weighting scheme.

Temporal aggregation in regression models, in contrast, is a well-studied area. The problem of temporal aggregation concerning dynamic models has been considered in the work of many authors, for example, [5, 11, 14, 21, 24, 26, 30–33, 36, 39–41].

In the first section, we draw comparisons between temporal aggregation and MIDAS. In the second section, we focus on a specific topic of special interest in finance–volatility forecasting.

Temporal Aggregation and MIDAS

MIDAS regressions not only share some features with distributed lag models but also have unique novel features. A stylized distributed lag model is a regression model of the following type: $y_t = \beta_0 + B(L)x_t + \varepsilon_t$, where $B(L)$ is some finite or infinite lag polynomial operator, usually parameterized by a small set of hyperparameters (see, e.g., [8] for a survey on distributed lag models). In a standard MIDAS regression model, the regressor is sampled m times more frequently than the regressand, where the latter has a sample of T observations. The asymptotics are based on T for a given m . Therefore, one studies the asymptotic properties of the estimators assuming that the span of the data set T grows and the high-frequency sample size of the regressors would be mT such that when $T \rightarrow \infty$ then both the low- and high-frequency samples tend to become large.

Andreou *et al.* [2] decompose the MIDAS regression model into two terms: the equal or flat aggregated term and a MIDAS term, which involves

weighted higher order differences of the high-frequency process. This decomposition allows one to link the MIDAS regression model with the traditional temporal aggregation literature, which omits the MIDAS term.

Consider the mixed data sampling data $\{y_t, \mathbf{x}_{t/m}^{(m)}\}$ where $\mathbf{x}_{t/m}^{(m)} = (1, x_{2,t/m}^{(m)}, \dots, x_{p,t/m}^{(m)})$ is a p -dimensional vector of higher frequency data observed at most m time between t and $t - 1$. The data are assumed to be weakly dependent. The conventional approach in empirical macroeconomics and finance employs the linear regression based on aggregated data, which are typically obtained by a pre-estimation procedure that computes simple averages to bring the high-frequency data to a common (low) frequency. In other words, the conventional approach employs the linear regression subject to the temporal aggregation restriction of an equal weighting scheme. Define the variable $\mathbf{x}_t^A = \frac{1}{m} \sum_{j=1}^m L^{j/m} \mathbf{x}_{t/m}^{(m)}$ that temporally aggregates the covariates $\mathbf{x}_{t/m}^{(m)}$ using equal weights (simple average), where q is the number of high-frequency lags used in the temporal aggregation of $\mathbf{x}_{t/m}^{(m)}$ with $L^{j/m}$ being the high-frequency lag operator. Then the conventional approach estimates by Least squares (LS) the following linear regression model:

$$y_t = \mathbf{x}_t^A \boldsymbol{\beta}^* + u_t^*, \quad q = 2, 3, \dots \quad (1)$$

where $E(u_t^* | \mathbf{x}_t^A) = 0$ and $E(u_t^{*2} | \mathbf{x}_t^A) = \sigma^{*2}$.

Consider now a projection of the high-frequency data onto y and let the linear polynomial of the projection be $\mathbf{W}(L^{1/m}; \boldsymbol{\theta})$, which is parameterized by a low-dimensional $r \times 1$ vector of parameters $\boldsymbol{\theta}$. Then we can define a MIDAS regression model as

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t(\boldsymbol{\theta}) + u_t, \quad t = 1, 2, \dots, T \quad (2)$$

where $E(u_t | \mathbf{x}_{t/m}^{(m)}) = 0$ and $E(u_t^2 | \mathbf{x}_{t/m}^{(m)}) = \sigma^2$ where $\mathbf{x}_t(\boldsymbol{\theta}) = \left(1, x_{2,t}^{(m_2)}(\boldsymbol{\theta}_2), \dots, x_{p,t}^{(m_p)}(\boldsymbol{\theta}_p)\right)'$ is a nonlinear function that maps the higher frequency data into a common low frequency and

$$\begin{aligned} x_{k,t}^{(m)}(\boldsymbol{\theta}_k) &= W(L^{1/m}; \boldsymbol{\theta}_k) x_{k,t/h}^{(m)} \\ &= \sum_{j=1}^q w_{j,k}(\boldsymbol{\theta}_k) L^{j/m} x_{k,t/m}^{(m)}, \\ k &= 2, \dots, p \end{aligned} \quad (3)$$

where $w_{j,k}(\boldsymbol{\theta}_k) \in (0, 1)$ and $\sum_{j=1}^m w_{j,k}(\boldsymbol{\theta}_k) = 1$. The latter assumption allows the identification of the slope

coefficient vector β . Equation (2) is a nonlinear regression that allows both the weighting scheme of the temporal aggregation as parameterized by the vector θ and the slope coefficient vector β to be estimated from the data. Following Ghysels *et al.* [17] the lag coefficients, $w_{j,k}(\theta_k)$, which determine the weighting scheme used, are parsimoniously parameterized *via* exponential Almon, Almon, Beta, or step function schemes.

Andreou *et al.* [2] study regression models that involve data sampled at different frequencies and compare them with a traditional model that involves temporal aggregation for the estimation of a model at the same sampling frequency that implies an equal weighting scheme. They derive the asymptotic properties of the MIDAS nonlinear least squares (MIDAS-NLS) estimators of regression models with mixed frequencies and general weighting schemes and show that the LS estimator is always relatively less efficient than the MIDAS-NLS estimator. Our analysis reveals the effects of the aggregation horizon on the properties of the MIDAS-NLS and temporally aggregated or FLAT-LS estimators. In particular, they show that the FLAT-LS estimator is asymptotically biased and inefficient when the high-frequency regressor exhibits linear temporal dependence as in the case of AR(1), while it is only inefficient in the case of *i.i.d.* and ARCH(1) high-frequency processes. Furthermore, they propose two types of aggregation bias tests: a simple LM test and a Wu–Hausman-type test that uses the high-frequency variables as instruments.

To conclude, it is worth discussing one particular application of regression models, namely, testing for Granger causality among variables—a topic of interest in many financial applications. Much has been written about the spurious effects temporal aggregation may have on testing for Granger causality; see, for example [4, 20, 22, 23, 27, 28, 34], among others. Ghysels and Valkanov [18] show that there are unexplored advantages in testing Granger causality in combining the data sampled at the different frequencies. They develop a set of Granger causality tests that explicitly take advantage of data sampled at different frequencies. Along similar lines, Andreou *et al.* (2009) document many predictable patterns between daily financial data and monthly and quarterly macroeconomic variables. Such patterns significantly improve macroeconomic forecasts. Recent work on this topic includes improving quarterly macroforecasts with

monthly data using MIDAS regressions [3, 7, 13, 35, 38] or improving quarterly and monthly macroeconomic predictions with daily financial data [2]; Ghysels and Wright (2008); [25, 37].

Volatility Forecasting with MIDAS

The original work on MIDAS focused on volatility predictions [1, 6, 10, 12, 15, 16], among others. The abundance of high-frequency data, combined with the desire to predict volatility over longer horizons, provides an ideal setting for applying such regressions. The original work on MIDAS was in fact a filtering and not a regression setting. Ghysels *et al.* [15] studied the risk-return trade-off—where returns were computed over longer horizons, whereas risk was measured by conditional variance over the same horizon but measured with a filter of weighted daily returns. In [16], the same approach was adopted in a regression format to predict multihorizon volatility. Chen and Ghysels [6] take this a step further and use a semiparametric MIDAS regression that related news impact curves applied to intradaily data to future realized volatility. They find that returns have significant asymmetric effects on future volatility—a topic somehow lost in the recent realized volatility literature—as originally emphasized in the context of ARCH-type models by Engle and Ng (1993). As mentioned in the previous section, this work distinguishes itself from the more standard temporal aggregation literature in the context of volatility [9, 29].

References

- [1] Alper, C., Fendoglu, S. & Saltoglu, B. (2008). *Forecasting Stock Market Volatilities Using MIDAS Regressions: An Application to the Emerging Markets*. MPRA Paper No. 7460.
- [2] Andreou, E., Ghysels, E. & Kourtellis, A. (2008). *Should Macroeconomic Forecasters look at Daily Financial Data?* Tech. rep., Discussion Paper, UNC and University of Cyprus.
- [3] Andreou, E., Ghysels, E. & Kourtellis, A. (2009). Regression models with mixed sampling frequencies, *Journal of Econometrics* (forthcoming).
- [4] Armesto, M., Hernandez-Murillo, R., Owyang, M. & Piger, J. (2008). Measuring the information content of the beige book: a mixed data sampling approach, *Journal of Money, Credit and Banking* forthcoming.

- [5] Breitung, J. and Swanson, N. (2002). Temporal aggregation and spurious instantaneous causality in multiple time series models, *Journal of Time Series Analysis* **23**, 651–665.
- [6] Brewer, K. (1973). Some consequences of temporal aggregation and systematic sampling for ARMA and ARMAX models, *Journal of Econometrics* **1**, 133–154.
- [7] Chen, X. & Ghysels, E. (2008). *News-Good or Bad and its Impact on Volatility Predictions Over Multiple Horizons*. Discussion Paper, UNC.
- [8] Clements, M. & Galvão, A. (2008). Macroeconomic forecasting with mixed frequency data: forecasting US output growth, *Journal of Business and Economic Statistics* **26**, 546–554.
- [9] Dhrymes, P. (1971). *Distributed Lags: Problems of Estimation and Formulation*, Holden-Day, San Francisco.
- [10] Drost, F. & Nijman, T. (1993). Temporal aggregation of GARCH processes, *Econometrica* **61**, 909–927.
- [11] Engle, R., Ghysels, E. & Sohn, B. (2008). *On the Economic Sources of Stock Market Volatility*. Discussion Paper, NYU and UNC.
- [12] Engle, R.F. & Ng, V.K. (1993). Measuring and testing the impact of news on volatility, *Journal of Finance* **48**(5), 1749–1778.
- [13] Engle, R. & Liu, T. (1972). Effects of aggregation over time on dynamic characteristics of an econometric model, in *Econometric Models of Cyclical Behaviors*, B. G. Hickman ed, National Bureau of Economic Research, Inc, Columbia, UP, Vol. 2.
- [14] Forsberg, L. & Ghysels, E. (2006). Why do absolute returns predict volatility so well? *Journal of Financial Econometrics* **6**, 31–67.
- [15] Galvão, A. (2006). *Changes in Predictive Ability with Mixed Frequency Data*. Discussion Paper, Queen Mary University.
- [16] Geweke, J. (1978). Temporal aggregation in the multiple regression model, *Econometrica* **46**, 643–661.
- [17] Ghysels, E., Santa-Clara, P. & Valkanov, R. (2005). There is a risk-return tradeoff after all, *Journal of Financial Economics* **76**, 509–548.
- [18] Ghysels, E., Santa-Clara, P. & Valkanov, R. (2006a). Predicting volatility: getting the most out of return data sampled at different frequencies, *Journal of Econometrics* **131**, 59–95.
- [19] Ghysels, E., Sinko, A. & Valkanov, R. (2006b). MIDAS regressions: further results and new directions, *Econometric Reviews* **26**, 53–90.
- [20] Ghysels, E. & Wright, J. (2007). Forecasting professional forecasters, *Journal of Business and Economic Statistics* forthcoming.
- [21] Ghysels, E. & Valkanov, R. (2008). *Granger Causality Tests with Mixed Data Frequencies*. Working paper, UNC and UCSD.
- [22] Ghysels, E. & Wright, J. (2009). Forecasting professional forecasters, *Journal of Business and Economic Statistics* (forthcoming).
- [23] Granger, C. (1980). Testing for causality: a personal viewpoint, *Journal of Economic Dynamics and Control* **2**, 329–352.
- [24] Granger, C. (1987). Implications of aggregation with common factors, *Econometric Theory* **3**, 208–222.
- [25] Granger, C. (1988). Some recent developments in a concept of causality, *Journal of Econometrics*, **39**, 199–211.
- [26] Granger, C. (1995). Causality in the long run, *Econometric Theory* **11**, 530–536.
- [27] Griliches, Z. (1967). Distributed lags: a survey, *Econometrica* **35**, 16–49.
- [28] Hamilton, J. (2006). *Daily Monetary Policy Shocks and the Delayed Response of New Home Sales*. Working paper, UCSD.
- [29] Hsiao, C. (1979). Linear regression using both temporally aggregated and temporally disaggregated data, *Journal of Econometrics* **10**(2), 243–252.
- [30] Lütkepohl, H. (1993). Testing for causation between two variables in higher dimensional VAR models, in *Studies in Applied Econometrics*, H. Schneeweiss & K. Zimmerman, eds, Springer-Verlag, Heidelberg, p. 75.
- [31] McCrorie, J. & Chambers, M. (2006). Granger causality and the sampling of economic processes, *Journal of Econometrics* **132**, 311–336.
- [32] Meddahi, N. & Renault, E. (2004). Temporal aggregation of volatility models, *Journal of Econometrics* **119**, 355–379.
- [33] Mundlak, Y. (1961). Aggregation over time in distributed lag models, *International Economic Review* **2**, 154–163.
- [34] Phillips, P. (1972). The structural estimation of a stochastic differential equation system, *Econometrica* **40**, 1021–1041.
- [35] Phillips, P. (1973). The problem of identification in finite parameter continuous time models, *Journal of Econometrics* **1**, 351–362.
- [36] Phillips, P. (1974). The estimation of some continuous time models, *Econometrica* **42**, 803–824.
- [37] Renault, E., Sekkat, K. & Szafarz, A. (1998). Testing for spurious causality in exchange rates, *Journal of Empirical Finance* **5**, 47–66.
- [38] Schumacher, C. & Breitung, J. (2008). Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data, *International Journal of Forecasting*, **24**, 386–398.
- [39] Sims, C. A. (1971). Discrete approximations to continuous time distributed lags in econometrics, *Econometrica* **39**, 545–563.
- [40] Tay, A. (2006). *Financial Variables as Predictors of Real Output Growth*. Discussion Paper, SMU.
- [41] Tay, A. (2007). *Mixing Frequencies: Stock Returns as a Predictor of Real Output Growth*. Discussion Paper, SMU.
- [42] Theil, H. (1954). *Linear Aggregation of Economic Relationships*, North Holland.

- [43] Tiao, G.C. & Wei W.S., (1976). Effect of temporal aggregation on the dynamic relationship of two time series variables, *Biometrika* **63**, 513–523.
- [44] Zellner, A. & Montmarquette, C. (1971). A study of some aspects of temporal aggregation problems in econometric analysis, *Review of Economic Statistics* **53**, 335–342.
- Colacito, R., Engle, R. & Ghysels, E. (2008). *A Component Model of Dynamic Correlations*, Discussion Paper UNC.
- Ghysels, E., Santa-Clara, P. & Valkanov, R. (2003). *The MIDAS Touch: Mixed Data Sampling Regression Models*, Working paper, UNC and UCLA.
- Ghysels, E. Rubia, A & Valkanov, R. (2009). Multi-Period Forecasts of Volatility: Direct, Iterated, and Mixed-Data Approaches, *Discussion Paper UNC*.
- Ghysels, E. & Sinko, A. (2009). Volatility forecasting and microstructure noise, *Journal of Econometrics* (forthcoming).

Further Reading

- Anderson, E., Ghysels, E. & Juergens, J. (2009). The impact of risk and uncertainty on expected returns, *Journal of Financial Economics* (forthcoming).
- Chen, X., Ghysels, E. & Wang, F. (2009). *The HYBRID GARCH Model*, Discussion Paper UNC.

ERIC GHYSELS

Pricing Kernels

Pricing kernels or stochastic discount factors (SDFs) (see **Stochastic Discount Factors**) are used to represent valuation operators in dynamic stochastic economies. A kernel is a commonly used mathematical term for representing an operator. The term *stochastic discount factor (SDF)* extends concepts from economics and finance to include adjustments for risk. As we will see, there is a close connection between the two terms. The terms *pricing kernel* and SDF are often used interchangeably.

After deriving convenient representations for prices, we provide several examples of stochastic discount factors and discuss econometric methods for estimating and testing asset pricing models that restrict stochastic discount factors.

Representing Prices

We follow Ross [49] and Harrison and Kreps [34] by exploring the implications of no arbitrage in frictionless markets to obtain convenient representations of pricing operators. We build these operators as mappings that assign prices that trade in competitive markets to payoffs on portfolios of assets. The payoffs specify how much a numeraire good is provided in alternative states of the world. To write these operators, we invoke the *law of one price*, which stipulates that any two assets with the same payoff must necessarily have the same price. This law is typically implied by but is weaker than the *principle of no arbitrage* (see **Fundamental Theorem of Asset Pricing**). Formally, the principle of no arbitrage stipulates that nonnegative payoffs that are positive with positive (conditional) probability command a strictly positive price. To see why the principle implies the law of one price, suppose that there is such a nonnegative portfolio payoff and call this portfolio a . If two portfolios, say b and c have the same payoff but the price of portfolio b is less than that of portfolio c , then an arbitrage opportunity exists. This can be seen by taking a long position on portfolio b , a short position on portfolio c , and using the proceeds to purchase portfolio a . This newly constructed portfolio has zero price but a positive payoff, violating the principle of no arbitrage.

Given that we can assign prices to payoffs, let $\pi_{t,t+1}$ be the date t valuation operator for payoffs (consumption claims) at date $t + 1$. This operator assigns prices or values to portfolio payoffs p_{t+1} in a space P_{t+1} suitably restricted. We extend Debreu's [17] notion of a valuation functional to a valuation operator to allow for portfolio prices to depend on date t information. With this in mind, let $\mathcal{F}_t$ be the sigma algebra used to depict the information available to all investors at date t . We use a representation of $\pi_{t,t+1}$ of the following form:

$$\pi_{t,t+1}(p_{t+1}) = E(s_{t+1}p_{t+1}|\mathcal{F}_t) \quad (1)$$

for some positive random variable s_{t+1} with probability one and any admissible payoff $p_{t+1} \in P_{t+1}$ on a portfolio of assets.

Definition 1 *The positive random variable s_{t+1} is the pricing kernel of the valuation operator $\pi_{t,t+1}$, provided that representation (1) is valid.*

If a kernel representation exists, the valuation operator $\pi_{t,t+1}$ necessarily satisfies the principle of no arbitrage.

How does one construct a representation of this type? There are alternative ways to achieve this. First, suppose that (i) P_{t+1} consists of all random variables that are $\mathcal{F}_{t+1}$ measurable and have finite conditional (on $\mathcal{F}_t$) second moments, (ii) $\pi_{t,t+1}$ is linear conditioned on $\mathcal{F}_t$, (iii) $\pi_{t,t+1}$ also satisfies a conditional continuity restriction, and (iv) the principle of no arbitrage is satisfied. The existence of a kernel s_{t+1} follows from a conditional version of the Riesz representation theorem [30]. Second, suppose that (i) P_{t+1} consists of all bounded random variables that are $\mathcal{F}_{t+1}$ measurable, (ii) $\pi_{t,t+1}$ is linear conditioned on $\mathcal{F}_t$, (iii) $E(\pi_{t,t+1})$ induces an absolutely continuous measure on $\mathcal{F}_{t+1}$, and (iv) the principle of no arbitrage is satisfied. We can apply the Radon–Nikodym theorem to justify the existence of a kernel. In both these cases, security markets are *complete* in that payoffs of any indicator function for events in $\mathcal{F}_{t+1}$ are included in the domain of the operator.^a Kernels can be constructed for other mathematical restrictions on asset payoffs and prices as well.

As featured by Harrison and Kreps [34] and Hansen and Richard [30], suppose that P_t is not so richly specified. Under the (conditional) Hilbert space formulation and appropriate restrictions on P_t , we may still apply a (conditional) version of the Riesz

representation theorem to represent

$$\pi_{t,t+1}(p_{t+1}) = E(q_{t+1} p_{t+1} | \mathcal{F}_t) \quad (2)$$

for some $q_{t+1} \in P_{t+1}$ and all $p_{t+1} \in P_{t+1}$. Even if the principle of no arbitrage is satisfied, there is no guarantee that the resulting q_{t+1} is positive with probability 1, however. Provided, however, that we can extend P_{t+1} to a larger space that includes indicator functions for events in $\mathcal{F}_{t+1}$ while preserving the principle of no arbitrage, when there exists a pricing kernel s_{t+1} for $\pi_{t,t+1}$. However, the pricing kernel may not be unique.

Stochastic Discounting

The random variable s_{t+1} is also called the *one-period stochastic discount factor*. Its multiperiod counterpart is as follows.

Definition 2 The stochastic discount process $\{S_{t+1} : t = 1, 2, \dots\}$ is

$$S_{t+1} = \prod_{j=1}^{t+1} s_j \quad (3)$$

where s_j is the pricing kernel used to represent the valuation operator between dates $j-1$ and j .

Thus, the $t+1$ period SDF compounds the corresponding one-period SDFs. The compounding is justified by the law of iterated values when trading is allowed at intermediate dates. As a consequence,

$$\pi_{0,t+1}(p_{t+1}) = E(S_{t+1} p_{t+1} | \mathcal{F}_0) \quad (4)$$

gives the date zero price of a security that pays p_{t+1} in the numeraire at date $t+1$. An SDF discounts the future in a manner that depends on future outcomes. This outcome dependence is included in order to make adjustments for risk. However, such adjustments are unnecessary for a discount bond because such a bond has a payoff that is equal to one independent of the state realized at date $t+1$. The price of a date $t+1$ discount bond is obtained from formula (4) by letting $p_{t+1} = 1$ and hence is given by $E[S_{t+1} | \mathcal{F}_0]$.

More generally, the date τ price of p_{t+1} is given as

$$\pi_{\tau,t+1}(p_{t+1}) = E\left[\left(\frac{S_{t+1}}{S_\tau}\right) p_{t+1} | \mathcal{F}_\tau\right] \quad (5)$$

for $\tau \leq t$. Thus, the ratio $\frac{S_{t+1}}{S_\tau}$ is the pricing kernel for the valuation operator $\pi_{\tau,t+1}$.

Risk-neutral Probabilities

Given a one-period pricing kernel, we build the so-called *risk-neutral probability measure* recursively as follows. First, rewrite the one-period pricing operator as

$$\pi_{t,t+1}(p_{t+1}) = E(s_{t+1} | \mathcal{F}_t) \tilde{E}(p_{t+1} | \mathcal{F}_t) \quad (6)$$

where

$$\tilde{E}(p_{t+1} | \mathcal{F}_t) = E\left[\left(\frac{s_{t+1}}{E(s_{t+1} | \mathcal{F}_t)}\right) p_{t+1} | \mathcal{F}_t\right] \quad (7)$$

Thus, one-period pricing is conveniently summarized by a one-period discount factor for a riskless payoff, $E(s_{t+1} | \mathcal{F}_t)$, and a change of the conditional probability measure implied by the conditional expectation operator $\tilde{E}(\cdot | \mathcal{F}_t)$. Risk adjustments are now absorbed into the change of probability measure.

Using this construction repeatedly, we decompose the date $t+1$ component of the SDF process:

$$S_{t+1} = \bar{S}_{t+1} M_{t+1} \quad (8)$$

where

$$\bar{S}_{t+1} = \left[\prod_{j=1}^{t+1} E(s_j | \mathcal{F}_{j-1}) \right] \quad (9)$$

is the discount factor constructed from the sequence of one-period riskless discount factors and

$$M_{t+1} = \prod_{j=1}^{t+1} \frac{s_j}{E(s_j | \mathcal{F}_{j-1})} = \frac{S_{t+1}}{\bar{S}_{t+1}} \quad (10)$$

is the positive martingale adapted to $\{\mathcal{F}_t : t = 0, 1, \dots\}$ with expectation unity. For each date t , M_t can be used to assign probabilities to events in $\mathcal{F}_t$. Since $E(M_{t+1} | \mathcal{F}_t) = M_t$, the date $t+1$ assignment of probabilities to events in $\mathcal{F}_{t+1}$ is compatible with the date t assignment to events in $\mathcal{F}_t \subset \mathcal{F}_{t+1}$. This follows from the law of iterated expectations. In effect, the law of iterated expectations enforces the law of iterated values.

Applying factorization (8), we have an alternative way to represent $\pi_{\tau,t+1}$:

$$\pi_{\tau,t+1}(p_{t+1}) = \tilde{E}[\bar{S}_{t+1} p_{t+1} | \mathcal{F}_\tau] \quad (11)$$

Pricing is reduced to riskless discounting and a distorted (risk-neutral) conditional expectation. The insight provided in [34, 49] is that dynamic asset pricing in the absence of arbitrage is captured by the existence of a positive (with probability one) martingale that can be used to represent prices in conjunction with a sequence of one-period riskless discount factors. Moreover, it justifies an arguably mild modification of the “efficient market hypothesis” that states that discounted prices should behave as martingales with the appropriate cash-flow adjustment. While Rubinstein [50] and Lucas [42] had clearly shown that efficiency should not preclude risk compensation, the notion of equivalent martingale measures reconciles the points of view under greater generality. The martingale property and associated “risk-neutral pricing” are recovered for some distortion of the historical probability measure that encapsulates risk compensation. This distortion preserves “equivalence” (the two probability measures agree about which events in $\mathcal{F}_t$ are assigned probability zero measure for each finite t) by ensuring the existence of a strictly positive SDF.

The concept of equivalent martingale measure (for each t) has been tremendously influential in derivative asset pricing. The existence of such a measure allows risk-neutral pricing of all contingent claims that are attainable because their payoff can be perfectly duplicated by self-financing strategies. Basic results from probability theory are directly exploitable in characterizing asset pricing in an arbitrage-free environment. More details can be found in **Econometrics of Option Pricing**.

Although we have developed this discussion for a discrete-time environment, there are well-known continuous-time extensions. In the case of continuous-time Brownian motion information structures, the change of measure has a particularly simple structure. In accordance to the Girsanov theorem, the Brownian motion under the original probability measure becomes a Brownian motion with drift under the risk-neutral measure. These drifts absorb the adjustments for exposure to Brownian motion risk.

Investors’ Preferences

Economic models show how exposure to risk is priced and what determines riskless rates of interest. As featured by Ross [49], the risk exposure that

is priced is the risk that cannot be diversified by averaging through the construction of portfolios of traded securities. Typical examples include macroeconomic shocks. Empirical macroeconomics seeks to identify macroeconomic shocks to quantify responses of macroeconomic aggregates to those shocks. Asset pricing models assign prices to exposure of cash flows to the identified shocks. The risk prices are encoded in SDF processes and hence are implicit in the risk-neutral probability measures used in financial engineering.

SDFs implied by specific economic models often reflect investor preferences. Subjective rates of discount and intertemporal elasticities appear in formulas for risk-free interest rates, and investors’ aversion to risk appears in formulas for the prices assigned to alternative risk exposures. Sometimes the reflection of investor preferences is direct and sometimes it is altered by the presence of market frictions. In what follows, we illustrate briefly some of the SDFs that have been derived in the literature.

Power Utility

When there are no market frictions and markets are complete, investors’ preferences can be subsumed into a utility function of a representative agent. In what follows, suppose that the representative investor has discounted time-separable utility with a constant elasticity of substitution (CES) $1/\rho$:

$$\frac{S_{t+1}}{S_t} = \exp(-\delta) \left(\frac{C_{t+1}}{C_t} \right)^{-\rho} \quad (12)$$

This is the SDF for the power utility specification of the model of Rubinstein [50] and Lucas [42].^b

Recursive Utility

Following Epstein and Zin [18], Kreps and Porteus [41] and Weil [53], preferences are specified recursively using continuation values for hypothetical consumption processes. Using a double CES recursion, the resulting SDF is

$$\frac{S_{t+1}}{S_t} = \exp(-\delta) \left(\frac{C_{t+1}}{C_t} \right)^{-\rho} \left[\frac{U_{t+1}}{R_t(U_{t+1})} \right]^{\rho-\gamma} \quad (13)$$

where $\gamma > 0$, $\rho > 0$, U_t is the continuation value associated with current and future consumption, and

$$R_t(U_{t+1}) \doteq (E[(U_{t+1})^{1-\gamma} | \mathcal{F}_t])^{1/(1-\gamma)} \quad (14)$$

is the risk-adjusted continuation value. The parameter ρ continues to govern the elasticity of substitution, whereas the parameter γ alters the risk preferences. See [25, 31, 52] for a discussion of the use of continuation values in representing the SDF and see [11, 25] for discussions of empirical implications. When $\rho = 1$, the recursive utility model coincides with a model in which investors have a concern about robustness as in [6].

The recursive utility model is just one of a variety of ways of altering investors' preferences. For instance, Constantinides [15] and Heaton [36] explore implications of introducing intertemporal complementarities in the form of "habit persistence".

Consumption Externalities and Reference Utility

Abel [1], Campbell and Cochrane [12] and Menzly *et al.* [45] develop asset pricing implications of models in which there are consumption externalities in models with stochastic consumption growth. These externalities can depend on a social stock of past consumptions. The implied one-period SDF in these models has the following form:

$$\frac{S_{t+1}}{S_t} = \exp(-\delta) \left(\frac{C_{t+1}}{C_t} \right)^{-\rho} \frac{\phi(H_t)}{\phi(H_0)} \quad (15)$$

The social stock of consumption is built as a possibly nonlinear function of current and past social consumption or innovations to social consumption. The construction of this stock differs across the various specifications. The process $\{H_t : t = 0, 1, \dots\}$ is the ratio of the consumption to the social stock and ϕ is an appropriately specified function of this ratio. In a related approach, Garcia *et al.* [21] consider investors' preferences in consumption as being evaluated relative to a reference level that is determined externally. The SDF is

$$\frac{S_{t+1}}{S_t} = \exp(-\delta) \left(\frac{C_{t+1}}{C_t} \right)^{-\rho} \left(\frac{H_t}{H_0} \right)^\eta \quad (16)$$

where the process $\{H_t : t = 0, 1, \dots\}$ is the ratio of the consumption to the reference level. In effect, the social externality in these models induces a preference shock or wedge to the SDF specification, which explicitly depends on the equilibrium aggregate consumption process. Finally, Bakshi and Chen [8] develop implications for a model in which relative wealth enters the utility function of investors.

Incomplete Markets

Suppose that individuals face private shocks that they cannot insure against but that they can write complete contracts over aggregate shocks. Let $\mathcal{F}_{t+1}$ denote the date $t + 1$ sigma algebra generated by aggregate shocks available up until date $t + 1$. This conditioning information set determines the type of risk exposure that can be traded as of date $t + 1$. Under this partial risk sharing and power utility, the SDF for pricing aggregate uncertainty is

$$\begin{aligned} \frac{S_{t+1}}{S_t} &= \exp(-\delta) \left(\frac{E[(C_{t+1}^j)^{-\rho} | \mathcal{F}_{t+1}]}{E[(C_t^j)^{-\rho} | \mathcal{F}_t]} \right) \\ &= \exp(-\delta) \left(\frac{C_{t+1}^a}{C_t^a} \right)^{-\rho} \\ &\quad \times \left(\frac{E[(C_{t+1}^j / C_{t+1}^a)^{-\rho} | \mathcal{F}_{t+1}]}{E[(C_t^j / C_t^a)^{-\rho} | \mathcal{F}_t]} \right) \end{aligned} \quad (17)$$

In this example and those that follow, C_{t+1}^j is consumption for individual j and C_{t+1}^a is aggregate consumption. Characterization (17) of the SDF displays the pricing implications of limited risk sharing in security markets. It is satisfied, for instance, in the model of Constantinides and Duffie [16].

Private Information

Suppose that individuals have private information about labor productivity that is conditionally independent, given aggregate information, leisure enters preferences in a manner that is additively separable and consumption allocations are Pareto optimal given the private information. As shown by Kocherlakota and Pistaferri [40], the SDF follows from the "inverse Euler equation" of [40, 47],

$$\begin{aligned} \frac{S_{t+1}}{S_t} &= \exp(-\delta) \left(\frac{E[(C_t^j)^\rho | \mathcal{F}_t]}{E[(C_{t+1}^j)^\rho | \mathcal{F}_{t+1}]} \right) \\ &= \exp(-\delta) \left(\frac{C_{t+1}^a}{C_t^a} \right)^{-\rho} \\ &\quad \times \left(\frac{E[(C_t^j / C_t^a)^\rho | \mathcal{F}_t]}{E[(C_{t+1}^j / C_{t+1}^a)^\rho | \mathcal{F}_{t+1}]} \right) \end{aligned} \quad (18)$$

where $\mathcal{F}_t$ is generated by the public information. As emphasized by Rogerson [47], this is a model with a form of “savings constraints”. While the stochastic discount factor for the incomplete information model is expressed in terms of the $(-\rho)^{\text{th}}$ moments of the cross-sectional distributions of consumption in adjacent time periods Kocherlakota and Pistaferri [40] show that in the private information model it is the ρ th moments of these same distributions.

Solvency Constraints

Luttmer [43], He and Modest [35], and Cochrane and Hansen [14] study asset pricing implications in models with limits imposed on the state contingent debt that is allowed. Alvarez and Jermann [4] motivate such constraints by appealing to limited commitment as in [38, 39]. When investors default, they are punished by excluding participation in asset markets in the future. Chien and Lustig [13] explore the consequences of alternative (out of equilibrium) punishments. Following [43], the SDF in the presence of solvency constraints and power utility is

$$\frac{S_{t+1}}{S_t} = \exp(-\delta) \left(\min_j \frac{C_{t+1}^j}{C_t^j} \right)^{-\rho} \quad (19)$$

and in particular,

$$\frac{S_{t+1}}{S_t} \geq \exp(-\delta) \left(\frac{C_{t+1}^a}{C_t^a} \right)^{-\rho} \quad (20)$$

Thus, the consumer with the smallest realized growth rate in consumption has a zero Lagrange multiplier on his or her solvency constraint, and hence the intertemporal marginal rate of substitution for this person is equal to the SDF. For the other consumers, the binding constraint prevents them shifting consumption from the future to the current period.

Long-term Risk

The SDF process assigns prices to risk exposures at alternative investment horizons. To study pricing over these horizons, Alvarez and Jermann [5], Hansen and Scheinkman [32], and Hansen [24] use a Markov structure and apply factorizations of the following form:

$$S_{t+1} = \exp(-\eta t) M_{t+1} \frac{\hat{f}(X_0)}{\hat{f}(X_{t+1})} \quad (21)$$

where $\{M_t : t = 0, 1, \dots\}$ is a *multiplicative martingale*, η is a positive number, $\{X_t : t = 0, 1, \dots\}$ is an underlying Markov process, and $\hat{f}$ is a positive function of the Markov state. The martingale is used as a convenient change of probability, one that is distinct from the “risk-neutral” measure described previously. Using this change of measure, asset prices can be depicted as

$$\pi_{0,t+1}(p_{t+1}) = \exp[-\eta(t+1)] \times \hat{E} \left[\frac{p_{t+1}}{\hat{f}(X_{t+1})} | X_0 \right] \hat{f}(X_0) \quad (22)$$

where $\hat{E}$ is the conditional expectation that is built with the martingale $\{M_t : t \geq 0\}$ in (21). The additional discounting is now constant and simple expectations can now be computed by exploiting the Markov structure. Hansen and Scheinkman [32] give necessary and sufficient conditions for the martingale in factorization (21) to imply a change of probability measure with stable stochastic dynamics.^c While Alvarez and Jermann [5] use this factorization to investigate the long-term links between the bond market and the macroeconomy, Hansen and Scheinkman [32] and Hansen [24] extend this factorization to study the valuation of cash flows that grow stochastically over time. As argued by Hansen and Scheinkman [32] and Hansen [24], these more general factorizations are valuable for the study of risk-return trade-offs for long investment horizons.

As argued by Bansal and Lehmann [9], many alterations to the power utility model in the section Power Utility can be represented as

$$\frac{S_{t+1}^*}{S_t^*} = \left(\frac{S_{t+1}}{S_t} \right) \left[\frac{f(X_{t+1})}{f(X_t)} \right] \quad (23)$$

for a positive function f . Transient components in asset pricing models are included to produce short-term alterations in asset prices and are expressed as the ratio of a function of the Markov state in adjacent dates. As shown by Bansal and Lehmann [9], this representation arises in models with habit persistence, or as shown in [24] the same is true for a limiting version of the recursive utility model. Combining equation (23) with equation (21) gives

$$S_{t+1}^* = \exp(-\eta) M_{t+1} \frac{f^*(X_0)}{f^*(X_{t+1})} \quad (24)$$

where $f^* = \hat{f}/f$.

Inferring Stochastic Discount Factors from Data

Typically a finite number of asset payoffs are used in econometric practice. In addition, the information used in an econometric investigation may be less than that used by investors. With this in mind, let Y_{t+1} denote an n -dimensional vector of asset payoffs observed by the econometrician such that

$$E(|Y_{t+1}|^2|\mathcal{G}_t) < \infty \quad (25)$$

with $E(Y_{t+1}Y_{t+1}'|\mathcal{G}_t)$ nonsingular with probability 1 and $\mathcal{G}_t \subset \mathcal{F}_t$. Let Q_t denote the corresponding price vector that is measurable with respect to $\mathcal{G}_t$, implying that

$$E(s_{t+1}Y_{t+1}|\mathcal{G}_t) = Q_t \quad (26)$$

where $s_{t+1} = S_{t+1}/S_t$. We may construct a counterpart to a (one-period) SDF by forming

$$p_{t+1}^* = Y_{t+1}' [E(Y_{t+1}Y_{t+1}'|\mathcal{G}_t)]^{-1} Q_t \quad (27)$$

Note that

$$E(p_{t+1}^* Y_{t+1} | \mathcal{G}_t) = Q_t \quad (28)$$

suggesting that we could just replace s_{t+1} in equation (26) with p_{t+1}^* . We refer to p_{t+1}^* as a *counterpart* to an SDF because we have not restricted p_{t+1}^* to be positive.

This construction is a special case of the representation of the prices implied by a conditional version of the Riesz representation theorem [30]. Since p_{t+1}^* is not guaranteed to be positive, if we used it to assign prices to derivative claims (nonlinear functions of Y_{t+1}), we might induce arbitrage opportunities. Nevertheless, provided that s_{t+1} has a finite conditional second moment,

$$E[(s_{t+1} - p_{t+1}^*) Y_{t+1} | \mathcal{G}_t] = 0 \quad (29)$$

This orthogonality indicates that p_{t+1}^* is the conditional least-squares projection of s_{t+1} onto Y_{t+1} . Although a limited set of asset price data will not reveal s_{t+1} , the data can provide information about the date $t+1$ kernel for pricing over a unit time interval.

Suppose that Y_{t+1} contains a conditionally riskless payoff. Then

$$E(s_{t+1}|\mathcal{G}_t) = E(p_{t+1}^*|\mathcal{G}_t) \quad (30)$$

By a standard least-squares argument, the conditional volatility of s_{t+1} must be at least as large as the conditional volatility of p_{t+1}^* . There are a variety of other restrictions that can be derived. For instance, see [9, 28, 51].^d

This construction has a direct extension to the case in which a complete set of contracts can be written over the derivative claims. Let H_{t+1} be the set of all payoffs that have finite second moments conditioned on $\mathcal{G}_t$ and are of the form $h_{t+1} = \phi(Y_{t+1})$ for some Borel measurable function ϕ . Then, we may obtain a kernel representation for pricing claims with payoffs in H_{t+1} by applying the Riesz representation theorem:

$$\pi_t(h_{t+1}) = E(h_{t+1}^* h_{t+1} | \mathcal{G}_t) \quad (31)$$

for some h_{t+1}^* in H_{t+1} . Then

$$E[(s_{t+1} - h_{t+1}^*) h_{t+1} | \mathcal{G}_t] = 0 \quad (32)$$

which shows that the conditional least-squares projection of s_{t+1} onto H_{t+1} is h_{t+1}^* . In this case, h_{t+1}^* will be positive. Thus, this richer collection of observed tradable assets implies a more refined characterization of s_{t+1} including the possibility that $s_{t+1} = h_{t+1}^*$, provided H_{t+1} coincides with the full collection of one-period payoffs that could be traded by investors. The estimation procedures of [3, 48] can be interpreted as estimating h_{t+1}^* projected on the return information.

Linear Asset Pricing Models

The long history of linear beta pricing can be fruitfully revisited within the SDF framework. Suppose that

$$Q_t = E[(\lambda_t \cdot z_{t+1} + \alpha_t) Y_{t+1} | \mathcal{G}_t] \quad (33)$$

for some vector λ_t and some scalar α_t that are $\mathcal{G}_t$ measurable. Then

$$E[Y_{t+1} | \mathcal{G}_t] = \alpha_t^* Q_t + \lambda_t^* \cdot \text{cov}(Y_{t+1}, z_{t+1} | \mathcal{G}_t) \quad (34)$$

where

$$\begin{aligned} \alpha_t^* &= \frac{1}{\alpha_t + \lambda_t \cdot E(z_{t+1} | \mathcal{G}_t)} \\ \lambda_t^* &= -\lambda_t \alpha_t^* \end{aligned} \quad (35)$$

which produces the familiar result that risk compensation expressed in terms of the conditional mean discrepancy: $E[Y_{t+1}|\mathcal{G}_t] - \alpha_t^* Q_t$ depends on the conditional covariances with the factors. When the entries of z_{t+1} are standardized to have conditional variances equal to unity, the entries of λ_t^* become the conditional regression coefficients, the “betas” of the asset payoffs onto the alternative observable factors. When z_{t+1} is the scalar market return, this specification gives the familiar CAPM from empirical finance. When z_{t+1} is augmented to include the payoff on a portfolio designed to capture risk associated with size (market capitalization) and the payoff on a portfolio designed to capture risk associated with book to market equity, this linear specification gives a conditional version of the [19] three-factor model designed to explain cross-sectional differences in expected returns (see **Factor Models**).

If the factors are among the asset payoffs and a conditional (on $\mathcal{G}_t$) riskless payoff is included in Y_{t+1} , then

$$p_{t+1}^* = \lambda_t \cdot z_{t+1} + \alpha_t \quad (36)$$

and thus $\lambda_t \cdot z_{t+1} + \alpha_t$ is the conditional least-squares regression of s_{t+1} onto Y_{t+1} .

For estimation and inference, consider the special case in which $\mathcal{G}_t$ is degenerate and the vector Y_{t+1} consists of excess returns (Q_t is a vector of zeros). Suppose that the data generation process for $\{(z_{t+1}, Y_{t+1})\}$ is independent and identically distributed (i.i.d.) and multivariate normal. Then α_t and λ_t are time invariant. The coefficient vectors can be efficiently estimated by maximum likelihood and the pricing restrictions tested by likelihood ratio statistics [22]. As MacKinlay and Richardson [44] point out, it is important to relax the i.i.d. normal assumption in many applications. In contrast with the parametric maximum likelihood approach, generalized method of moments (GMM) provides an econometric framework that allows conditional heteroskedasticity and temporal dependence [23]. Moreover, the GMM approach offers the important advantage to provide a unified framework for the testing of conditional linear beta pricing models. See [37] for an extensive discussion of the application of GMM to linear factor models. We will have more to say about estimation when $\mathcal{G}_t$ is not degenerate in our subsequent discussion.

Suppose that the conditional linear pricing model is misspecified. Hansen and Jagannathan [29] show

that choosing (λ_t, α_t) to minimize the maximum pricing error of payoffs with $E(|p_{t+1}|^2|\mathcal{G}_t) = 1$ is equivalent to solving the least-squares problem:^e

$$\begin{aligned} \min_{\lambda_t, \alpha_t, v_{t+1}, E((v_{t+1})^2|\mathcal{G}_t) < \infty} & E[(\lambda_t \cdot z_{t+1} + \alpha_t - v_{t+1})^2|\mathcal{G}_t] \\ \text{subject to } & Q_t = E(v_{t+1} Y_{t+1}|\mathcal{G}_t) \end{aligned} \quad (37)$$

This latter problem finds a random variable v_{t+1} that is close to $\lambda_t \cdot z_{t+1} + \alpha_t$, allowing for departures from pricing formula (33) where v_{t+1} is required to represent prices correctly. For a fixed (λ_t, α_t) , the “concentrated” objective is

$$\begin{aligned} & [E[\lambda_t \cdot z_{t+1} + \alpha_t] Y_{t+1}|\mathcal{G}_t] - Q_t]' \\ & \times [E(Y_{t+1} Y_{t+1}'|\mathcal{G}_t)]^{-1} \\ & \times [E[\lambda_t \cdot z_{t+1} + \alpha_t] Y_{t+1}|\mathcal{G}_t] - Q_t] \end{aligned} \quad (38)$$

which is a quadratic form of the vector of pricing error. The random vector (λ_t, α_t) is chosen to minimize this pricing error. In the case of a correct specification, this minimization results in a zero objective; otherwise, it provides a measure of model misspecification.

Estimating Parametric Models

In examples of SDFs like those given in the section Investors’ Preferences, there are typically unknown parameters to estimate. This parametric structure often permits the identification of the SDF even with limited data on security market payoffs and prices. To explore this approach, we introduce a parameter θ that governs say investors’ preferences. It is worth noting that inference about θ is a semiparametric statistical problem since we avoid the specification of the law of motion of asset payoffs and prices along with the determinants of SDFs. In what follows, we sketch the main statistical issues in the following simplified context.

Suppose that $s_{t+1} = \phi_{t+1}(\theta_o)$ is a parameterized model of an SDF that satisfies

$$E[\phi_{t+1}(\theta_o) Y_{t+1}|\mathcal{G}_t] - Q_t = 0 \quad (39)$$

where ϕ_{t+1} can depend on observed data but the parameter vector θ_o is unknown. This conditional

moment restriction implies a corresponding unconditional moment condition:

$$E[\phi_{t+1}(\theta_o)Y_{t+1} - Q_t] = 0 \quad (40)$$

As emphasized by Hansen and Richard [30], conditioning information can be brought in through the back door by scaling payoffs and their corresponding prices by random variables that are $\mathcal{G}_t$ measurable. Hansen and Singleton [33] describe how to use the unconditional moment condition to construct a GMM estimator of the parameter vector θ_o with properties characterized by Hansen [23] (see also **Generalized Method of Moments (GMM)**).

As an alternative, the parameterized family of models might be misspecified. Then the approach of [29] described previously could be used, whereby an estimator of θ_o is obtained by minimizing

$$\left[\frac{1}{N} \sum_{t=1}^N \phi_{t+1}(\theta) Y_{t+1} - Q_t \right] \left[\frac{1}{N} \sum_{t=1}^N Y_{t+1} Y_{t+1}' \right]^{-1} \times \left[\frac{1}{N} \sum_{t=1}^N \phi_{t+1}(\theta) Y_{t+1} - Q_t \right] \quad (41)$$

with respect to θ . This differs from a GMM formulation because the weighting matrix in the quadratic form does not depend on the SDF, but only on the second moment matrix of the payoff vector Y_{t+1} . This approach suffers from a loss of statistical efficiency in estimation when the model is correctly specified, but it facilitates comparisons across models because the choice of an SDF does not alter how the overall magnitude of the pricing errors is measured.

As in [26, 29], the analysis can be modified to incorporate the restriction that pricing errors should also be small for payoffs on derivative claims. This can be formalized by computing the time series approximation to the least-squares distance between a candidate, but perhaps misspecified, SDF from the family of strictly positive random variables z_{t+1} that solve the pricing restriction:

$$E(z_{t+1}Y_{t+1} - Q_t) = 0 \quad (42)$$

See [25] for more discussion of these alternative approaches for estimation.

So far we have described econometric methods that reduce conditional moment restrictions (33) and (40)

to their unconditional counterparts. From a statistical perspective, it is more challenging to work with the original conditional moment restriction. Along this vein, Ai and Chen [2] and Antoine *et al.* [7] develop nonparametric methods to estimate conditional moment restrictions used to represent pricing restrictions. In particular, Antoine *et al.* [7] develop a conditional counterpart to the continuously updated GMM estimator introduced by Hansen *et al.* [27] in which the weighting matrix in GMM objective explicitly depends on the unknown parameter to be estimated.^f These same methods can be adapted to explicitly take account of conditioning information while allowing for misspecification as in [29].

Acknowledgment

Jarda Borovicka provided suggestions that were very helpful for this article.

End Notes

^aSee [46] for a dynamic extension using this approach.

^bSee [10] for a continuous-time counterpart.

^cAlthough Hansen and Scheinkman [32] use a continuous-time formulation, discrete-time counterparts to their analysis are straightforward to obtain.

^dThese authors do not explicitly use conditioning information. In contrast, Gallant *et al.* [20] estimate conditional moment restrictions using a flexible parameterization for the dynamic evolution of the data.

^eHansen and Jagannathan [29] abstract from conditioning information, but what follows is a straightforward extension.

^fThe continuously updated estimator is also closely related to the “minimum entropy” and “empirical likelihood” approaches (see **Entropy-based Estimation**).

References

- [1] Abel, A.B. (1990). Asset prices under habit formation and catching up with the Joneses, *American Economic Review* **80**(2), 38–42.
- [2] Ai, C. & Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions, *Econometrica* **71**(6), 1795–1843.
- [3] Ait-Sahalia, Y. & Lo, A.W. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices, *Journal of Finance* **53**(2), 499–547.
- [4] Alvarez, F. & Jermann, U.J. (2000). Efficiency, equilibrium, and asset pricing with risk of default, *Econometrica* **68**(4), 775–798.

- [5] Alvarez, F. & Jermann, U.J. (2005). Using asset prices to measure the persistence of the marginal utility of wealth, *Econometrica* **73**(6), 1977–2016.
- [6] Anderson, E.W., Hansen, L.P. & Sargent, T.J. (2003). A quartet of semigroups for model specification, robustness, prices of risk, and model detection, *Journal of the European Economic Association* **1**, 69–123.
- [7] Antoine, B., Bonnal, H. & Renault, E. (2007). On the efficient use of the informational content of estimating equations: implied probabilities and euclidean empirical likelihood, *Journal of Econometrics* **138**(2), 461–487.
- [8] Bakshi, G.S. & Chen, Z. (1996). The spirit of capitalism and stock-market prices, *American Economic Review* **86**(1), 133–157.
- [9] Bansal, R. & Lehmann, B.N. (1997). Growth-optimal portfolio restrictions on asset pricing models, *Macroeconomic Dynamics* **1**(2), 333–354.
- [10] Breeden, D.T. (1979). An intertemporal asset pricing model with stochastic consumption and investment opportunities, *Journal of Financial Economics* **7**(3), 265–296.
- [11] Campbell, J.Y. (2003). Consumption-based asset pricing, in G.M. Constantinides, M. Harris & R.M. Stulz, eds, *Handbook of the Economics of Finance*, Elsevier, Chapter 13, Vol. 1, pp. 803–887.
- [12] Campbell, J.Y. & Cochrane, J.H. (1999). By force of habit, *Journal of Political Economy* **107**(2), 205–251.
- [13] Chien, Y.L. & Lustig, H.N. (2008). The market price of aggregate risk and the wealth distribution, forthcoming in the *Review of Financial Studies*.
- [14] Cochrane, J.H. & Hansen, L.P. (1992). Asset pricing explorations for macroeconomics, *NBER Macroeconomics Annual* **7**, 115–165.
- [15] Constantinides, G.M. (1990). Habit formation: a resolution of the equity premium puzzle, *Journal of Political Economy* **98**(3), 519–543.
- [16] Constantinides, G.M. & Duffie, D. (1996). Asset pricing with heterogeneous consumers, *Journal of Political Economy* **104**(2), 219–240.
- [17] Debreu, G. (1954). Valuation equilibrium and Pareto optimum, *Proceedings of the National Academy of Sciences* **40**(7), 588–592.
- [18] Epstein, L. & Zin, S. (1989). Substitution, risk aversion and the temporal behavior of consumption and asset returns: A theoretical framework, *Econometrica* **57**(4), 937–969.
- [19] Fama, E.F. & French, K.R. (1992). The cross-section of expected stock returns, *Journal of Finance* **47**, 427–465.
- [20] Gallant, A.R., Hansen, L.P. & Tauchen, G. (1990). Using conditional moments of asset payoffs to infer the volatility of intertemporal marginal rates of substitution, *Journal of Econometrics* **45**(1–2), 141–179.
- [21] Garcia, R., Renault, E. & Semenov, A. (2006). Disentangling risk aversion and intertemporal substitution through a reference utility level, *Finance Research Letters* **3**, 181–193.
- [22] Gibbons, M.R., Ross, S.A. & Shanken, J. (1989). A test of the efficiency of a given portfolio, *Econometrica* **57**(5), 1121–1152.
- [23] Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators, *Econometrica* **50**(4), 1029–1054.
- [24] Hansen, L.P. (2008). Modeling the long run: valuation in dynamic stochastic economies, *Fisher-Schultz Lecture at the 2006 Meetings of the Econometric Society*, Vienna, Austria, 2006.
- [25] Hansen, L.P., Heaton, J. Lee, J. & Roussanov, N. (2007). *Handbook of Econometrics*, Intertemporal substitution and risk aversion, in J.J. Heckman & E.E. Leamer, eds, *Handbook of Econometrics*, Elsevier, Chapter 61, Vol. 6.
- [26] Hansen, L.P., Heaton, J. & Luttmer, E.G.J. (1995). Econometric evaluation of asset pricing models, *Review of Financial Studies* **8**(2), 237–274.
- [27] Hansen, L.P., Heaton, J. & Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators, *Journal of Business and Economic Statistics* **14**(3), 262–280.
- [28] Hansen, L.P. & Jagannathan, R. (1991). Implications of security market data for models of dynamic economies, *Journal of Political Economy* **99**(2), 225–262.
- [29] Hansen, L.P. & Jagannathan, R. (1997). Assessing specification errors in stochastic discount factor models, *Journal of Finance* **52**(2), 557–590.
- [30] Hansen, L.P. & Richard, S.F. (1987). The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models, *Econometrica* **50**(3), 587–614.
- [31] Hansen, L.P., Sargent, T.J. & Tallarini, T.D. Jr. (1999). Robust permanent income and pricing, *The Review of Economic Studies* **66**(4), 873–907.
- [32] Hansen, L.P. & Scheinkman, J.A. (2009). Long-term risk: an operator approach, *Econometrica* **77**(1), 177–234.
- [33] Hansen, L.P. & Singleton, K.J. (1982). Generalized instrumental variables estimation of nonlinear rational expectations models, *Econometrica* **50**(5), 1269–1286.
- [34] Harrison, J.M. & Kreps, D.M. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- [35] He, H. & Modest, D.M. (1995). Market frictions and consumption-based asset pricing, *Journal of Political Economy* **103**(1), 94–117.
- [36] Heaton, J. (1995). An empirical investigation of asset pricing with temporally dependent preference specifications, *Econometrica* **63**(3), 681–717.
- [37] Jagannathan, R., Soukakis, G. & Wang, Z. (2009). The analysis of the cross section of security returns, in forthcoming in the *Handbook of Financial Econometrics*, Y. Ait-Sahalia & L.P. Hansen, Elsevier.
- [38] Kehoe, T.J. & Levine, D.K. (1993). Debt-constrained asset markets, *Review of Economic Studies* **60**(4), 865–888.

-
- [39] Kocherlakota, N.R. (1996). Implications of efficient risk sharing without commitment, *Review of Economic Studies* **63**(4), 595–609.
- [40] Kocherlakota, N. & Pistaferri, L. (2009). Asset pricing implications of Pareto optimality with private information, *Journal of Political Economy* **117**(3), 555–590.
- [41] Kreps, D.M. & Porteus, E.L. (1978). Temporal resolution of uncertainty and dynamic choice, *Econometrica* **46**(1), 185–200.
- [42] Lucas, R.E. (1978). Asset prices in an exchange economy, *Econometrica* **46**(6), 1429–1445.
- [43] Luttmer, E.G.J. (1996). Asset pricing in economies with frictions, *Econometrica* **64**(6), 1439–1467.
- [44] MacKinlay, A.C. & Richardson, M.P. (1991). Using generalized method of moments to test mean-variance efficiency, *Journal of Finance* **46**, 511–527.
- [45] Menzly, L., Santos, T. & Veronesi, P. (2004). Understanding predictability, *Journal of Political Economy* **112**(1), 1–47.
- [46] Rogers, L.C.G. (1998). The origins of risk-neutral pricing and the Black-Scholes formula, in *Risk Management and Analysis*, C.O. Alexander, ed., Wiley, New York, Chapter 2, pp. 81–94.
- [47] Rogerson, W.P. (1985). The first-order approach to principal-agent problems, *Econometrica* **53**(6), 1357–1367.
- [48] Rosenberg, J.V. & Engle, R.F. (2002). Empirical pricing kernels, *Journal of Financial Economics* **64**(3), 341–372.
- [49] Ross, S.A. (1976). The arbitrage theory of capital asset pricing, *Journal of Economic Theory* **13**, 341–360.
- [50] Rubinstein, M. (1976). The valuation of uncertain income streams and the valuation of options, *Bell Journal* **7**, 407–425.
- [51] Snow, K. (1991). Diagnosing asset pricing using the distribution of asset returns, *Journal of Finance* **46**(3), 955–983.
- [52] Tallarini, T.D. (2000). Risk-sensitive real business cycles, *Journal of Monetary Economics* **45**(3), 507–532.
- [53] Weil, P. (1990). Nonexpected utility in macroeconomics, *The Quarterly Journal of Economics* **105**(1), 29–42.

Related Articles

Arrow–Debreu Prices; Complete Markets; Fixed Mix Strategy; Fundamental Theorem of Asset Pricing; Stochastic Discount Factors.

LARS PETER HANSEN & ERIC RENAULT

Entropy-based Estimation

Let P, μ be two probability measures on a probability space, with P absolutely continuous with respect to μ (see **Equivalence of Probability Measures**), and denote the density of P with respect to μ by $dP/d\mu > 0$. The *relative entropy* $D(P\|\mu)$, sometimes called the *Kullback–Leibler information* [9] or divergence in statistics, is defined as

$$\begin{aligned} D(P\|\mu) &= \int_{\Omega} \log\left(\frac{dP}{d\mu}\right) dP \\ &= \int_{\Omega} \frac{dP}{d\mu} \log\left(\frac{dP}{d\mu}\right) d\mu \end{aligned} \quad (1)$$

Jensen's inequality can be used to show $D(P\|\mu) \geq 0$ and $D(P\|\mu) = 0$ if and only if $P \equiv \mu$. $D(P\|\mu)$ is thus an (asymmetric) index of the discrepancy between P and μ , and its definition is grounded in information theory and statistical mechanics. When $P = (P_1, \dots, P_H)$ and $\mu = (\mu_1, \dots, \mu_H)$ are probability measures on a discrete set with H elements, $D(P\|\mu) = \sum_{h=1}^H P_h \log\left(\frac{P_h}{\mu_h}\right)$. In the special case where $\mu_h \equiv 1/H$ is the uniform distribution, $D(P\|\mu)$ is the *Shannon entropy* $-\sum_h P_h \log P_h$.

Estimation by Entropy Minimization under Constraints

Let x denote a (vector) of random variable(s), β a (vector) of parameter(s), and $f(x, \beta)$ a (column vector) of real-valued function(s). Let $E_P[f(x, \beta)]$ denote its expectation computed with probability measure P . For a specific value of β , define the set of probability measures:

$$\mathcal{P}(\beta) \equiv \{P : E_P[f(x, \beta)] = 0\} \quad (2)$$

which additionally are absolutely continuous with respect to a distinguished measure μ that is determined by the application. Selection of a particular probability measure in $\mathcal{P}(\beta)$ is an example of *linear inverse problem*. One way to select a solution in stable manner is by picking the probability measure,

which minimizes the relative entropy with respect to μ under constraints

$$\begin{aligned} &\min_{P \in \mathcal{P}(\beta)} D(P\|\mu) \\ &\equiv \min_P \int \log(dP/d\mu) dP \quad \text{s.t. } E_P[f(x, \beta)] \\ &= 0 \end{aligned} \quad (3)$$

When μ is the uniform distribution on a finite set with H elements, $\mu_h \equiv 1/H$ and the constrained minimization of $D(P\|\mu)$ is equivalent to maximization of the *Shannon entropy* $-\sum_h P_h \log P_h$.

The solution to equation (3) is well known ([10], sec.3(A)) to have the following *Gibbs Canonical* or *Esscher Transformed* density (see **Esscher Transform**)

$$\frac{dP(\beta)}{d\mu} = \frac{e^{\gamma(\beta)' f(x, \beta)}}{E_{\mu}[e^{\gamma(\beta)' f(x, \beta)}]} \quad (4)$$

To compute the coefficient vector $\gamma(\beta)$ in equation (4), solve the following problem

$$\gamma(\beta) = \arg \max_{\gamma} I \equiv -\log E_{\mu}[e^{\gamma' f(x, \beta)}] \quad (5)$$

where the maximized value of I in equation (5) turns out to be the minimal relative entropy $D(P(\beta)\|\mu)$ in equation (3). This constrained minimum of relative entropy also has a frequentist interpretation from the large deviations (see **Large Deviations**) theory of IID processes, which will prove useful in motivating the parameter estimation procedure described later.

Application to Derivative Security Valuation

Consider the simplest problem of option pricing: value a European call option, written on a single underlying stock that pays no dividends, whose price at expiration T -periods ahead is denoted $x(T)$. The riskless, continuously compounded gross interest rate r is constant between now and expiration. Under the familiar assumptions of complete and frictionless markets that do not admit arbitrage opportunities, there is a *risk-neutral* probability measure P under which the call option's price C is the expected

value of its risklessly discounted payoff at expiration, that is,

$$C = E_P[\max[x(T) - K, 0]/r^T] \quad (6)$$

where the risk-neutral probabilities P satisfy the “martingale” constraint $x(0) = E_P\left[\frac{x(T)}{r^T}\right]$, rewritten as

$$E_P\left[\frac{x(T)}{x(0)r^T} - 1\right] = 0 \quad (7)$$

Risk-neutral pricing of options proceeds by specifying a parametric model for the risk-neutral stochastic price process of the underlying stock. Parameter values are found that make the model’s computed stock prices and/or option prices close (e.g., in the least-squares sense) to observed stock and/or option prices [4]. In the simplest case (the Black–Scholes model), this procedure requires an estimate of the volatility parameter, found either from past stock returns (i.e., historical volatility) or from market option prices (i.e., a best-fitting implied volatility).

However, suppose one has doubts about the correct parametric model. The formalism of the previous section provides an alternative. Let the scalar function $f(x, \beta) = \frac{x(T)}{x(0)r^T} - 1$, where $\beta = r$. The distribution μ is the forecast distribution of $x(T)$. To estimate this, one could just use a histogram of past T -period stock returns as in [27, 29], a conditional histogram [28], or a more complex forecasting model [13]. The distribution is then substituted into equation (5) and solved to find the $\gamma(\beta)$ needed to estimate the density $P(\beta)$ in equation (4), which is required to compute the option valuation (6). It is possible to extend the approach to handle stochastic dividends and interest rates.

Another approach presumes a particular form for μ , and defines a vector $f(x, \beta)$ with i th component $\max[x(T) - K_i, 0]/r^T - C_i$, where C_i is the observed market price for a call with exercise price K_i , as in [5, 19]. Then, equations (4)–(5) are used to study the nature of the measure $P(\beta)$ implied by those options’ market prices, while equation (6) could be used to value options *other* than those present in f . See [3, 7, 8] for some theoretical

and applied extensions of this approach. (*see also Weighted Monte Carlo*).

Further refinements were developed by Gray *et al.* [16], showing that the associated dynamic hedge (i.e., entropic hedge ratio) outperformed hedging benchmarks, and by Alcock and Carmichael [1], extending the concept to enable valuation of American options.

The entropic approaches are not necessarily inconsistent with the conventional approach. In fact, extant closed-form option pricing models can be analytically derived by specification of the process generating the stock price distributions, and systematically applying the **Esscher transform** (4) as in [12, 14].

Application to Parameter Estimation

A research program initiated by Hansen [17] first developed the generalized methods of moments (GMM) parameter estimator. Two entropy-based alternatives are defined, motivated by alternative frequentist estimation criteria suggested by large deviations theory (*see Cramér’s Theorem; Large Deviations*). We begin with this alternative frequentist criterion, and finish by describing its connection to entropy. Interpret β in equation (2) to be an unknown vector of parameters from a set Θ of possible parameter vectors, and $f(x; \beta) = (f_1, \dots, f_r)$ to be an r -component column vector of observable, real-valued functions. A time series of the random variable x is denoted $x_1, \dots, x_T$. The corresponding (random) sample average of the vector f is denoted $f_T(\beta) \equiv \sum_{i=1}^T f(x_i; \beta)/T$.

The empirical asset pricing literature considers processes for which a typical realization produces $\lim_{T \rightarrow \infty} f_T(\beta) = E[f(x; \beta)]$, where the expectations operator is, unless otherwise subscripted, taken with respect to the invariant measure μ of the process. So the probability of observing $f_T(\beta)$ in a compact neighborhood that does *not* contain the population mean $E[f(x; \beta)]$ must approach zero as $T \rightarrow \infty$. The *asymptotic rate* at which this probability goes to zero can be calculated by use of part of a powerful result of Ellis [11], expounded in [6], (pp. 20–22), and perhaps first used in asset pricing by Stutzer [26], (Appendix). Define the following extended, real-valued function of an r -component

vector γ :

$$\begin{aligned}\phi_T(\gamma; \beta) &\equiv \frac{1}{T} \log E \left[e^{\gamma' \sum_{t=1}^T f(x_t, \beta)} \right] \\ &= \frac{1}{T} \log E \left[\prod_{t=1}^T e^{\gamma' f(x_t, \beta)} \right] \quad (8)\end{aligned}$$

Define the asymptotic limit of equation (8) to be

$$\phi(\gamma; \beta) \equiv \lim_{T \rightarrow \infty} \phi_T(\gamma; \beta) \quad (9)$$

and the nonnegative, probability decay rate function

$$I(c) \equiv \sup_{\gamma} [\gamma' c - \phi(\gamma; \beta)] \quad (10)$$

$$\arg \min_{\beta} \left[I(0) = \sup_{\gamma} - \lim_{T \rightarrow \infty} \left(\phi_T(\gamma; \beta) \equiv \frac{1}{T} \log E \left[e^{\gamma' \sum_{t=1}^T f(x_t, \beta)} \right] \right) \right] \quad (13)$$

Then the probability of observing a value of the sample average lying within a closed (open) set C (O) that does *not* contain the population mean $E[f(x; \beta)]$, *decays* toward zero at a positive asymptotic *rate* that must be weakly larger (smaller) than $\inf_{c \in C} I(c; \beta)$ ($\inf_{c \in O} I(c; \beta)$).

Hansen [17] GMM framework estimates parameters and tests theoretical implications represented by the constraints:

$$\exists ! \beta^* : E[f(x; \beta^*)] = 0 \quad (11)$$

where β^* is the unique parameter vector (to be estimated) that satisfies the moment constraints.

Define $\mathcal{O}_\epsilon$ to be the open ϵ -ball about $E[f(x; \beta^*)] = 0$. Because the identification condition (11) implies that $E[f(x; \beta)] \neq 0$ when $\beta \neq \beta^*$, the probability of finding $f_T(\beta) \in \mathcal{O}_\epsilon$ decays to zero at an asymptotic rate weakly lower than $\inf_{c \in \mathcal{O}_\epsilon} I(c)$. Because the limit of this as $\epsilon \downarrow 0$ is $I(0)$, which is 0 if and only if $\beta = \beta^*$, a sensible estimation criterion is to search for a value of β making it as small as possible, that is, find $\arg \min_{\beta} I(0)$ [15]. Kitamura and Stutzer [23] show that when the vector f is Gaussian with stationary covariance sequence $E[(f(x_t; \beta) - E[f(x_t; \beta)])(f(x_{t-\tau}; \beta) - E[f(x_{t-\tau}; \beta)])'] \equiv \Gamma_\tau(\beta)$, $j = -\infty, \dots, +\infty$, this estimation criterion is

identical to the optimally weighted GMM criterion:

$$\begin{aligned}\arg \min_{\beta} I(0) &= \frac{1}{2} E[f(x; \beta)]' \\ &\times \left(\sum_{\tau=-\infty}^{+\infty} \Gamma_\tau(\beta) \right)^{-1} E[f(x; \beta)] \quad (12)\end{aligned}$$

forming the foundation for different feasible estimators, corresponding to different ways of estimating the population moments and infinite sum present in equation (12), though the large deviation interpretation (*see Large Deviations*) of equation (12) no longer holds outside of Gaussianity.^a

What can be done in more realistic cases where the researcher does not want to commit to specific distributional assumptions? The general criterion is

so a feasible estimator of equation (13) is needed. Fortunately, Kitamura [20] developed a smoothing approach to estimation, which was used in ([22], equation (9)) to produce an analog estimator [24] of the minimizing β in equation (13) (for proof, see [23]). Under regularity conditions listed there, they showed that this estimator is asymptotically equivalent to feasible, optimally weighted GMM in non-IID cases.

Entropic Interpretation

When the process is IID (but not necessarily Gaussian), $\phi_T(\gamma; \beta) = \log E[e^{\gamma' f(x; \beta)}]$, so $I(0) \equiv \max_{\gamma} -\log E[e^{\gamma' f(x; \beta)}]$ and the large deviations estimation (*see Large Deviations*) criterion (13) simplifies to the *exponential tilting* criterion, that is, $\arg \min_{\beta} [I(0) = \max_{\gamma} -\log E[e^{\gamma' f(x; \beta)}]]$. From equation (5) and the comment immediately following it, we see that the optimized value of the criterion may also be interpreted as the moment-constrained minimum of relative entropy in equation (3), in the IID special case.

In the general case (13), the aforementioned analog estimator developed in ([22], equation (9)) also has an entropic interpretation. It is the estimator that minimizes the relative entropy between the empirical

distribution of smoothed observations and the distribution of smoothed observations under the moment restriction.

Alternatively, when one substitutes $D(\mu \parallel P)$ for $D(P \parallel \mu)$ in equation (3) with no change in the rest of the formalism, the result is the *empirical likelihood* criterion [25]. From the view point of statistical inference and estimation, the empirical likelihood criterion $D(\mu \parallel P)$ achieves various optimality properties, especially in terms of large deviations [21]. First-order Taylor approximation of $D(\mu \parallel P)$ results in the χ^2 distance (see [9], p. 333), so substituting it for $D(\mu \parallel P)$ yields the *Euclidean Empirical Likelihood* criterion [2, 18]. Of course, the large deviations interpretation is different, due to the interchange of P and μ in the formalism.

End Notes

^aAlso, the choice of criterion function impacts the finite sample behavior of feasible estimators even under Gaussianity.

References

- [1] Alcock, J. & Carmichael, T. (2008). Nonparametric American option pricing, *Journal of Futures Markets* **28**, 717–748.
- [2] Antoine, B., Bonnal, H. & Renault, E. (2007). On the efficient use of informational content of estimating equations: implied probabilities and Euclidean empirical likelihood, *Journal of Econometrics* **138**, 461–486.
- [3] Avellaneda, M. (1998). Minimum relative-entropy calibration of asset-pricing models, *Journal of Theoretical and Applied Finance* **1**, 447–472.
- [4] Bates, D.S. (1996). Testing option pricing models, in *Handbook of Statistics 15*, G.S. Maddala & C.R. Rao, eds, North-Holland.
- [5] Buchen, P.W. & Kelly, M. (1996). The maximum entropy distribution of an asset inferred from option prices, *Journal of Financial and Quantitative Analysis* **31**, 143–159.
- [6] Bucklew, J. (1990). *Large Deviation Techniques in Decision, Simulation, and Estimation*, Wiley.
- [7] Cont, R. & Tankov, P. (2004). Nonparametric calibration of jump-diffusion processes, *Journal of Computational Finance* **7**(3), 1–49.
- [8] Cont, R. & Tankov, P. (2006). Retrieving Levy processes from option prices: regularization of a nonlinear inverse problem, *SIAM Journal of Control and Optimization* **45**, 1–25.
- [9] Cover, T.M. & Thomas, J.A. (1991). *Elements of Information Theory*, Wiley.
- [10] Csiszar, I. (1975). I-divergence geometry of probability distributions and minimization problems, *Annals of Probability* **3**, 146–158.
- [11] Ellis, R. (1984). Large deviations for a general class of random vectors, *Annals of Probability* **12**, 1–12.
- [12] Eberlein, E., Keller, U. & Prause, K. (1998). New insights into smile, mispricing, and value at risk: The Hyperbolic model, *Journal of Business* **71**, 371–405.
- [13] Foster, F.D. & Whiteman, C.H. (1999). An application of Bayesian option pricing to the soybean market, *American Journal of Agricultural Economics* **81**, 222–272.
- [14] Gerber, H. & Shiu, E. (1994). Option pricing by Esscher Transforms, *Transactions of the Society of Actuaries* **46**, 99–140.
- [15] Glasserman, P. & Jin, Y. (2000). *Comparing Stochastic Discount Factors Through their Implied Measures*, mimeo, Graduate School of Business, Columbia University.
- [16] Gray, P., Edwards, S. & Kalotay, E. (2007). Canonical pricing and hedging of index options, *Journal of Futures Markets* **27**, 771–790.
- [17] Hansen, L. (1982). Large sample properties of generalized methods of moments estimators, *Econometrica* **50**, 1029–1054.
- [18] Hansen, L., Heaton, J. & Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators, *Journal of Business and Economic Statistics* **14**, 262–280.
- [19] Hawkins, R.J., Rubinstein, M. & Daniell, G.J. (1996). Reconstruction of the probability density function implicit in option prices from incomplete and noisy data, in *Maximum Entropy and Bayesian Methods*, K. Hanson & R. Silver, eds, Kluwer.
- [20] Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes, *Annals of Statistics* **25**, 2084–2102.
- [21] Kitamura, Y. (2007). Empirical likelihood methods in econometrics, in *Advances in Economics and Econometrics*, R. Blundell, W. Newey & T. Persson, eds, Cambridge University Press.
- [22] Kitamura, Y. & Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation, *Econometrica* **65**, 861–874.
- [23] Kitamura, Y. & Stutzer, M. (2002). Connections between entropic and linear projections in asset pricing estimation, *Journal of Econometrics* **107**, 159–174.
- [24] Manski, C. (1988). *Analog Estimation Methods in Economics*, Chapman and Hall, London.
- [25] Owen, A. (2001). Chapman and Hall, *Empirical Likelihood*.
- [26] Stutzer, M. (1995). A Bayesian approach to diagnosis of asset pricing models, *Journal of Econometrics* **68**, 367–397.

- [27] Stutzer, M. (1996). A simple nonparametric approach to derivative security valuation, *Journal of Finance* **51**, 1633–1652.
- [28] Stutzer, M. & Chowdhury, M. (1999). A simple nonparametric approach to bond futures option pricing, *Journal of Fixed Income* **8**, 67–76.
- [29] Zou, J. & Derman, E. (1999). *Strike Adjusted Spread: A New Metric for Estimating the Value of Equity Options*, *Quantitative Strategies Research Note*, Goldman, Sachs and Co.

Further Reading

Osborne, M.F.M. (1970). Brownian motion in the stock market, in *The Random Character of the Stock Market*, P. Cootner, ed., MIT Press.

Stutzer, M. (2003). Portfolio choice with endogenous utility: a large deviations approach, *Journal of Econometrics* **116**, 365–386.

Related Articles

Cramér’s Theorem; Esscher Transform; Generalized Method of Moments (GMM); Large Deviations; Minimal Entropy Martingale Measure; Model Calibration; Weighted Monte Carlo.

YUICHI KITAMURA & MICHAEL STUTZER

Risk Measures: Statistical Estimation

The recent interest in the estimation of risk measures is explained by changes in regulations in finance and insurance industries [4]. Ten years ago, the regulation was applied to a rather limited scope of assets and firms and only consisted in fixing minimal reserve level to hedge risky investments. However, these levels were computed without clear reference to the underlying risk. The new regulations, known as *Basel 2* for the finance industry and *Solvency 2* for the insurance industry, have proposed to account for risk in computing the reserves (Pillar 1). They also insist on the need for introduction of internal models by banks or insurance companies to follow their risk (Pillar 2) and on transparency (Pillar 3). This is a coherent program, which concerns risk modeling as well as the creation or the improvement of databases, and has to account for the heterogeneous technological level in these industries. In this article, we focus on risk measures, their computation, and updating.

In the current implementation of Basel 2 and Solvency 2, the risk measure suggested by regulators and generally retained by firms is the Value-at-Risk (VaR). This measure has several drawbacks, but has the advantage of being easily understood by practitioners. It can be presented as follows. Let us consider a given portfolio of assets. This portfolio can concern stocks, corporate bonds, consumer loans, or life insurance contracts. In practice, it corresponds to a business line of a bank, a credit institution, or an insurance company. The value of this portfolio at date t is denoted by W_t , but at this date its future values W_{t+h} are unknown.^a To hedge this uncertainty, some reserves R are introduced. With these reserves, the future portfolio value increases to $W_{t+h} + R$, which diminishes the potential loss. Let us fix a probability of loss $\alpha = 1, 5$, or 10% , say; then, the reserve level can be chosen such that

$$P_t[W_{t+h} + R < 0] = \alpha \quad (1)$$

where P_t denotes the probability given the information available at time t . By solving equation (1), we see that the reserve level is the opposite of the α -quantile of the distribution of the future portfolio value. It depends on the selected loss probability

α , investment horizon h , date t , information available at this date, and the set of selected assets. It can be denoted by $R(t, h; \alpha)$ and is written in a monetary unit as the dollar or the euro. For the determination of reserves, the admissible sets of assets, the level α , the horizon h , and the date of computations are imposed by the regulators. Other values can be selected by firms for internal use, especially for fixing and real-locating the so-called economic capital.

However, the portfolio value (W_t) generally features a nonstationary evolution, which can render difficulty in the determination of R . To circumvent this difficulty, it has been proposed to introduce the notion of VaR defined as

$$VaR(t, h; \alpha) = R(t, h; \alpha) + W_t \quad (2)$$

Thus, the VaR defined as

$$P_t[W_{t+h} - W_t < -VaR(t, h; \alpha)] = \alpha \quad (3)$$

is the opposite of the conditional α -quantile of the change in portfolio value ($W_{t+h} - W_t$), which is a much more stationary process.^b This definition highlights the importance of the (conditional) quantile function for building risk measures.

In the following section, we discuss properties that are suitable for risk measures. This allows to define a large class of risk measures, called *distortion risk measures (DRMs)*, which is flexible and includes the VaR as a special case. In the section Nonparametric Estimation in the IID Framework, we consider the estimation of DRMs, when the portfolio returns are independent of the past and identically distributed (IID). Under the IID assumption, the estimated risk measures have simple expressions in terms of returns, and it is easy to estimate their accuracy. These simple computations, which involve only a ranking of returns plus some averaging, explains the success of this practice, called *historical simulation*.

Nevertheless, the IID assumption is not realistic (see **Stylized Properties of Asset Returns**). For instance, it would be preferable to account for volatility persistence, for cycle effects, and so on. The introduction of serial dependences when computing and analyzing conditional risk measures is not an easy task and is only in its infancy for regulation purposes. Finally, we discuss the validation of the approaches used to compute the reserves and the coherent definitions of reserves for several business lines.

Risk Measures

First, we discuss the construction and the comparison of risk measures for a given date and a given horizon. For this reason, we do not indicate indexes t and h . Two axiomatic theories have been introduced in the literature to construct risk measures.

1. Coherent risk measures

The first constructive approach to evaluate the capital requirement was proposed by Artzner *et al.* [2]. $R(W)$ denotes the required reserve amount for a portfolio W , that is, for the future portfolio value W_{t+h} .

The axioms are as follows:

Distribution dependence: $R(W)$ depends only on the distribution of W .

Monotonicity: If W is more risky than W^* by the stochastic dominance^c of order 1, then $R(W) \geq R(W^*)$.

Drift invariance: $R(W + c) = R(W) - c$, for any W and any deterministic c .

Positive homogeneity: $R(\lambda W) = \lambda R(W)$, for any W and any nonnegative scalar λ .

Subadditivity: $R(W + W^*) \leq R(W) + R(W^*)$, for any W and W^* .

Note that the homogeneity and subadditivity properties imply the convexity of the coherent risk measure (see **Convex Risk Measures**).

2. The Wang axioms

The next set of axioms were introduced in a series of papers by Wang and Young [27], for the purpose of analyzing losses in insurance. Thus, we restrict ourselves to losses that are negative W . The axioms are as follows:

Distribution dependence: $R(W)$ depends on the distribution of W only.

Positivity: $R(W) \geq 0$.

Deterministic loss: $R(c) = -c$, for any deterministic loss $c < 0$.

Positive homogeneity: $R(\lambda W) = \lambda R(W)$ for any loss $W < 0$, and any nonnegative scalar λ .

Additivity for comonotonic risk: $R(W + W^*) = R(W) + R(W^*)$ for any losses W and W^* , which are increasing deterministic transformations of the same underlying risk Z .

Some of the above-mentioned conditions are easily written in terms of quantile functions, by noting that

- the quantile function of λW is λ times the quantile function of W and
- the quantile function of the sum $W + W^*$ of comonotonic risks is the sum of the quantile functions of W and W^* .

3. Distortion risk measures

A reserve level can be considered as a measure of risk. The interpretation in terms of quantile functions of the homogeneity and additivity for comonotonic risk assumptions implies that a risk measure satisfying Wang's set of axioms can be written as a linear functional of the quantile function. By applying the Riesz theorem, this leads to the introduction of the so-called distortion risk measures (DRM) defined by Wang [25, 26]:

$$\Pi(Q, H) = - \int_0^1 Q(u) dH(u) \quad (4)$$

where Q denotes the quantile function of the uncertain loss and H is a measure, called the *distortion measure*. The positivity axiom implies that H is a positive measure and the assumption of certain loss implies that H is a probability measure. Moreover, the DRM is coherent if H is concave. DRMs are the basis of the so-called “dual theory of choices under risk” [22, 27, 28]. Thus, a DRM is simply a weighted average of opposite quantile values. Well-known risk measures are derived by considering appropriate DRMs.

(a) VaR

When $H(u; \alpha) = \mathbf{1}_{(u \geq \alpha)}$, with $\alpha \in (0, 1)$, we get

$$\Pi(Q, H(\cdot; \alpha)) = -Q(\alpha) = VaR(\alpha) \quad (5)$$

(b) Tail-VaR

When $H(u, \alpha) = \min(u/\alpha, 1)$, with $\alpha \in [0, 1]$, we get the equally weighted average of VaR on the interval $[0, \alpha]$:

$$\begin{aligned} \Pi[Q, H(\cdot, \alpha)] &= - \int_0^\alpha \frac{Q(u)}{\alpha} du \\ &= - \frac{1}{\alpha} \int_0^\alpha u dF(u) \\ &= E[-W | W < -VaR(\alpha)] \end{aligned} \quad (6)$$

which is the opposite of the expected portfolio value when there is a loss; this measure is called the *Tail-VaR* at level α . This is a coherent risk measure since H is a concave function.

(c) Proportional hazard DRM

This risk measure is obtained for the power law transformation $H(u; p) = u^p$, where $p \geq 0$ and has been introduced for the definition of insurance premium.

Nonparametric Estimation in the IID Framework

In this section, we assume a sequence of IID returns concerning a given portfolio, and consider the estimation of the risk measures by their sample counterpart. Since these measures are defined from the quantile function, we first recall the asymptotic behavior of the empirical quantile. Then, we deduce the properties of estimated distortion risk measures.

The Empirical Quantile Function

$y_1, \dots, y_T$ denotes the sequence of IID returns, with a common distribution with pdf f , cdf F , and quantile function Q . The empirical cdf is defined as

$$\hat{F}_T(y) = \frac{1}{T} \sum_{i=1}^T \mathbf{1}_{y_i \leq y} \quad (7)$$

whereas the empirical quantile function is

$$\hat{Q}_T(u) = \inf\{y : \hat{F}_T(y) \geq u\} \quad (8)$$

When the distribution is continuous with positive pdf, the empirical cdf and quantile functions are asymptotically related by means of the Bahadur representation [19, Section 4.3]:

$$\begin{aligned} \sqrt{T}[\hat{Q}_T(u) - Q(u)] \\ = -\frac{1}{f[Q(u)]} \sqrt{T}[\hat{F}_T[Q(u)] - u] + o_P(1) \end{aligned} \quad (9)$$

where $o_P(1)$ is negligible in probability. This asymptotic equivalence allows for deriving the asymptotic behavior of the empirical quantile from the asymptotic behavior of the empirical cdf.

It is well known that the empirical cdf is strongly consistent with the true underlying cdf and converges weakly, that is, in distribution, to a Brownian bridge up to an appropriate standardization [23]:

$$\sqrt{T}[\hat{F}_T(Q(u)) - u] \Rightarrow B(u), u \in [0, 1] \quad (10)$$

where $\Rightarrow$ denotes weak convergence and B is a Brownian bridge, that is, a Gaussian process with zero mean and covariance: $\text{Cov}[B(u_1), B(u_2)] = \min(u_1, u_2) - u_1 u_2$. By applying the Bahadur representation (9), we deduce that

$$\sqrt{T}[\hat{Q}_T(u) - Q(u)] \Rightarrow -\frac{1}{f[Q(u)]} B(u) \quad (11)$$

Empirical Distortion Risk Measures

Let us now consider a given risk measure:

$$\Pi(Q, H) = \int_0^1 Q(u) dH(u) \quad (12)$$

where H is the distortion measure. Its empirical counterpart is given by

$$\hat{\Pi}_T(H) = \Pi(\hat{Q}_T, H) = -\int_0^1 \hat{Q}_T(u) dH(u) \quad (13)$$

It is easily written in terms of observed returns. Let us consider the returns ranked in increasing order as $y_1^* \leq y_2^* < \dots < y_T^*$, called the *order statistic*. We have

$$\hat{\Pi}_T(H) = -\sum_{i=1}^T \{y_i^* [H(i/T) - H((i-1)/T)]\} \quad (14)$$

This estimator is a linear combination of the order statistics $y_1^* < \dots < y_T^*$, with weights depending on the sample size.^d

Formula (14) is especially simple for estimating a VaR, that is, the opposite of the value of the α -quantile, when the sample size is well chosen. Indeed, let us assume, for instance, a risk level $\alpha = 5\%$ and $T = 200$; the estimated VaR (5%) is simply

$$\widehat{VaR}_T(5\%) = -y_{10}^* \quad (15)$$

(since $5\% \times 200 = 10$).

Asymptotic Behavior of Estimated DRM

The asymptotic behavior of the estimated DRM is directly deduced from the asymptotic behavior of the empirical quantile function. The estimator $\hat{\Pi}_T(H)$

converges to the theoretical DRM, $\Pi(Q; H)$, and we have

$$\begin{aligned} & \sqrt{T}[\hat{\Pi}_T(H) - \Pi(Q, H)] \\ & \Rightarrow \int_0^1 \frac{1}{f[Q(u)]} B(u) dH(u) \end{aligned} \quad (16)$$

This means that the estimated DRM is asymptotically Gaussian, with zero mean and a variance given by

$$\begin{aligned} & V_{as}\{\sqrt{T}(\hat{\Pi}_T(H) - \Pi(Q, H))\} \\ & = \int_0^1 \int_0^1 \frac{\min(u_1, u_2) - u_1 u_2}{f[Q(u_1)]f[Q(u_2)]} dH(u_1) dH(u_2) \end{aligned} \quad (17)$$

This expression of the asymptotic variance can be difficult to use in practice, since it involves the pdf, which is rather difficult to estimate. It can

$$\begin{aligned} & V_{as}\{\sqrt{T}[T\widehat{VaR}(\alpha) - TVaR(\alpha)]\} \\ & = \frac{V[Y|Y \leq -VaR(\alpha)] + (1 - \alpha)[TVaR(\alpha) - VaR(\alpha)]^2}{\alpha} \end{aligned} \quad (21)$$

be simplified when the DRM is continuous with distortion density h . Indeed, we get

$$\begin{aligned} & V_{as}\{\sqrt{T}[\hat{\Pi}_T(H) - \Pi(Q, H)]\} \\ & = \int_0^1 \int_0^1 \frac{[\min(u_1, u_2) - u_1 u_2]}{f[Q(u_1)]f[Q(u_2)]} \\ & \quad \times h(u_1)h(u_2) du_1 du_2 \\ & = \int_0^1 \int_0^1 [\min(u_1, u_2) - u_1 u_2] \\ & \quad \times h(u_1)h(u_2) dQ(u_1) dQ(u_2) \end{aligned} \quad (18)$$

by noting that $1/f[Q(u)]$ is the quantile density. Therefore, we get

$$\begin{aligned} & V_{as}\{\sqrt{T}[\hat{\Pi}_T(H) - \Pi(Q, H)]\} \\ & = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [\min(F(y_1), F(y_2)) - F(y_1)F(y_2)] \\ & \quad \times h[F(y_1)]h[F(y_2)] dy_1 dy_2. \end{aligned} \quad (19)$$

Example 1 (Sample VaR). The $VaR(\alpha)$ is obtained with the distortion measure equal to the point mass at α . The distortion measure is not continuous and formula (19) cannot be applied. From the general result, we know that

$$\begin{aligned} & \sqrt{T}[\widehat{VaR}_T(\alpha) - VaR(\alpha)] = -\sqrt{T}[\hat{Q}_T(\alpha) \\ & - Q(\alpha)] \rightsquigarrow N\left[0, \frac{1}{f[Q(\alpha)]}\alpha(1 - \alpha)\right] \end{aligned} \quad (20)$$

The accuracy of this estimator depends not only on the risk level α but also on the tail behavior of the distribution, since the density f (generally) tends to zero at infinity. Thus, the estimator of the VaR is not very accurate for small probability of loss α , and the lack of accuracy increases with increase in magnitude of the tail.

Example 2 (Sample Tail-VaR). By applying formula (19), the asymptotic variance of the estimated Tail-VaR [15] is

where TVaR denotes the Tail-VaR.

This asymptotic variance has two components. The first one measures the variability of the loss when a loss occurs, and the second one accounts for the expected loss beyond the VaR.

Estimated Accuracy of a Sample DRM

It is important to distinguish the case of continuous distortion measures, which exclude the VaR.

1. Continuous distortion measure

When H is continuous with density h , we can directly use the sample counterpart of formula (19) to derive a consistent estimator of the variance of $\hat{\Pi}_T(H)$ [15, 18]. We get

$$\begin{aligned} & \hat{V}_{as}(\sqrt{T}[\hat{\Pi}_T(H) - \Pi(Q; H)]) \\ & = \sum_{i=1}^{T-1} \sum_{j=1}^{T-1} \left\{ \left(\min\left(\frac{i}{T}, \frac{j}{T}\right) - \frac{i}{T} \frac{j}{T} \right) \right\} \end{aligned}$$

$$\times h\left(\frac{i}{T}\right)h\left(\frac{j}{T}\right)(y_{i+1}^* - y_i^*)(y_{j+1}^* - y_j^*)\} \quad (22)$$

2. Discrete distortion measure

Formula (19) cannot be applied to the VaR. Since

$$V_{\text{as}}\left(\sqrt{T}[\hat{Q}_T(\alpha) - Q(\alpha)]\right) = \frac{1}{f[Q(\alpha)]}\alpha(1-\alpha) \quad (23)$$

this variance can be estimated by

$$\hat{V}_{\text{as}}\left(\sqrt{T}[\hat{Q}_T(\alpha) - Q(\alpha)]\right) = \frac{1}{\hat{f}_T(\hat{Q}_T(\alpha))}\alpha(1-\alpha) \quad (24)$$

where $\hat{f}_T$ is, for instance, a kernel estimator of the density.

Dynamic Analysis

Despite its success among practitioners, the computation and analysis of VaR and DRM in an IID framework is not realistic, and the (nonlinear) return dynamics should be taken into account. It arises in the computation of the risk measure itself, which depends on the conditional quantile function.^c In this conditional framework, a nonparametric approach is difficult to follow, since the number of conditioning variables can be rather large, including current and lagged values of several asset returns, for instance. This explains why parametric or semiparametric approaches are generally considered, with the risk of using a misspecified model, that is, of introducing an additional model risk. In the first subsection, we consider a conditionally Gaussian model and derive the corresponding estimators of the DRM. The second subsection concerns the term structure of VaR. In the third subsection, we discuss dynamic models recently introduced for the purpose of computing dynamic VaR. These models are directly written in terms of conditional quantile functions.

Parametric Dynamic Models

1. Definition of the dynamic DRM

The dynamic model is generally written for the portfolio returns of interest $y_1, \dots, y_T$. Let us assume that

the conditional distribution of y_t , given observable variables z_{t-1} , say, is such that

$$y_t = m(z_{t-1}; \theta) + \sigma(z_{t-1}; \theta)u_t \quad (25)$$

where (u_t) are the IID standard Gaussian variables and θ is an unknown parameter. $m(z_{t-1}; \theta)$ (respectively, $\sigma(z_{t-1}; \theta)$) is the conditional drift (respectively, conditional volatility). The conditional quantile function is

$$Q_t(u) = m(z_{t-1}; \theta) + \sigma(z_{t-1}; \theta)\Phi^{-1}(u) \quad (26)$$

where Φ denotes the cdf of the standard normal. We deduce that a dynamic DRM can be written as

$$\begin{aligned} \Pi(Q_t, H) &= - \int_0^1 Q_t(u) dH(u) = -m(z_{t-1}; \theta) \\ &\quad - \sigma(z_{t-1}; \theta) \int_0^1 \Phi^{-1}(u) dH(u) \end{aligned} \quad (27)$$

The DRM depends on the information by means of the conditional mean and volatility and is a linear combination of these summary statistics.

This type of analysis is usually applied with autoregressive conditionally heteroscedastic (ARCH) models or switching regime models [6], and is easily extended to error terms with no Gaussian distribution, such as the Student distribution.

For instance, such an approach is followed by J.P. Morgan with an IGARCH model defined by

$$y_t = \sigma_t u_t \quad (28)$$

where

$$\sigma_t^2 = \theta \sigma_{t-1}^2 + (1-\theta)y_{t-1}^2 = (1-\theta) \sum_{j=1}^{\infty} \theta^{j-1} y_{t-j}^2 \quad (29)$$

and $\theta = 0.95$. Such an automatic dynamic, assumed valid for any portfolio return, is likely misspecified, and it is preferable to estimate the return dynamics separately for each portfolio.

2. Estimation of the dynamic DRM

The DRM is usually estimated in two steps. In the first step, the parameter θ is estimated by (conditional) maximum likelihood (ML), that is, by maximizing

$$\hat{\theta}_T = \arg \max_{\theta} \sum_{t=1}^T \left\{ -\frac{1}{2} \log \sigma^2(z_{t-1}; \theta) - \frac{1}{2} \frac{[y_t - m(z_{t-1}; \theta)]^2}{\sigma^2(z_{t-1}; \theta)} \right\} \quad (30)$$

This estimator has the standard asymptotic properties of an ML estimator. Then, in the second step, the DRM is estimated as

$$\hat{\Pi}_{t,T} = -m(z_{t-1}; \hat{\theta}_T) - \sigma(z_{t-1}; \hat{\theta}_T) \int_0^1 \Phi^{-1}(u) dH(u) \quad (31)$$

For instance, we have

$$\widehat{VaR}_T(t; \alpha) = -m(z_{t-1}; \hat{\theta}_T) - \sigma(z_{t-1}; \hat{\theta}_T) \Phi^{-1}(\alpha) \quad (32)$$

The first step is valid for any type of DRM and can be applied to get a coherent set of estimated risk measures, such as $VaR(1\%)$, $VaR(5\%)$, $VaR(10\%)$, and $TVaR(10\%)$.

The Term Structure of VaR

Until now, we focused on the evolution of the dynamic risk measure, that is, on its dependence with respect to time for a fixed horizon $h = 1$. It is interesting to consider how the risk measure depends on the horizon h for a fixed date t . If y_{t+1} denotes the opposite change of portfolio value between t and $t + 1$, the opposite change of portfolio value between t and $t + h$ is $y_{t,h} = y_{t+1} + \dots + y_{t+h}$. Therefore, the conditional quantile function of interest $Q_{t,h}$ is now associated with the conditional distribution of $y_{t,h}$ given the information at time t .

The term structure of VaR is the mapping

$$h \rightarrow VaR(t, h; \alpha) = -Q_{t,h}(\alpha) \quad (33)$$

which can be defined for any date t and probability level α .

In simple cases, the term structure of VaR is derived explicitly. For instance, let us assume a Gaussian autoregressive process of order 1 for the short-term opposite portfolio change

$$y_t = \mu + \varphi(y_{t-1} - \mu) + \sigma u_t \quad (34)$$

where μ, φ, σ are the mean, autoregressive coefficient, and innovation variance, respectively, and (u_t) is a sequence of IID standard Gaussian variables. Thus, the conditional distribution of $y_{t,h}$ is Gaussian, with mean

$$m(y_t, h) = \mu h + \varphi \frac{1 - \varphi^h}{1 - \varphi} (y_t - \mu) \quad (35)$$

and variance

$$\sigma^2(h) = \frac{\sigma^2}{(1 - \varphi^2)} \left\{ h - 2\varphi(1 - \varphi^h) + \varphi^2 \frac{1 - \varphi^{2h}}{1 - \varphi^2} \right\} \quad (36)$$

We deduce the VaR at horizon h as

$$VaR(t, h; \alpha) = -m(y_t, h) - \sigma(h) \Phi^{-1}(\alpha) \quad (37)$$

The IID Gaussian framework is obtained when $\varphi = 0$. In this case, we have $VaR(t, h; \alpha) = -\mu h - \sigma \sqrt{h} \Phi^{-1}(\alpha)$, and the term structure involves both a linear and square-root function of the horizon h . This formula is often used by practitioners with $\mu = 0$ to derive automatically the VaR at horizon h from the VaR at horizon 1 by $VaR(t, h; \alpha) = \sqrt{h} VaR(t, 1; \alpha)$. As seen earlier, this simplified formula is valid under very restrictive assumptions and has to be systematically avoided.

In fact, it is necessary to carefully compute the VaR at the different horizons case by case. In general, the term structure of VaR has no explicit expression, but can be derived by simulation, whenever the short-term quantile function $Q_{t,1}$ has an explicit expression: $Q_{t,1}(u) = Q(u; y_{t-1})$, say, if we assume for illustration, a Markov process. The simulation procedure is as follows:

1. Consider independent drawings $u_{t+1}^s, \dots, u_{t+h}^s$ in the standard uniform distribution on $(0, 1)$

2. Apply the recursive formula

$$\begin{aligned} y_{t+1}^s &= Q(u_{t+1}^s, y_t) \\ y_{t+2}^s &= Q(u_{t+2}^s, y_{t+1}^s) \\ y_{t+h}^s &= Q(u_{t+h}^s, y_{t+h-1}^s) \end{aligned} \quad (38)$$

$(y_{t+1}^s, \dots, y_{t+h}^s)$ has the same conditional distribution as $(y_{t+1}, \dots, y_{t+h})$.

3. Replicate $s = 1, \dots, S$, and approximate $Q_{t,h}(\alpha)$ by the empirical quantile distribution of $y_{t,h}^s = y_{t+1}^s + \dots + y_{t+h}^s$, $s = 1, \dots, S$.

Dynamic Quantile Models

Specification. Since the dynamic DRM is easily written in terms of the conditional quantile function, it has been recently proposed to directly specify the dynamic model for (y_t) in terms of function Q_t , instead of using more standard (nonlinear) autoregressive specifications [11, 13]. For instance, a dynamic additive quantile (DAQ) model is

$$Q_t(u; \theta) = \sum_{k=1}^K a_k(\underline{y}_{t-1}; \alpha_k) Q_k(u; \beta_k) + a_0(\underline{y}_{t-1}; \alpha_0) \quad (39)$$

where $\underline{y}_{t-1} = (y_{t-1}, y_{t-2}, \dots)$, Q_k are path-independent baseline quantile functions with identical range $(-\infty, +\infty)$ and $a_k(\underline{y}_{t-1}; \alpha_k)$, $k = 1, \dots, K$ are non-negative functions of the past. $\theta = (\alpha'_0, \dots, \alpha'_K, \beta'_1, \dots, \beta'_K)'$ is the vector of parameters. The positivity restrictions on functions a_k ensure that $Q_t(\cdot; \theta)$ is an increasing function and thus is compatible with the quantile interpretation.

The DAQ model encompasses simple specifications of interest. For instance, specifications of the type

$$\begin{aligned} Q_t(u) &= Q_0(u) + Q_1(u) \max(y_{t-1}, 0) \\ &\quad + Q_2(u) \max(-y_{t-1}, 0) \end{aligned} \quad (40)$$

or

$$\begin{aligned} Q_t(u) &= \gamma_0(u) + \gamma_1(u) Q_{t-1}(u) + \gamma_2(u) |y_{t-1}| \\ &= \frac{\gamma_0(u)}{1 - \gamma_1(u)} + \sum_{k=1}^{\infty} \gamma_2(u) \gamma_1(u)^k |y_{t-k}| \end{aligned} \quad (41)$$

are special cases of the conditional autoregressive variance at risk (CAViaR) models introduced by Engle and Manganelli [9]. A specification such as

$$\begin{aligned} Q_t(u) &= m_0 + m_1 |y_{t-1} - \mu| \\ &\quad + (\sigma_{0,0} + \sigma_{0,1} |y_{t-1} - \mu|)^{1/2} \Phi^{-1}(u) \\ &\quad + (\sigma_{1,0} + \sigma_{1,1} |y_{t-1} - \mu|)^{1/2} tg[\pi(u - 1/2)] \end{aligned} \quad (42)$$

defines a quantile mixture of a Gaussian distribution (with baseline quantile function Φ^{-1}) and a Cauchy distribution (with baseline quantile function $tg[\pi(u - 1/2)]$), with ARCH effects in both scale parameters and a risk premium in the location parameter.

For a DAQ model, the dynamic DRM are immediately derived as

$$\begin{aligned} \Pi(Q_t, H) &= \sum_{k=1}^K a_k(\underline{y}_{t-1}; \alpha_k) \int Q_k(u; \beta_k) dH(u) \\ &\quad + a_0(\underline{y}_{t-1}; \alpha_0) \end{aligned} \quad (43)$$

They depend on the same set of dynamic summaries of the past $a_k(\underline{y}_{t-1}; \alpha_k)$, $k = 0, \dots, K$, with sensitivity coefficients equal to the baseline DRMs. These explicit expressions (see examples (40)–(42)) are appropriate for easily deriving the term structure of VaR by simulation (see the section The Term Structure of VaR).

Estimation. When the dynamic model is directly written in terms of a quantile function with a simple expression of Q_t , it can be numerically cumbersome to derive the conditional pdf $f_t(y) = \left(\frac{dQ_t}{du} [Q_t^{-1}(y)] \right)^{-1}$, which requires the inversion of function Q_t at each observation date. As a consequence, it can be difficult to derive the maximum likelihood estimate of the parameter θ .

Different estimation approaches have been proposed in the statistical literature.

1. When $Q_t(u; \theta) = Q(u; \theta)$ is independent of the path, the parameter θ can be estimated by calibrating on L -moments or trimmed L -moments [8, 17]. The theoretical L -moments or trimmed L -moments are of the type $\int Q(u; \theta) P_l(u) du$, $l = 1, \dots, L$, where P_l are well-chosen

polynomials and their sample counterparts are $\frac{1}{T} \sum_{i=1}^T y_i^* P_i(t/T)$, where $y_1^* < \dots < y_T^*$ is the order statistic, that is, the observations $y_1, \dots, y_T$ are ranked in increasing order. This approach is easy to implement, but does not provide efficient estimators, and is difficult to extend to a dynamic framework [see the Generalized Method of L -Moments in [13, Section 4].

2. In the dynamic framework, the parameter θ can also be consistently estimated by minimizing an objective function of the type in [20]:

$$\begin{aligned} \hat{\theta}_T &= \arg \min_{\theta} \sum_{t=1}^T \{ \alpha \max[y_t - Q_t(\theta), 0] \\ &\quad + (1 - \alpha) \max[Q_t(\theta) - y_t, 0] \} \\ &= \arg \min_{\theta} \sum_{t=1}^T \{ \alpha [y_t - Q_t(\theta)]^+ \\ &\quad + (1 - \alpha) [y_t - Q_t(\theta)]^- \} \end{aligned} \quad (44)$$

where α is a scalar between 0 and 1 (see the section The Control Problem for the interpretation of the objective function).

3. The estimation methods above are easy to apply, but are generally not efficient. They can be used as starting values in the optimization algorithm of the estimated inverse Kullback–Leibler information criterion (KLIC) :

$$\begin{aligned} \hat{\theta}_T &= \arg \max_{\theta} \sum_{i=1}^T \left\{ \int_0^1 \log \left[\frac{dQ}{du}(u|y_{t-1}; \theta) \right] \right. \\ &\quad \left. + \int_0^1 \log \hat{f}_T[Q(u|y_{t-1}; \theta)|y_{t-1}] \right\} du \end{aligned} \quad (45)$$

where $\hat{f}_T$ is a kernel estimator of the conditional pdf of y_t given y_{t-1} . The inverse KLIC estimator is asymptotically equivalent to the ML estimator and asymptotically efficient [13].

Control and Out-of-sample Validation

The current regulations have at least three limitations: (i) the large use of the VaR to measure the risk and fix the required capital. This measure

accounts for the probability of loss, but not for the magnitude of the loss beyond the VaR. This drawback can be corrected by considering another DRM such as the Tail-VaR. (ii) The lack of coherency between the suggested (but often not followed) computation of the VaR as a conditional quantile, and the *ex post* validation of this computation by the regulator, which is generally based on a historical (that is unconditional) analysis. (iii) The computation of VaR is performed independently for each business line without taking into account the potential dependence between the risks of different lines.

The two latter limitations can be solved by considering the problem of the determination of reserves as a control problem [12, 16].

The Control Problem

The concept of VaR has been initially defined in a rather *ad hoc* way, but is the solution of an optimal control problem. More precisely, let us consider the objective function

$$\psi(t, z_t) = E_t[\alpha(y_{t+1} + z_t)^+ + (1 - \alpha)(y_{t+1} + z_t)^-] \quad (46)$$

where α is the announced risk level. The decision criterion corresponds to an asymmetric loss function, which distinguishes two types of errors. When $y + z < 0$, the reserve level is not large enough to hedge the loss; when $y + z > 0$, the reserve level is too high. The type I error has to be avoided from the prudential point of view of the regulator, whereas the type II error has to be avoided for the firm, since the reserves receive no the interest rate in the current regulation.

The solution of the optimization problem (46) [19],

$$\begin{aligned} \hat{z}_t &= \arg \max_{z_t} E_t[\alpha(y_{t+1} + z_t)^+ \\ &\quad + (1 - \alpha)(y_{t+1} + z_t)^-] \end{aligned} \quad (47)$$

performed over the variables z_t function of the information, is

$$\hat{z}_t = -Q_t(\alpha) = VaR_t(\alpha) \quad (48)$$

Thus, the reserve level can be interpreted as a control variate and the conditional VaR as the optimal control variate.

In this control framework, the optimal value of the criterion function is given as [24]

$$\begin{aligned}\psi(t, \hat{z}_t) &= \alpha P_t[y_{t+1} > -\hat{z}_t] \\ &\times E_t[(y_{t+1} + \hat{z}_t)^+ | y_{t+1} > -\hat{z}_t] + (1 - \alpha) \\ &\times P_t[y_{t+1} < \hat{z}_t] E_t[(y_{t+1} + \hat{z}_t)^- | y_{t+1} < -\hat{z}_t] \\ &= \alpha(1 - \alpha)[TVaR^+(\alpha) + TVaR^-(\alpha)]\end{aligned}\quad (49)$$

where $TVaR^+$ and $TVaR^-$ are the Tail-VaR associated with the right and left tails of the conditional distribution, respectively. Thus, if the VaR is used as the reserve level, the Tail-VaRs have also to be computed and provide the optimal value of the decision criterion.

Specification Test

In practice, it is usual to check the validity of the proposed reserve levels $\widehat{VaR}_t$ by comparing the announced level α with the frequency of exceedance dates with $y_t + \widehat{VaR}_t < 0$. This practice is misleading. Indeed, two different procedures for determining the reserves can give the same frequency, 5%, say, but the dates at which the reserves are not sufficient can be equally located with the first procedure and come in cluster with the second one. Clearly, if $T = 100$, the second procedure, in which losses will cumulate over five periods, say, is more risky. An appropriate validation has to distinguish these situations.

In fact, the first-order conditions of the control problem are as follows:

$$\frac{\partial \psi}{\partial z}(t, \hat{z}_t) = 0 \Leftrightarrow E_t[\mathbf{1}_{y_{t+1} + \hat{z}_t < 0}] = \alpha \quad (50)$$

which is the definition of the conditional VaR, that is,

$$P_t[y_{t+1} + \hat{z}_t < 0] = \alpha \quad (51)$$

The first-order condition (50) is a conditional moment restriction. As usual, it can be replaced by a set of unconditional moment restrictions by introducing instrumental variables g_t , which are observable functions of the past. We get

$$E[g_t(\mathbf{1}_{y_{t+1} + \hat{z}_t < 0} - \alpha)] = 0 \quad \forall g_t \quad (52)$$

Specification tests can be constructed from the sample counterparts of these restrictions; the test

compares the statistics $(1/T) \sum_{t=1}^T g_t(\mathbf{1}_{y_{t+1} + \hat{z}_t < 0} - \alpha)$ to zero, after an appropriate standardization.

The current practice, considering the historical frequency $(1/T) \sum_{t=1}^T (\mathbf{1}_{y_{t+1} + \hat{z}_t < 0} - \alpha)$, thus, selects $g_t = 1$ as the single instrumental variable. Clearly, other instrumental variables have also to be considered to detect the persistence in extreme losses, such as

$$g_{1,t} = \mathbf{1}_{y_t + \hat{z}_{t-1} < 0}, g_{2,t} = \mathbf{1}_{y_{t-1} + \hat{z}_{t-2} < 0}, \dots \quad (53)$$

which indicate the lack of reserve in the past.

Dependent Business Lines

Let us consider two business lines, say, $y_{1,t}$, $y_{2,t}$. The current regulation suggests to compute separately the VaR for the two lines, with the same α , that is, to compute $VaR_{1,t}(\alpha)$, $VaR_{2,t}(\alpha)$. Then, the total reserve is $VaR_{1,t}(\alpha) + VaR_{2,t}(\alpha)$. This regulation does not account for the dependence between the two risks. The question of reserve allocations between business lines is difficult. A possible solution introduced by Gourioux and Liu [16] focuses on the definition of an appropriate decision criterion. By analogy with the control problem described in the section The Control Problem, we may consider the objective function:

$$\begin{aligned}\psi(t; z_{1t}, z_{2t}) &= E_t\{[\alpha(y_{1,t+1} + z_{1t})^+ + (1 - \alpha)(y_{1,t+1} + z_{1t})^-] \\ &\times [\alpha(y_{2,t+1} + z_{2t})^+ + (1 - \alpha)(y_{2,t+1} + z_{2t})^-]\}\end{aligned}\quad (54)$$

which distinguished four types of errors, which are type I and II errors cross with business lines. If the loss variables $y_{1,t+1}$ and $y_{2,t+1}$ are conditionally independent, the objective function is equal to the product:

$$\begin{aligned}E_t[\alpha(y_{1,t+1} + z_{1t})^+ (1 - \alpha)(y_{1,t+1} + z_{1t})^-] \\ \times E_t[\alpha(y_{2,t+1} + z_{2t})^+ + (1 - \alpha)(y_{2,t+1} + z_{2t})^-]\end{aligned}\quad (55)$$

and the solutions of the control problem are simply the VaR computed per business line.

$$\hat{z}_{1,t} = VaR_{1,t}(\alpha), \hat{z}_{2,t} = VaR_{2,t}(\alpha) \quad (56)$$

If the loss variables are dependent, the optimization of the objective function (54) provides another

solution, which accounts for the dependence between extreme risks. The advantage of this approach, based on the direct optimization of the control decision criterion, is that it is easy to extend to any number of business lines and can provide a solution without having to estimate or specify the joint distribution of $(y_{1,t+1}, y_{2,t+1})$.

Concluding Remarks

The new regulations have led to a rethink on the whole question of risk measures in term of reserve levels and to see the determination of reserves as a control problem. As a consequence, the standard volatility risk measure is progressively being replaced by the VaR, Tail-VaR, or another DRM for several financial problems such as portfolio management [1, 5, 10, 14, 21], performance comparisons [7], or even derivative pricing, especially in insurance [3, 25].

Acknowledgments

The author gratefully acknowledges financial support of NSERC Canada and of the chair AXA/Risk Foundation “Large Risks in Insurance”.

End Notes

^aThus, we implicitly assume that the current portfolio value is known. This assumption is satisfied for assets that are highly traded on the markets, but should be discussed for over-the-counter (OTC) products.

^bThe portfolio returns $(W_{t+h} - W_t)/W_t$ can be even more stationary; the corresponding α -quantile becomes $-VaR(t, h; \alpha)/W_t$.

^cEquivalently if the cumulative distribution function (cdf) of W is smaller than the cdf of W^* .

^dThis is called an *L-statistic* in the statistical literature.

^eIt also arises when computing the VaR by historical simulation, since the asymptotic properties of the sample quantile, especially the asymptotic variance–covariance matrix, are modified under serial dependence.

References

- [1] Acerbi, C. & Simonetti, P. (2002). *Portfolio Optimization with Spectral Measure of Risk*. DP. Abaxbank.
- [2] Artzner, P., Delbaen, F., Eber, J. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**, 203–228.
- [3] Basak, S. & Shapiro, A. (2001). Value-at-risk based risk management: optimal policies and asset prices, *The Review of Financial Studies* **14**, 371–405.
- [4] Basel Committee on Banking Supervision (1995). *An Internal Model Based Approach to Market Risk Capital Requirements*. Bank of International Settlements, Working Paper.
- [5] Bassi, F., Embrechts, P. & Kafetzaki, M., (1997). Risk management and quantile estimation, in *Practical Guide to Heavy Tails*, Adler, R., Feldman, K. & M., Taqqu, eds, Birkhauser, Boston.
- [6] Billio, M. & Pelizzon, L. (2000). Value-at-risk: a multivariate switching regime approach, *Journal of Empirical Finance* **7**, 531–554.
- [7] Darolles, S., Gouriéroux, C. & Jasiak, J. (2009). *L-Performance Measures with Application to Hedge Funds*. CREST DP.
- [8] Elamir, E. & Seheult, A. (2003). Trimmed L-moments, *Computational Statistics and Data Analysis* **43**, 299–314.
- [9] Engle, R. & Manganelli, S. (2004). CAViaR: conditional autoregressive value-at-risk by regression quantile, *Journal of Business and Economic Statistics* **22**, 367–381.
- [10] Foellmer, H. & Leukert, P. (1999). Quantile hedging, *Finance and Stochastics* **3**, 251–273.
- [11] Giacomini, R. & Komunjer, I. (2005). Evaluation and combination of conditional quantile forecasts, *Journal of Business and Economic Statistics* **23**, 416–431.
- [12] Giacomini, R. & White, H. (2006). Test of conditional predictive ability, *Econometrica* **74**, 1545–1578.
- [13] Gouriéroux, C. & Jasiak, J. (2008). Dynamic quantile model, *Journal of Econometrics* **147**, 198–205.
- [14] Gouriéroux, C., Laurent, J.P. & Scaillet, O. (2000). Sensitivity analysis of value-at-risk, *Journal of Empirical Finance* **7**, 225–245.
- [15] Gouriéroux, C. & Liu, W. (2007). *Converting Tail-VaR to VaR: an Econometric Study*. CREST-DP.
- [16] Gouriéroux, C. & Liu, W. (2009). Control and out-of-sample validation of dependent risks, *Journal of Risk and Insurance* **75**, 683–707.
- [17] Hosking, J. (1990). L-moments: analysis and estimation of distribution using linear combinations of order statistics, *Journal of the Royal Statistical Society, B* **52**, 105–124.
- [18] Jones, B. & Zitikis, R. (2003). Empirical estimation of risk measures and related quantities, *North-American Actuarial Journal* **7**, 44–54.
- [19] Koenker, R. (2005). *Quantile Regression*, Cambridge University Press.
- [20] Koenker, R. & Xiao, Z. (2006). Quantile autoregression, *JASA* **101**, 980–990.
- [21] Liu, W. (2007). *Analysis of Risk Measures and Multidimensional Risk Dependence*. PhD dissertation, University of Toronto.
- [22] Schmeidler, D. (1989). Subjective probability and expected utility without additivity, *Econometrica* **57**, 571–587.

- [23] Shorack, G. & Wellner, J. (1986). *Empirical Processes with Applications to Statistics*, Wiley, New-York.
- [24] Uryasev, S. & Rockafellar, R. (2000). Optimization of conditional value-at-risk, *Journal of Risk* **2**, 21–41.
- [25] Wang, S. (1996). Premium calculation by transforming the layer premium density, *ASTIN Bulletin* **26**, 71–92.
- [26] Wang, S. (2000). A class of distortion operators for pricing financial and insurance risks, *Journal of Risk and Insurance* **67**, 15–36.
- [27] Wang, S. & Young, V. (1998). Ordering risks: expected utility theory versus Yaari’s dual theory of risk, *Insurance: Mathematics and Economics* **22**, 145–161.
- [28] Yaari, M. (1987). The dual theory of choice under risk, *Econometrica* **55**, 95–115.

Further Reading

Cont, R., Deguest, R. & Scandolo, G. (2007). *Robustness and Sensitivity Analysis of Risk Measurement Procedures*,

Columbia University Financial Engineering Report No. 2007-06.

Koenker, R. & Zhao, Q., (1996). Conditional quantile estimation and inference for ARCH models, *Econometric Theory* **12**, 793–813.

Related Articles

Backtesting; Convex Risk Measures; Expected Shortfall; Market Risk; Model Validation; Multivariate Distributions; Regulatory Capital; Simulation-based Estimation; Spectral Measures of Risk; Stylized Properties of Asset Returns; Value-at-Risk.

CHRISTIAN GOURIEROUX

Mixture of Distribution Hypothesis

The mixed distribution hypothesis (MDH) in finance maintains heavy tails in price fluctuations $\Delta X(t)$ that can be explained by embedding randomized time $\xi(t)$. Intuitively, the net price change $\Delta X(t)$ over period t is hypothesized to be the sum of $\xi(t)$ incremental interperiod steps, say $\Delta X_{t,i}$, giving

$$\Delta X(\xi(t)) = \sum_{i=1}^{\xi(t)} \Delta X_{t,i} \quad (1)$$

Randomized time $\xi(t)$ can be thought of as the amount of economic time or the number of information events that traders actually experience during t , typically associated with volume-traded $V(t)$. The resulting convolution $\Delta X(\xi(t))$ has a mixed distribution that can capture leptokurtosis, explain heterogeneity, including conditional heteroscedasticity, and improve model fitting similar to nonparametrics. Applications abound in finance (equity returns, bond prices) and macroeconomics (income, unemployment, exchange rates).

The statistical foundations of mixtures are well known [7]. If $y(t) := \Delta X(t)$ has density $f_y(y; \xi)$ and economic time $\xi(t) \geq 0$ has density $f_\xi(\xi)$, then the observed convolution $y(\xi(t))$ has a density mixture

$$f_{y(\xi)}(y) = \int_0^\infty f_y(y; \xi) f_\xi(\xi) d\xi \quad (2)$$

See [22] for a survey of mixture models, including generalized gamma and beta, log-Cauchy and lognormal.

The use of compound or mixed distributions to explain leptokurtosis in price fluctuations $\Delta X(t)$, while retaining a finite variance, first appears in the finance literature in the 1960s^a. A sequence of papers, including [25, 26], and prominently Clark [4], exploited in various ways a subordinated finite variance process model $\Delta X(\xi(t))$ as a direct response to Mandelbrot's [20] controversial α -stable hypothesis of aggregate prices. The latter "extreme value theory" framework permits infinite variance, but at a cost of estimation and inference challenges (*see* [5, 17]). Nevertheless, only the α -stable laws form a

closed distribution class under affine transformations: finite variance α -stable laws are the Gaussian laws that are necessarily symmetric, and infinite variance α -stable laws permit asymmetry, making this class attractive for applied statistical analysis in finance (*see* [28] for an encyclopedic treatment of the α -stable laws).

Tests of MDH against nonmixtures, including normality or α -stability, have mixed results [12, 21, 30], and there exists a sharp identification problem that is typically ignored (*consider* [19]). The subordinated process premise has successfully evolved into randomized volatility [1, 3, 11], which now receives vigorous attention in the finance literature [8, 9, 11, 27] and extreme value theory literature [14, 15, 18]; some applications of distribution mixtures in finance are motivated by specification flexibility rather than explaining heavy tails [10].

In the remainder of this article, we discuss at length Clark's [4] seminal model, noting the early attempts to improve on it, and conclude with remarks on recent extensions in the stochastic volatility literature.

Randomized Time and Subordination

Clark's [4] influential model of asset returns exploits Bochner's [2] and Feller's [7] work on subordinated Gaussian processes. Denote by μ_y , σ_y^2 , and κ_y respectively the mean, variance, and kurtosis of y , and $\sigma_{y|z}^2(t)$ is the conditional variance, given random z .

If $\{X(t)\} \equiv \{X(t) : -\infty < t < \infty\}$ is a stochastic process and $\{\xi(t)\}$, $\xi : \mathbb{N} \rightarrow \mathbb{R}_+$, is a "driving process" then the stochastically indexed $X(\xi(t))$ is "subordinated" to $X(t)$. If $\{\Delta\xi(t)\}$ is stationary and independent with mean $\mu_{\Delta\xi} > 0$, and^b $\Delta X(t) \stackrel{\text{iid}}{\sim} (0, \sigma_{\Delta X}^2)$, then the subordinated fluctuations $\Delta X(\xi(t))$ are stationary and independent, and distributed:

$$\Delta X(\xi(t)) \stackrel{\text{iid}}{\sim} (0, \mu_{\Delta\xi} \times \sigma_{\Delta X}^2) \quad (3)$$

Since $\mu_{\Delta\xi}$ reflects the arrival of new information, in Clark's model only the average amount of new information effects the dispersion of embedded returns $\sigma_{\Delta X(\xi)}^2$. If, in addition, price fluctuations are normally distributed $\Delta X(t) \stackrel{\text{iid}}{\sim} N(0, \sigma_{\Delta X}^2)$ and $\sigma_{\Delta\xi}^2 < \infty$ and $\Delta\xi(t)$ are independent of $X(t)$, then price

fluctuations $\Delta X(\xi(t))$ have kurtosis

$$\kappa_{\Delta X(\xi)} = 3 \left(1 + \frac{\sigma_{\Delta \xi}^2}{\mu_{\Delta \xi}^2} \right) > 3 \quad (4)$$

a simple positive function of $\sigma_{\Delta \xi}^2$, and larger than 3, the benchmark value for normals and therefore for what is considered canonically to be heavy tailed. The tails of price fluctuations $\Delta X(\xi(t))$ grow heavier as the variance $\sigma_{\Delta \xi}^2$ of information flow increases, and are heavy tailed in the strict sense of being thicker than normals for any directing process $\{\xi(t)\}$ with $\sigma_{\Delta \xi}^2 > 0$. Further, as in [4], if iid $\Delta \xi(t)$ are lognormal with parameters $\{c, v^2\}$ then

$$\begin{aligned} \mu_{\Delta \xi} &= \exp\{c + v^2/2\} \text{ and } \sigma_{\Delta \xi}^2 \\ &= \exp\{2c + v^2\} \times [\exp\{v^2\} + 1] \end{aligned} \quad (5)$$

and from equation (1) the convolution $y(\xi(t)) \equiv \Delta X(\xi(t))$ is lognormal-normally distributed as follows:

$$\begin{aligned} f_y(y) &= \frac{1}{2\pi\sigma_{\Delta X}^2\sigma_{\Delta \xi}^2} \int_0^\infty u^{-3/2} \\ &\times \exp \left\{ \frac{-(\ln u - c)^2}{2v^2} - \frac{y^2}{2u\sigma_{\Delta X}^2} \right\} du \end{aligned} \quad (6)$$

Volume-traded $V(t)$ may be positively associated with economic time $\xi(t)$, in which case equity returns should be heavier tailed during periods of substantial trader activity. Heimstra and Jones [13], for example, use this prediction to test for variance causality from volume to returns under a finite variance assumption.

Many models for relating randomized time to volume have been suggested. See [16] for a survey. Clark [4] used the parametric specification

$$\sigma_{\Delta X|\xi}^2(t) = \beta V(t)^\gamma, \text{ where } \beta, \gamma > 0 \quad (7)$$

for the log-price $X(t)$ of cotton future prices. Epps and Epps [6] use a parsimonious parametric asset price model to support Clark's mixtures of distribution model. Tauchen and Pitts [29] propose an asset price model that relates price fluctuations to volume via a variance-components framework with trader-specific and traderwide shocks. Unlike Clark their model predicts $\Delta X(t)$ and $V(t)$ are independent

for each transaction, but have moments with common parameters. For example, $\sigma_{\Delta X}^2$ and μ_V are both increasing functions of the variance of the trader-specific shock and therefore exhibit spurious positive association (as opposed to Clark's *ad hoc* association).

Stochastic volatility models for asset returns inherently exploit the premise that a mixture of distributions captures the arrival of news, making this literature the latest offshoots of Clark's model [1, 11, 27]. The Brownian semimartingale framework for stochastic volatility has become dominant in the option pricing literature by permitting volatility clustering and accounting for market incompleteness (which abnegates option redundancy), providing flexible extensions of the Black–Scholes option pricing model. See, especially, **Stochastic Volatility Models** and **Econometrics of Option Pricing** in this volume, and consult [9, 23].

End Notes

^aEvidently Newcomb [24] first used a mixture of normal distributions to account for numerous “discordant” observations (i.e., heavy tails), with an application to planetary orbits.

^b $y \sim (\mu, \sigma^2)$ implies y is distributed with mean μ and variance σ^2 . iid stands for “independent and identically distributed”.

References

- [1] Anderson, T. (1996). Return volatility and trading volume: an information flow interpretation of stochastic volatility, *Journal of Finance* **51**, 169–204.
- [2] Bochner, S. (1960). *Harmonic Analysis and the Theory of Probability*, University of California Press, Berkeley.
- [3] Carr, P., Geman, H., Madan, D.B. & Yor, M. (2003). Stochastic volatility for Lévy processes, *Mathematical Finance* **13**, 345–382.
- [4] Clark, P.K. (1973). Subordinated stochastic process model with finite variance for speculative processes, *Econometrica* **41**, 133–153.
- [5] Davis, R.A. (2009). Heavy tailed processes, in *Encyclopedia of Quantitative Finance*, R. Cont, ed., Wiley, New York.
- [6] Epps, T.W. & Epps, M.L. (1976). The stochastic dependence of security price changes and transaction volumes: implications for the mixture-of-distributions hypothesis, *Econometrica* **44**, 305–321.
- [7] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Wiley, New York, Vol. II.

-
- [8] Fouque, J.-P. (2009). Option pricing with stochastic volatility, in *Encyclopedia of Quantitative Finance*, R. Cont, ed., Wiley, New York.
 - [9] Garcia, R., Ghysels, E. & Renault, E. (2009). Econometrics of option pricing models, in *Handbook of Financial Econometrics*, Y. Ait-Sahalia & L.P. Hansen, eds, Elsevier, North Holland, forthcoming.
 - [10] Geweke, J. & Keane, M. (2007). Smoothly mixing regressions, *Journal of Econometrics* **138**, 252–290.
 - [11] Ghysels, E., Harvey, A. & Renault, E. (1995). Stochastic Volatility, in *Handbook of Statistics 15, Statistical Methods in Finance*, G.S. Maddala & C.R. Rao, eds, North Holland, Amsterdam.
 - [12] Hall, J.A., Brorsen, W. & Irwin, S.H. (1989). The distribution of futures prices: a test of the stable paretian and mixture of normals hypotheses, *Journal of Financial and Quantitative Analysis* **24**, 105–116.
 - [13] Heimstra, C. & Jones, J.D. (1994). Testing for linear and nonlinear granger causality in the stock price-volume relation, *Journal of Finance* **49**, 1639–1664.
 - [14] Hill, J.B. (2008). *Robust Estimation and Inference for Extremal Dependence in Time Series*, Department of Economics, University of North Carolina, Chapel Hill, submitted.
 - [15] Hill, J.B. (2009). On functional central limit theorems for dependent, heterogeneous arrays with applications to tail index and tail dependence estimation, *Journal of Statistical Planning and Inference* **139**, 2091–2110.
 - [16] Karpoff, J.M. (1987). The relation between price changes and trading volume: a survey, *Journal of Financial and Quantitative Analysis* **22**, 109–126.
 - [17] Klüppelberg, C. (2009). Extreme value theory, in *Encyclopedia of Quantitative Finance*, R. Cont, ed., Wiley, New York.
 - [18] Klüppelberg, C. & Lindner, A. (2008). *Extremes of Random Volatility Models*, Department of Mathematics, Universität München, submitted.
 - [19] Li, L.A. & Sedransk, N. (1988). Mixtures of distributions: a topological approach, *Annals of Statistics* **16**, 1623–1634.
 - [20] Mandelbrot, B. (1963). The variation of certain speculative prices, *Journal of Business* **36**, 394–419.
 - [21] Mandelbrot, B. (1973). Comments on: “A subordinated stochastic process model with finite variance for speculative processes” by P.K. Clark, *Econometrica* **41**, 157–159.
 - [22] McDonald, J.B. & Butler, R.J. (1987). Generalized mixture distributions with an application to unemployment duration, *Review of Economics and Statistics* **69**, 232–240.
 - [23] Mykland, P. & Zhang, L. (2008). *Inference for Continuous Semimartingales Observed at High Frequency: A General Approach*, mimeo, University of Chicago.
 - [24] Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result, *American Journal of Mathematics* **8**, 343–366.
 - [25] Praetz, P.D. (1972). The distribution of share price changes, *Journal of Business* **45**, 49–55.
 - [26] Press, J.S. (1967). A compound events model for security prices, *Journal of Business* **40**, 317–335.
 - [27] Renault, E. (2009). Econometrics of option pricing, in *Encyclopedia of Quantitative Finance*, R. Cont, ed., Wiley, New York.
 - [28] Samorodnisky, G. & Taqqu, M. (1994). *Stable Non-Gaussian Random Processes*, Chapman and Hall, New York.
 - [29] Tauchen, G.E. & Pitts, M. (1983). The price variability-volume relationship on speculative markets, *Econometrica* **51**, 485–505.
 - [30] Upton, D.E. & Shannon, D.S. (1979). The stable paretian distribution, subordinated stochastic processes, and asymptotic lognormality: an empirical investigation, *Journal of Finance* **34**, 1031–1039.

Related Articles

Stylized Properties of Asset Returns; Time Change; Time-changed Lévy Process.

JONATHAN B. HILL

Market Microstructure Effects

Most asset pricing models assume that markets are in equilibrium, but not how asset prices move from one equilibrium to another. Market microstructure is the area of economics that studies how prices adjust to new information and how the trading mechanism affects this adjustment process. In a perfect market, new information would be immediately disseminated and interpreted by all market participants. In this full information setting, prices would immediately adjust to a new equilibrium value determined by the agents' preferences and the content of the information. This immediate adjustment, however, is not likely to hold in practice. Not all relevant information is known by all market participants at the same time. Furthermore, information that becomes available is not processed at the same speed by all market participants, implying variable lag time between a news announcement and the agents' realization of price implications. Much of modern microstructure theory is, therefore, driven by models of asymmetric information and the learning process of the market.

In this asymmetric information setting, there are two general areas of interest. First, we can study the process by which the market prices converge to prices that accurately reflect new information. This is termed the *price discovery process*. Learning occurs through time and therefore empirical analysis of this process examines time series properties of asset prices.

The second area of interest studies market quality. In an ideal market, all trades would occur costlessly at a single price where that price accurately reflects all relevant information. In reality, prices for buyers and sellers differ. This difference is determined by the cost of making market (day-to-day operations costs), the risk faced by a market maker that the price will move against their current inventory position, and the risk of trading against better informed agents. All these risks imply that a market maker must be compensated for his/her risk. Higher quality markets will exhibit smaller cost. O'Hara [5] provides a thorough introduction to the economics of market microstructure.

Both these areas, price discovery and market quality, require the study of the fine grain, detailed

price data. In an asymmetric information setting, agents learn about the value of the asset from order flow. A trader with information that says the asset is undervalued will tend to buy the asset and *vice versa*. In this spirit, price impact studies measure how prices adjust to characteristics of order flow. Market quality measures assess the difference between available prices (such as quotes) or trade prices and a notional fair market price. Much of empirical market microstructure uses high-frequency, intraday and often trade-by-trade data.

Here, we focus on econometric issues and models for high-frequency data. We begin with a discussion of the types of data available. Properties of the data are discussed and relevant models are presented.

Data

High-frequency data generally refer to data that are collected at a very rapid rate. The most fundamental data are prices and quantities; however, there might be more than one type of price and more than one type of quantity that can be reported. Many data sets report transaction prices that are the price paid for in a given trade and the quantity (the number of shares) transacted. These are usually referred to as *transaction prices* and *trade size*. The second type of price and quantity data is the limit order book. The limit order book refers to a set of prices and quantities available for sale (on the ask side of the market) and buy (on the bid side of the market). New information arrival might correspond to a change in any of these prices or quantities. These data sets can be quite large. A frequently traded asset such as IBM might have more than 30 000 trades in a single day and significantly many more updates to the limit order book.

In most markets, trades or updates to the limit order book do not occur at regularly spaced intervals, but rather at the pace of the market. Most data sets contain one observation each time that a new piece of information arrives and a time stamp indicating a recorded time at which the transaction or change took place.

The most popular data source for high-frequency US equities is the trades and quotes (TAQ) data set available through Wharton Research Data Services. High-frequency foreign exchange data are becoming more readily available now. One source that has

been around for a while is Olsen and Associates. Other markets such as options and futures markets are slowly becoming available as well. High-frequency data sets are now more readily available for fixed income, such as GovPX, which offers tick-by-tick US treasury prices and volume.

Features of High-frequency Data

Figure 1 presents 2 h of trade-by-trade price data for the stock Airgas (ticker symbol ARG). The horizontal axis is the local time of day. Price data are probably the most fundamental data and there are several things that become immediately apparent. First, the data are not regularly spaced in time. From 10:30 to 11:00 there are about eight trades that occur, but from 11:00 to 11:30 there are many more. Trades occur when individuals make trade decisions, not on a regular schedule.

The second feature that is apparent is that the prices fall on a grid. The smallest increment in the grid is referred to as the *tick size*. US stocks have recently undergone a transition from trading in one-eighth of a dollar to decimalization. This transition was initially tested for 7 NYSE stocks in August 2000 and was completed for the NYSE listed stocks on January 29, 2001. NASDAQ began testing with 14 stocks on March 12, 2001 and completed the transition on April 9, 2001. In June 1997, NYSE permitted 1/16th prices.

When high-frequency data are viewed across multiple days, many relevant variables exhibit diurnal,

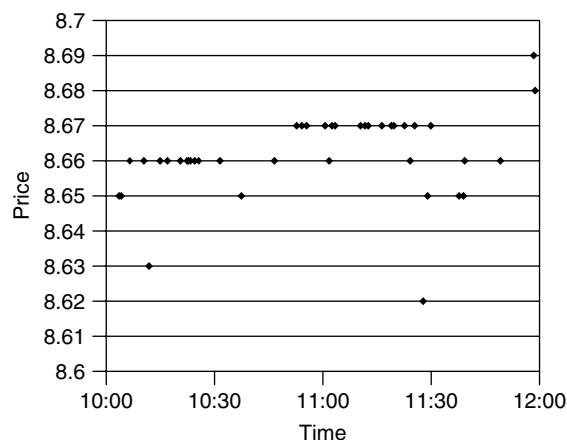


Figure 1 Trade-by-trade transaction prices for airgas

or within day patterns. For example, the volatility of asset price returns over fixed intervals tends to exhibit a U shape in most markets. Volatility is highest in just after the open and just before the close. Other examples of a similar pattern driven by the open and close of the market can be found in bid–ask spreads, trading activity, and traded volume.

Unlike their lower frequency counterparts, market microstructure effects sometimes induce strong correlations in observed, high-frequency returns. Trade-to-trade returns often exhibit negative correlation as the price “bounces” back and forth between trades at the ask price and trades at the bid price.

Some Useful Modeling Approaches

High-frequency data have interesting features to challenge the empirical investigator. In this section, we provide a brief introduction to different modeling approaches that have proven useful. Engle and Russell [3] provide a more detailed discussion of high-frequency data and modeling approaches.

Fixed Interval Models

Most econometric techniques are designed for models with equally spaced data. There is a natural inclination to aggregate irregularly spaced data into the familiar fixed interval setting and apply standard techniques. Returns can be computed over fixed intervals such as 5 min. Since prices are not continuous, there will, in general, not be a price posted exactly at the end of each interval. Hence, a choice will have to be made as to how to define the prices. The prevailing price method uses the last available price before, or at, the end of the interval. Another choice is to interpolate a value for the price given the two prices immediately surrounding the end of the interval. For very frequently traded assets (relative to the interval size), there will generally be a price near the end of the interval, but for short intervals or infrequently traded assets this is not guaranteed. The lack of well-defined prices becomes particularly important when multivariate models are considered. Prices may be stale (old) if the prevailing price method is used so that prices do not reflect recent information if a trade has not occurred in some time, or prices will reflect future information if interpolation is used. Either case can result in erroneous relationships.

Once the prices are aggregated to the fixed interval, then standard techniques can be applied. This temporal aggregation has proven particularly attractive in modeling volatility. Once the spacing of the data is regular, standard volatility models such as GARCH models (*see GARCH Models*) can be applied. Dynamics of the mean can be studied using traditional ARMA or vector autoregression (VAR) time series approaches.

Event Time Models

This class of models indexes the observation in event time. If the data are trade-by-trade data, then these models index time by the trade number. The first observation is the first trade, the second observation is the second trade, and so on. This approach neglects the temporal (clock time) spacing of the data and treats the data series as if the clock that determines the updating to the process is a “trade clock”. If traders update their beliefs about prices each time a trade is made, then there is no role for clock time and this approach appears reasonable. Again, standard time series techniques can be applied to the series such as GARCH models (*see GARCH Models*) or ARMA models (*see Autoregressive Moving Average (ARMA) Processes*).

This approach is not well suited for simultaneously modeling more than one return series, as each will be operating on their own trade time. For this reason, multivariate analysis of return series for more than one asset generally relies on fixed interval approaches. A notable exception is multivariate models for characteristics of a trade (like volume or whether the trade is a buyer or seller initiated) and returns for the same stock. Since both series are updated at the same random time intervals, they both operate on the same time scale. A seminal paper by Hasbrouck [4] uses VAR models for TAQ to study the price impact of a trade on future prices.

Tick time models are also unlikely to be suitable for volatility modeling. Clearly, the longer the time interval between price observations, the larger the volatility should be. While a component of volatility is related to the transaction process and therefore is updated in trade time, another component of volatility is due to the passage of time and potential (nontrade) information.

Point Process Models

A different approach views the data as a point process (*see Point Processes*). A point process is simply a sequence of random arrival times $t_1, t_2, t_3, \dots$, where $t_i \geq t_{i-1}$. These arrival times might correspond to trade times so that t_1 is the time (clock time) that the first trade occurred, t_2 is the time of the second trade, and so on. There may be information associated with each arrival time such as the price or the volume associated with the trade. If r_i denotes the return from the $i - 1$ th trade to the i th trade, then the sequence of arrival times and prices (t_i, r_i) forms a marked point process. Many questions of economic interest can be formulated in the context of the joint conditional distribution $f(t_i, r_i | F_{i-1})$, where F_{i-1} is information available at the time of the $i - 1$ th trade. This distribution completely describes the probability of when and what will happen to the price.

Clearly, the sequence event arrival times can be described by t_i or by the durations between events $x_i = t_i - t_{i-1}$. It may be difficult to directly specify this joint conditional distribution, so Engle [1] proposed decomposing the distribution into $f(x_i, r_i | F_{i-1}) = g(r_i | x_i, F_{i-1})h(x_i | F_{i-1})$, where g is the distribution of the price given the time the event happened (and the history F_{i-1}) and h is the marginal distribution of when the event occurs. Engle and Russell [2] proposed the autoregressive conditional duration (ACD) model for the distribution of the arrival times $h(x_i | F_{i-1})$ (*see Duration Models*). Models for the mark (the price in this case) can then be fit using standard techniques conditional on the duration. The model operates in event time, but as the distribution is conditional on the time since the last trade, calendar time effects can be captured.

Engle [1] proposed a GARCH model for $g(r_i | x_i, F_{i-1})$ where the volatility explicitly depends on the time passed since the last trade. Russell and Engle [6] proposed a model for discrete returns $g(r_i | x_i, F_{i-1})$ that explicitly accounts the time passed since the last trade. Both models suggest that the passage of calendar time is important in models of returns.

The three modeling strategies outlined here (fixed interval, event time, and point processes) comprise the bulk of the approaches but there are others. We do not discuss unobserved component models where the price is modeled as a latent process with bid and ask prices determined by noisy perturbations of a true

price. We also did not touch on an important and growing literature that links market microstructure effects to asset prices. That is, if investors have preferences for assets that can be easily traded, then the ease of trade of the asset may affect the equilibrium price. Finally, we do not discuss optimal execution strategies or models of transaction cost.

References

- [1] Engle, R. (2000). The econometrics of ultra-high frequency data, *Econometrica* **68**(1), 1–22.
- [2] Engle, R. & Russell, J. (1998). Autoregressive conditional duration: a new model for irregularly spaced data, *Econometrica* **66**(5), 1127–1162.
- [3] Engle, R. & Russell, J. (2009). Analysis of high frequency data, Forthcoming in *Handbook of Financial Econometrics*, Y. Ait-Sahalia & L. Hansen, eds, Elsevier.

- [4] Hasbrouck, J. (1991). Measuring the information content of stock trades, *The Journal of Finance* **66**(1), 179–207.
- [5] O'Hara, M. (1995). *Market Microstructure Theory*, Blackwell Publishers.
- [6] Russell, J. & Engle, R. (2005). A discrete-state continuous time model for transaction prices and times: the ACM-ACD model, *Journal of Business Economics and Statistics* **23**, 166–180.

Related Articles

Algorithmic Trading; Bid–Ask Spreads; Duration Models; GARCH Models; Order Flow; Pricing Kernels.

JEFFREY R. RUSSELL

Simulation-based Estimation

Simulation-based estimation methods were first used in microeconomic models involving qualitative or limited dependent endogenous variables and, possibly, lagged endogenous variables (see, in particular, [13, 14]) and in macroeconomic models [11, 12]. In this literature, the role of simulations was to approximate likelihood functions or moments which are not computable exactly, and the origin of the computational problem was often the occurrence of large-size integrals when passing from characteristics of the latent model to those of the observable model.

In financial econometrics, the computational problem comes from the simultaneous presence of dynamics and latent variables. Such a situation may imply, for instance, a likelihood function defined as a multivariate integral the size of which is astronomical. Let us look more precisely at the genesis of the problem.

Many financial econometric models can be written in the form

$$\begin{cases} y_t = r_{1t}(\underline{y}_{t-1}, \underline{y}_t^*, \varepsilon_{1t}; \theta) \\ y_t^* = r_{2t}(\underline{y}_{t-1}, \underline{y}_{t-1}^*, \varepsilon_{2t}; \theta) \end{cases} t = 1, \dots, T \quad (1)$$

where y_t is an observable endogenous vector, y_t^* is an unobservable endogenous vector, θ is an unknown vector, ε_{1t} and ε_{2t} are independent unobservable error white noises (the distributions of which can be assumed to be known without loss of generality), $\underline{y}_t$ is a notation for $(y_t, y_{t-1}, \dots)$.

From system (1) it is, in general, easy to derive the conditional probability density function (pdf):

$$f(y_t / \underline{y}_{t-1}, \underline{y}_t^*; \theta) \\ \text{and} \quad f(y_t^* / \underline{y}_{t-1}, \underline{y}_{t-1}^*; \theta)$$

and, therefore, the joint pdf:

$$f(\underline{y}_T, \underline{y}_T^*; \theta) = \prod_{t=1}^T f(y_t / \underline{y}_{t-1}, \underline{y}_t^*, \theta) f(y_t^* / \underline{y}_{t-1}, \underline{y}_{t-1}^*, \theta) \quad (2)$$

However since $\underline{y}_T^*$ is unobserved the likelihood function is

$$l_T(\underline{y}_T; \theta) = \int f(\underline{y}_T, \underline{y}_T^*; \theta) d\underline{y}_T^* \quad (3)$$

Even if y_T^* is scalar, we are facing an integral of size T , to which the standard numerical methods cannot be applied. For instance, in the simplest case where each y_t^* can take only two values the previous integral is a sum of 2^T terms. If $T = 100$, which is very small in finance, we have 2^{100} terms. So standard numerical procedures are of no use. Note, however, that recursive algorithms are available in some special cases (the Kalman filter in the linear case and the Kitagawa–Hamilton in the case where y_t^* takes a finite number of values and appears in the model with a fixed number of lags).

Examples of Models of Type Governed by Equations (1)

1. Stochastic volatility models

$$y_t = \exp(y_t^*) \varepsilon_{1t} \quad (4)$$

$$y_t^* = \theta_1 + \theta_2 y_{t-1}^* + \varepsilon_{2t} \quad (5)$$

$$\varepsilon_{1t} \sim N(0, 1), \varepsilon_{2t} \sim N(0, \theta_3) \quad (6)$$

2. Factor ARCH model

$$y_t = \theta_1 y_t^* + \varepsilon_{1t}, \quad y_t : \text{multivariate} \\ (\theta_{11} = 1) \quad (7)$$

$$y_t^* = (\theta_2 + \theta_3 y_{t-1}^{*2})^{1/2} \varepsilon_{2t} \quad (8)$$

$$\varepsilon_{1t} \sim N(0, \Sigma(\theta_4)) \quad (9)$$

$$\varepsilon_{2t} \sim N(0, 1) \quad (10)$$

where y_t is a vector and y_t^* a univariate latent ARCH driving the whole dynamics.

3. Diffusion process observed at discrete dates

If we observe the diffusion,

$$d\tilde{y}_t = a(\tilde{y}_t, \theta) + \sigma(\tilde{y}_t, \theta) dW_t \quad (11)$$

(where W_t is a standard Brownian motion) at $t = 1, \dots, T$, we can consider the Euler approximation corresponding to time unit $\frac{1}{n}$ (n is an integer chosen by the econometrician to have a good

approximation):

$$\tilde{y}_\tau = \tilde{y}_{\tau-\frac{1}{n}} + \frac{1}{n}a(\tilde{y}_{\tau-\frac{1}{n}}, \theta) + \frac{1}{\sqrt{n}}\sigma(\tilde{y}_{\tau-\frac{1}{n}}, \theta)\varepsilon_\tau \quad (12)$$

where τ takes the values $\frac{k}{n}$, $k = 1, 2, \dots$ and the sequence ε_τ is a standard Gaussian white noise. The $\tilde{y}_\tau$ are observed only if τ is an integer. Renaming $\tilde{y}_\tau = y_k^*$, $\varepsilon_\tau = \varepsilon_{1k}$, we have

$$y_k^* = y_{k-1}^* + \frac{1}{n}a(y_{k-1}^*, \theta) + \frac{1}{\sqrt{n}} \times \sigma(y_{k-1}^*, \theta)\varepsilon_{1k} \quad (13)$$

$$y_k = y_k^*, \text{ if } k \text{ multiple of } n \\ = 0, \text{ otherwise (by convention)} \quad (14)$$

which is of type (1) (k replacing t).

Simulation Methods

The econometric methods described below are based on different simulation methods, which can be partitioned into three categories.

The first category is made of *unconditional simulations* obtained in the following way. Drawing a sequence $(\tilde{\varepsilon}_{1t}^s, \tilde{\varepsilon}_{2t}^s, t = 1, \dots, T)$ in the known distribution of $(\varepsilon_{1t}, \varepsilon_{2t})$, choosing initial values for y_t and y_t^* , and using equation (1), we can, for any value of θ , compute simulated paths $\{y_t^s(\theta), y_t^{*s}(\theta), t = 1, \dots, T\}$. These simulations are unconditional in the sense that they do not depend on the observed values $y_1, \dots, y_T$.

The second category is made of *sequentially conditional simulations*. Using formulas (2) and (3) we see that if, for a given value of θ , we draw sequentially $y_t^{*s} = 1, \dots, T$ in $f(y_t^*/y_{t-1}, y_{t-1}^*; \theta)$, where y_{t-1} are the observed values, we get an *unbiased simulator* of the likelihood function $f(y_T; \theta)$, namely, $\Pi_{t=1}^T f(y_t/y_{t-1}, y_t^{*s}, \theta)$, since

$$l(y_T; \theta) = E \Pi_{t=1}^T f(y_t/y_{t-1}, y_t^{*s}, \theta) \quad (15)$$

the expectation being taken with respect to the probability with pdf

$\Pi_{t=1}^T f(y_t^{*s}/y_{t-1}, y_{t-1}^{*s}; \theta)$. Therefore, $f(y_T; \theta)$ is the limit when S goes to infinity of

$$\frac{1}{S} \sum_{s=1}^S f(y_t^{*s}/y_{t-1}, y_{t-1}^{*s}; \theta) \quad (16)$$

However, in practice, this convergence is often very slow and must be improved by importance sampling methods [1, 4].

In the third category, we have the *globally conditional simulations*. The aim is to simulate in the global conditional distribution with pdf $f(y_T^*/y_T, \theta)$, for a given θ . Note that the previous sequential drawing does not provide such a global drawing; a sequential procedure giving such a drawing should be based on the sequence of pdfs $f(y_t^*/y_T, y_{t-1}^*; \theta)$ (y_T replacing y_{t-1}), but this is not feasible since $f(y_t^*/y_T, y_{t-1}^*; \theta)$ is not computable. More generally, the problem seems to have no solution since

$$f(y_T^*/y_T; \theta) = \frac{f(y_T, y_T^*; \theta)}{f(y_T; \theta)} \quad (17)$$

and the denominator is not computable. The key remark to solve the problem is that this denominator is constant, for given (y_T, θ) , and therefore we can use simulation techniques that only require us to know the pdf up to a multiplicative constant. The more important techniques satisfying these conditions are Monte Carlo Markov chain (MCMC) techniques like the Hastings–Metropolis algorithm.

In a first method, the likelihood function, which is not easily computable, is replaced by another objective function, the optimization of which gives consistent asymptotically normal estimators. Examples of such a method are the indirect inference (II) method [9], briefly described below, and the method of simulated moments [5], which is a particular case of the II method.

The second kind of methods is based on an approximation of the maximum likelihood estimator. For instance, the simulated likelihood ratio method [2] is based on the identity

$$\frac{l(y_T, \theta)}{l(y_T; \bar{\theta})} = E_{\bar{\theta}} \left[\frac{l(y_T, y_T^*; \theta)}{l(y_T, y_T^*; \bar{\theta})} / y_T \right] \quad (18)$$

for any given $\bar{\theta}$. Therefore a globally conditional simulation $\underline{y}_T^{*s}$ in $f(\underline{y}_T^*/\underline{y}_T; \theta)$ provides the unbiased simulator $\frac{l(\underline{y}_T, \underline{y}_T^{*s}; \theta)}{l(\underline{y}_T, \underline{y}_T^{*s}; \bar{\theta})}$ of $\frac{l(\underline{y}_T; \theta)}{l(\underline{y}_T; \bar{\theta})}$ and the ratio can be consistently approximated by an empirical mean of such unbiased simulations. Another example is the simulated expectation maximization (EM) method [15]. The EM algorithm is based on the maximization in θ at iteration m of $E_{\theta^{(m)}}[\log f(\underline{y}_T, \underline{y}_T^*, \theta)/\underline{y}_T]$; this conditional expectation is, in general, not computable but, again, can be approximated using globally conditional simulations.

The third approach is the Bayesian approach. Given a prior distribution of θ , with pdf $\pi(\theta)$, the Bayesian approach is based on the posterior pdf $\pi(\theta/\underline{y}_T)$. Since

$$\pi(\theta/\underline{y}_T) = \frac{f(\theta, \underline{y}_T)}{f(\underline{y}_T)} = \frac{f(\underline{y}_T; \theta)\pi(\theta)}{f(\underline{y}_T)} \quad (19)$$

the computation of $\pi(\theta/\underline{y}_T)$ seems impossible and, moreover, a simulation in this pdf seems also impossible since it is not even known up to a multiplicative constant.

The key idea here is to consider the joint posterior pdf of θ and $\underline{y}_T^*$ given $\underline{y}_T$, which is

$$f(\theta, \underline{y}_T^*/\underline{y}_T) = \frac{f(\underline{y}_T, \underline{y}_T^*; \theta)\pi(\theta)}{f(\underline{y}_T)} \quad (20)$$

The numerator of equation (20) is now computable and therefore, simulations in this joint distribution can be made, giving approximations of not only $\pi(\theta/\underline{y}_T)$ but also of the *smoothing* distribution $f(\underline{y}_T^*/\underline{y}_T)$. See [3].

Indirect Inference

Let us consider an auxiliary criterion $Q_T(\underline{y}_T, \beta)$, where β is an auxiliary parameter. We assume that, when T goes to infinity, $Q_T(\underline{y}_T, \beta)$ converges to a deterministic function $Q_\infty(\theta_0, \beta)$, and we denote by $b(\theta)$ the function obtained by $b(\theta) = \arg \max_\beta Q_\infty(\theta, \beta)$. The function $b(\cdot)$ is called a *binding function* and is assumed to be injective. The II estimators are obtained in the following way:

- First step Compute

$$\hat{\beta}_T = \arg \max_\beta Q_T(\underline{y}_T, \beta) \quad (21)$$

- Second step Compute

$$\hat{\theta}_{ST}(\Omega) = \arg \min_\theta [\hat{\beta}_T - \hat{\beta}_{ST}(\theta)]' \Omega [\hat{\beta}_T - \hat{\beta}_{ST}(\theta)] \quad (22)$$

where

$$\hat{\beta}_{ST} = \arg \max_\beta \sum_{s=1}^S Q_T[\underline{y}_T^s(\theta), \beta] \quad (23)$$

$\underline{y}_T^s(\theta)$ being an unconditional simulated path corresponding the value θ of the parameter of interest and Ω a symmetric positive definite matrix.

Two asymptotic equivalent versions of $\hat{\theta}_{ST}(\Omega)$ are obtained if in equation (21) $\hat{\beta}_{ST}(\theta)$ is replaced either by

$$\frac{1}{S} \sum_{s=1}^S \beta_T^s(\theta) \quad (24)$$

with $\beta_T^s(\theta) = \arg \max_\beta Q_T[\underline{y}_T^s(\theta), \beta]$ or by

$$\tilde{\beta}_{ST}(\theta) = \arg \max_\beta Q_{ST}[\underline{y}_{ST}^s(\theta), \beta] \quad (25)$$

Note, however, that the latter equivalence requires that

$$\frac{\partial Q_T}{\partial \beta}(\underline{y}_T, \beta) = \sum_{t=1}^T \frac{\partial Q_1(y_t; \beta)}{\partial \beta} \quad (26)$$

Finally we also obtain a class of estimators $\hat{\theta}_{ST}(\Sigma)$, which is globally asymptotically equivalent to the class $\hat{\theta}_{ST}(\Omega)$ [6] by

$$\hat{\theta}_{ST}(\Sigma) = \arg \min_\theta \left(\sum_{s=1}^S \frac{\partial Q_T[\underline{y}_T^s(\theta), \hat{\beta}_T]}{\partial \beta'} \right) \Sigma \left(\sum_{s=1}^S \frac{\partial Q_T[\underline{y}_T^s(\theta), \hat{\beta}_T]}{\partial \beta} \right) \quad (27)$$

A clear advantage of these methods is that the basic model has only been used through the path simulations $y_T^s(\theta)$. Therefore the only requirement made on the model is the possibility of performing these path simulations, which is generally satisfied.

When S is fixed and T goes to infinity the II estimators are consistent and asymptotically normal. The asymptotic variance–covariance matrix depends on Ω and this matrix can be optimally chosen. Note, however, that, if θ and β have the same size, Ω does not matter asymptotically. Specification tests can be based on the optimal values of equation (21) for the optimal matrix Ω , and the classical asymptotic significance tests (score test, Wald test, or a test based on the optimal values of the criteria providing the estimators) are also available.

The general ideas of the II can also be used in other domains: corrections for second-order bias [10], notions of indirect information and indirect identification [8, 9], and encompassing [7].

An important issue of the II methodology is the choice of Q_T and natural candidates are log-likelihood functions of approximations of the model retained or approximations of the log-likelihood function of this model.

References

- [1] Billio, M. & Monfort, A. (1998). Switching state space models likelihood function, filtering and smoothing, *Journal of Statistical Planning and Inference* **68**(1), 64–103.
- [2] Billio, M., Monfort, A. & Robert, C. (1997). *The Simulated Likelihood Ratio Method*. CREST Discussion paper no. 9821.
- [3] Billio, M., Monfort, A. & Robert, C. (1999). Bayesian methods for switching ARMA models, *Journal of Econometrics* **93**, 229–255.
- [4] Danielsson, J. & Richard, J.F. (1995). Accelerated Gaussian importance sampler with application to dynamic latent variable model, in *Econometric Inference using Simulation Techniques*, H. Van Dijk, A. Monfort & B.W. Brown, eds, John Wiley & Sons.
- [5] Duffie, D. & Singleton, K. (1993). Simulated moments estimation of Markov models of asset prices, *Econometrica* **61**, 929–952.
- [6] Gallant, A.R. & Tauchen, G. (1996). Which moments to match? *Econometric Theory* **12**, 657–668.
- [7] Gouriéroux, C. & Monfort, A. (1995). Testing encompassing, and simulating dynamic econometric models, *Econometric Theory* **11**, 195–228.
- [8] Gouriéroux, C. & Monfort, A. (1996). *Simulation Based Econometric Methods*, Oxford University Press.
- [9] Gouriéroux, C., Monfort, A. & Renault, E. (1993). Indirect inference, *Journal of Applied Econometrics* **8**, 85–118.
- [10] Gouriéroux, C., Renault, E. & Touzi, N. (1999). Calibration by simulation for small sample bias correction, in *Simulation Based Inference in Econometrics, Methods and Applications*, R. Mariano, T. Schuermann & M. Weeks, eds, Cambridge University Press.
- [11] Laroque, G. & Salanié, B. (1989). Estimation of multimarket fixed-price models: an application of pseudo-maximum likelihood methods, *Econometrica* **57**, 831–860.
- [12] Laroque, G. & Salanié, B. (1995). Simulation-based estimation of models with lagged latent variables, in *Econometric Inference Using Simulation Techniques*, H. Van Dijk, A. Monfort, B.W. Brown, eds, John Wiley & Sons.
- [13] Lerman, S. & Manski, C. (1981). On the use of simulated frequencies to approximate choice probabilities, in *Structural Analysis of Discrete Data with Econometric Applications*, C. Manski & D. McFadden, eds, MIT Press, Cambridge, MA, pp. 305–319.
- [14] McFadden, D. (1989). A method of simulated moments for estimation of discrete response model without numerical integration, *Econometrica* **57**, 995–1026.
- [15] Shephard, N. (1995). Fitting nonlinear time series with applications to stochastic variance model, in *Econometric Inference Using Simulation Techniques*, H.A. Van Dijk, B.W. Monfort & B.W. Brown, eds, John Wiley & Sons.

Related Articles

Generalized Method of Moments (GMM); Monte Carlo Simulation.

ALAIN MONFORT

Econometrics of Diffusion Models

Many models in finance take the form of stochastic differential equations, where the variable of interest X_t is a vector of asset prices, interest rates, and so on, with dynamics of the form

$$dX_t = \mu(X_t; \theta)dt + \sigma(X_t; \theta)dW_t \quad (1)$$

where W_t is an m -dimensional standard Brownian motion. In a typical parametric situation, the functions μ and σ are known, but not the parameter vector θ , which is to be estimated. In the base case, the data are discrete observations on the process X_t sampled at dates $\Delta, 2\Delta, \dots, n\Delta$, on a time interval $[0, T]$ with $T = n\Delta$ and Δ fixed. The case where Δ is either deterministic and time varying or random (as long as independent from X) introduces no further fundamental difficulties. This article focuses exclusively on the case where Δ need not be small, that is, exclude the vast array of methods specifically designed only for high-frequency data.

A number of methods are available to estimate θ . First maximum-likelihood estimation is discussed, followed by nonlikelihood alternatives.

Maximum-likelihood Estimation

Well-established principles from statistics indicate that in this context maximum-likelihood estimation of θ is the method of choice. One should write down as a function of θ the probability of collecting the observations we have actually obtained, and pick the value of θ that maximizes this function, known as the *likelihood function*. In other words, the resulting estimator of θ is the one that makes it most likely that the model would have produced the data we obtained.

Furthermore, the form of the likelihood function is particularly simple due to the Markovian nature of the model (1). The probability of observing the data $\{X_0, X_\Delta, \dots, X_{n\Delta}\}$, given that the parameter vector is θ , is

$$\begin{aligned} & \mathbb{P}(X_{n\Delta}, X_{(n-1)\Delta}, \dots, X_0; \theta) \\ &= \mathbb{P}(X_{n\Delta} | X_{(n-1)\Delta}, \dots, X_0; \theta) \end{aligned}$$

$$\begin{aligned} & \times \mathbb{P}(X_{(n-1)\Delta}, \dots, X_0; \theta) \\ &= \mathbb{P}(X_{n\Delta} | X_{(n-1)\Delta}; \theta) \\ & \times \mathbb{P}(X_{(n-1)\Delta}, \dots, X_0; \theta) \\ &= \mathbb{P}(X_{n\Delta} | X_{(n-1)\Delta}; \theta) \times \dots \\ & \times \mathbb{P}(X_\Delta | X_0; \theta) \mathbb{P}(X_0; \theta) \end{aligned} \quad (2)$$

where the first equality is simply Bayes' rule and the second is a consequence of Markov property. Repeating the process for each successive observation yields the third equality. The objective is then to maximize over values of θ , this function or equivalently its log

$$\ell_N(\theta) \equiv \sum_{i=1}^n l_X(X_{i\Delta} | X_{(i-1)\Delta}, \Delta; \theta) \quad (3)$$

where $l_X(x | x_0, \Delta; \theta) = \ln p_X(x | x_0, \Delta; \theta)$ and $p_X(x | x_0, \Delta; \theta)$ is the transition density of the process, or the probability that X will go from x_0 to x in the amount of time Δ , when the parameter vector is θ .

All this is indeed exceedingly simple, except for one not so minor detail: with few exceptions, we do not know in closed form the expression of the transition density p_X corresponding to a given model (1)! In finance, the exceptions are the well-known models of Black and Scholes [10] (the geometric Brownian motion $dX_t = \beta X_t dt + \sigma X_t dW_t$), Vasicek [37] (the Ornstein–Uhlenbeck process $dX_t = \beta(\alpha - X_t)dt + \sigma dW_t$) and Cox *et al.* [18] (Feller's square-root process $dX_t = \beta(\alpha - X_t)dt + \sigma X_t^{1/2} dW_t$) which rely on these existing closed-form expressions. The fact that these models have closed-form derivative prices is, of course, not by coincidence, since a derivative price is the expected value of its payoff under the density p_X corresponding to the risk-neutral version of the model. So knowledge of the density in closed form and closed-form option prices is closely related.

In many cases that are relevant in finance, however, the transition function is unknown: see, for example, the models used in [16] ($dX_t = \beta(\alpha - X_t)dt + \sigma X_t dW_t$), [31] ($dX_t = (\alpha X_t^{-(1-\delta)} + \beta)dt + \sigma X_t^{\delta/2} dW_t$), [17]; the more general version of Chan *et al.* [11] ($dX_t = \beta(\alpha - X_t)dt + \sigma X_t^\gamma dW_t$), [14] ($dX_t = (\alpha_0 + \alpha_1 X_t + \alpha_2 X_t^2)dt + (\sigma_0 + \sigma_1 X_t) dW_t$); the affine class of models in [20] and [19] ($dX_t = \beta(\alpha - X_t)dt + \sqrt{\sigma_0 + \sigma_1 X_t} dW_t$); and the nonlinear mean reversion model in [1] ($dX_t = (\alpha_0 + \alpha_1 X_t + \alpha_2 X_t^2 + \alpha_{-1}/X_t)dt + (\beta_0 + \beta_1 X_t + \beta_2 X_t^{\beta_3})dW_t$).

A method to construct accurate approximations in closed form to the unknown transition density $p_X(x|x_0, \Delta; \theta)$ for an arbitrary diffusion model (1) is described later.

Constructing the Likelihood Function

The construction of the function l_X to be used in equation (3) follows [2] (examples and application to interest rate data), [3] (univariate theory), and [4] (multivariate theory). The method proceeds by constructing a closed-form expansion to approximate the transition density p_X , or equivalently l_X directly. While the description below may seem involved, in practice these calculations can be implemented in software such as *Mathematica*. Alternative methods would involve solving numerically a partial differential equation, or employing simulations, neither one delivering a closed-form approximation to l_X . The closed-form method has been shown to be substantially more accurate and faster in practical implementations than the alternatives [28, 29].

Reducibility

Whenever possible, it is advisable to start with a change of variable. A diffusion X is *reducible* if and only if there exists a one-to-one transformation of the diffusion X into a diffusion Y whose diffusion matrix σ_Y is the identity matrix. That is, there exists an invertible function $\gamma(x; \theta)$ such that $Y_t \equiv \gamma(X_t; \theta)$ satisfies the stochastic differential equation:

$$dY_t = \mu_Y(Y_t; \theta)dt + dW_t \quad (4)$$

Every univariate diffusion is reducible, through the simple transformation $Y_t = \int^{X_t} du/\sigma(u; \theta)$. However, not every multivariate diffusion is reducible (a necessary and sufficient condition for reducibility is given in [4]). Whenever a diffusion is reducible, an expansion can be computed for the transition density p_X of X by first computing it for the density p_Y of Y and then transforming Y back into X .

The idea is, in fact, to make two successive transformations $X \mapsto Y \mapsto Z$ such that Z is sufficiently close to a Normal random variable, then construct a sequence for p_Z around a Normal and revert the transformation $Z \mapsto Y \mapsto X$. This will yield an expansion for p_X around a deformed Normal, unless

the volatility function σ is constant (independent of X_t).

When a diffusion is not reducible, the situation is more involved (see the section The Irreducible Case). The expansion for l_Y is of the form

$$l_Y^{(K)}(y|y_0, \Delta; \theta) = -\frac{m}{2} \ln(2\pi \Delta) + \frac{C_Y^{(-1)}(y|y_0; \theta)}{\Delta} + \sum_{k=0}^K C_Y^{(k)}(y|y_0; \theta) \frac{\Delta^k}{k!} \quad (5)$$

The coefficients of the expansion are given explicitly by

$$C_Y^{(-1)}(y|y_0; \theta) = -\frac{1}{2} \sum_{i=1}^m (y_i - y_{0i})^2 \quad (6)$$

$$C_Y^{(0)}(y|y_0; \theta) = \sum_{i=1}^m (y_i - y_{0i}) \int_0^1 \mu_{Yi} \times (y_0 + u(y - y_0); \theta) du \quad (7)$$

and, for $k \geq 1$,

$$C_Y^{(k)}(y|y_0; \theta) = k \int_0^1 G_Y^{(k)}(y_0 + u(y - y_0)|y_0; \theta) u^{k-1} du \quad (8)$$

where $G_Y^{(k)}(y|y_0; \theta)$ is given explicitly as a function of μ_{Yi} and $C_Y^{(j-1)}$, $j = 1, \dots, k$.

Given an expansion for the density p_Y of Y , an expansion for the density p_X of X can be obtained by a direct application of the Jacobian formula:

$$l_X^{(K)}(x|x_0, \Delta; \theta) = -\frac{m}{2} \ln(2\pi \Delta) - D_v(x; \theta) + \frac{C_Y^{(-1)}(\gamma(x; \theta)|\gamma(x_0; \theta); \theta)}{\Delta} + \sum_{k=0}^K C_Y^{(k)}(\gamma(x; \theta)|\gamma(x_0; \theta); \theta) \frac{\Delta^k}{k!} \quad (9)$$

from $l_Y^{(K)}$ given in equation (5), using the coefficients $C_Y^{(k)}$, $k = -1, 0, \dots, K$ given above, and where

$$D_v(x; \theta) \equiv \frac{1}{2} \ln(\text{Det}[v(x; \theta)]) \quad (10)$$

with $v(x; \theta) \equiv \sigma(x; \theta)\sigma^T(x; \theta)$.

The Irreducible Case

In the irreducible case, no such Y exists and the expansion of the log likelihood l_X has the form

$$l_X^{(K)}(x|x_0, \Delta; \theta) = -\frac{m}{2} \ln(2\pi \Delta) - D_v(x; \theta) + \frac{C_X^{(-1)}(x|x_0; \theta)}{\Delta} + \sum_{k=0}^K \times C_X^{(k)}(x|x_0; \theta) \frac{\Delta^k}{k!} \quad (11)$$

The approach is to calculate a Taylor series in $(x - x_0)$ of each coefficient $C_X^{(k)}$, at order j_k in $(x - x_0)$. Such an expansion will be denoted by $C_X^{(j_k, k)}$ at order $j_k = 2(K - k)$, for $k = -1, 0, \dots, K$.

The resulting expansion will then be

$$\tilde{l}_X^{(K)}(x|x_0, \Delta; \theta) = -\frac{m}{2} \ln(2\pi \Delta) - D_v(x; \theta) + \frac{C_X^{(j_{-1}, -1)}(x|x_0; \theta)}{\Delta} + \sum_{k=0}^K \times C_X^{(j_k, k)}(x|x_0; \theta) \frac{\Delta^k}{k!} \quad (12)$$

Such a Taylor expansion was unnecessary in the reducible case: the expressions given above provide the explicit expressions of the coefficients $C_Y^{(k)}$ and then in equation (9) we have the corresponding ones for $C_X^{(k)}$. However, even for an irreducible diffusion, it is still possible to compute the coefficients $C_X^{(j_k, k)}$ explicitly.

For each $k = -1, 0, \dots, K$, the coefficient $C_X^{(k)}(x|x_0; \theta)$ in equation (12) solves an equation of the form

$$f_X^{(k-1)}(x|x_0; \theta) = C_X^{(k)}(x|x_0; \theta) - \sum_{i=1}^m \sum_{j=1}^m \times v_{ij}(x; \theta) \frac{\partial C_X^{(-1)}(x|x_0; \theta)}{\partial x_i} \times \frac{\partial C_X^{(k)}(x|x_0; \theta)}{\partial x_j} - G_X^{(k)}(x|x_0; \theta) = 0 \quad (13)$$

Each function $G_X^{(k)}$, $k = 0, 1, \dots, K$, involves only the coefficients $C_X^{(h)}$ for $h = -1, \dots, k - 1$, so this

system of equation can be utilized to solve recursively for each coefficient at a time. Specifically, the equation $f_X^{(-2)} = 0$ determines $C_X^{(-1)}$, given $C_X^{(-1)}$, $G_X^{(0)}$ becomes known and the equation $f_X^{(-1)} = 0$ determines $C_X^{(0)}$, given $C_X^{(-1)}$ and $C_X^{(0)}$, $G_X^{(1)}$ becomes known and the equation $f_X^{(0)} = 0$ then determines $C_X^{(1)}$, and so on. It turns out that this results in a system of linear equations in the coefficients of the polynomials $C_X^{(j_k, k)}$, so each one of these equations can be solved explicitly in the form of the Taylor expansion $C_X^{(j_k, k)}$ of the coefficient $C_X^{(k)}$, at order j_k in $(x - x_0)$.

Applications: Stochastic Volatility and Term Structure Models

In a stochastic volatility model, the asset price process S_t follows

$$dS_t = (r - \delta)S_t dt + \sigma_1(X_t; \theta) dW_t^Q \quad (14)$$

where r is the riskfree rate, δ is the dividend yield paid by the asset (both taken to be constant only for simplicity), σ_1 denotes the first row of the matrix σ , and Q denotes the equivalent martingale measure [25]. The volatility state variables V_t then follow a stochastic differential equation (SDE) on their own. For example, in the [26] model, $m = 2$ and $q = 1$:

$$dX_t = d \begin{bmatrix} S_t \\ V_t \end{bmatrix} = \begin{bmatrix} (r - \delta)S_t \\ \kappa(\gamma - V_t) \end{bmatrix} dt + \begin{bmatrix} \sqrt{(1 - \rho^2)V_t}S_t & \rho\sqrt{V_t}S_t \\ 0 & \sigma\sqrt{V_t} \end{bmatrix} \times d \begin{bmatrix} W_1^Q(t) \\ W_2^Q(t) \end{bmatrix} \quad (15)$$

The model is completed by the specification of a vector of market prices of risk for the different sources of risk (W_1 and W_2 here), such as

$$\Lambda(X_t; \theta) = \begin{bmatrix} \lambda_1 \sqrt{(1 - \rho^2)V_t} & \lambda_2 \sqrt{V_t} \end{bmatrix}' \quad (16)$$

which characterizes the change of measure from Q back to the physical probability measure P .

Likelihood inference for this and other stochastic volatility models is discussed in [7]. Given a time series of observations of both the asset price, S_t , and a vector of option prices (which, for simplicity,

we take to be call options) C_t , the time series of V_t can then be inferred from the observed C_t . If V_t is multidimensional, sufficiently many options are required with varying strike prices and maturities to allow extraction of the current value of V_t from the observed stock and call prices. Otherwise, only a single option is needed. For reasons of statistical efficiency, we seek to determine the joint likelihood function of the observed data, as opposed to, for example, conditional or unconditional moments. We employ the closed-form approximation technique described above, which yields in closed form the joint likelihood function of $[S_t; V_t]'$. From there, the joint likelihood function of the observations on $G_t = [S_t; C_t]' = f(X_t; \theta)$ is obtained simply by multiplying the likelihood of $X_t = [S_t; V_t]'$ by the Jacobian term J_t :

$$\begin{aligned} \ln p_G(g|g_0, \Delta; \theta) = & -\ln J_t(g|g_0, \Delta; \theta) \\ & + l_X(f^{-1}(g; \theta)|f^{-1} \\ & \times (g_0; \theta); \Delta, \theta) \end{aligned} \quad (17)$$

with l_X obtained as described above.

Another example of practical interest in finance consists of term structure models. A multivariate term structure model specifies that the instantaneous riskless rate r_t is a deterministic function of an m -dimensional vector of state variables, X_t :

$$r_t = r(X_t; \theta) \quad (18)$$

An affine yield model is any model where the short rate (18) is an affine function of the state vector and the risk-neutral dynamics (1) are affine:

$$dX_t = (\tilde{A} + \tilde{B}X_t)dt + \Sigma\sqrt{S(X_t; \alpha, \beta)}dW_t^Q \quad (19)$$

where $\tilde{A}$ is an m -element column vector, $\tilde{B}$ and Σ are $m \times m$ matrices, and $S(X_t; \alpha, \beta)$ is the diagonal matrix with elements $S_{ii} = \alpha_i + X_t' \beta_i$, with each α_i a scalar and each β_i an $m \times 1$ vector, $1 \leq i \leq m$ [19].

It can then be shown that, in affine models, bond prices have the exponential affine form $\exp(-\gamma_0(\tau; \theta) - \gamma(\tau; \theta)'x)$, where $\tau = T - t$ is the bond's time to maturity. Ait-Sahalia and Kimmel [6] consider likelihood inference for affine term structure models on the basis of observations on the bond yields. For other implementations of the method in various contexts and with various datasets [8, 13, 21, 32, 36]. Methods to do small sample bias

corrections, relevant in particular for the speed of mean reversion parameter when near a unit root, are provided by Phillips and Yu [33] and Tang and Chen [35].

Nonlikelihood Methods

Because of the Cramer–Rao lower bound, maximum-likelihood estimation will, at least asymptotically, be efficient and should therefore be the method of choice for estimating θ given that we now have available closed-form, numerically tractable, likelihood expansions. There are, however, a number of alternative methods available to estimate the parameter vector. This discussion will again not include methods appropriate only when the sampling interval Δ tends to 0.

The method of moments is generally not available with standard moments such as the mean, variance, and so on. Hansen and Scheinkman [24] provide moment functions in the form of expectations of the infinitesimal generator that are applied to test functions, while Bibby and Sørensen [9] and Kessler and Sørensen [30] consider other constructions of moment functions.

A different type of moment functions is obtained by simulation. The moment function is the expected value, under the model whose parameters one wishes to estimate, of the score function from an auxiliary model. The idea is that the auxiliary model should be easier to estimate; in particular, the parameters of the auxiliary model appearing in the moment function are replaced by their quasi-maximum likelihood estimates. The parameters of the model of interest are then estimated by minimizing a generalized method of moments (GMM)-like criterion. This method has the advantage of being applicable to a wide class of models. On the other hand, the estimates are obtained by simulations that are computationally intensive and they require the selection of an arbitrary auxiliary model, so the parameter estimates for the model of interest will vary based on both the set of simulations and the choice of auxiliary model. Related implementations of this idea are due to Smith [34], Gouriéroux *et al.* [23], and Gallant and Tauchen [22].

Nonparametric methods are appropriate when the drift and diffusion functions in equation (1) are specified as $\mu(x)$ and $\sigma(x)$ instead of being dependent

upon a finite-dimensional vector of parameters θ . They are particularly useful for the purpose of testing whether a given parametric model for $\mu(x; \theta)$ and $\sigma(x; \theta)$ is appropriate, by comparing the parametric and nonparametric estimates of the corresponding densities.

Let $\pi_X(\cdot, \theta)$ denote the marginal density implied by the parametric model and $p_X(\cdot|\cdot, \cdot, \theta)$ the transition density. Under regularity assumptions, $(\mu(\cdot, \theta), \sigma^2(\cdot, \theta))$ will uniquely characterize the marginal and transition densities over discrete time intervals. For example, the Ornstein–Uhlenbeck process $dX_t = \beta(\alpha - X_t)dt + \gamma dW_t$ generates Gaussian marginal and transitional densities. The square-root process $dX_t = \beta(\alpha - X_t)dt + \gamma X_t^{1/2}dW_t$ yields a Gamma marginal and noncentral chi-squared transitional densities.

More generally, any parametrization of μ and σ^2 corresponds to a parametrization of the marginal and transitional densities. While the direct estimation of μ and σ^2 with discrete data is problematic, the estimation of the densities explicitly takes into account the discreteness of the data. The basic idea in [1] is to use the mapping between the drift and diffusion, on the one hand, and the marginal and transitional densities, on the other, to test the model's specification using densities at the observed discrete frequency (π_X, p_X) instead of the infinitesimal characteristics of the process (μ, σ^2) .

Ait-Sahalia *et al.* [5] develop an alternative specification test for the transition density of the process, based on a direct comparison of the nonparametric estimate of the transition density to the assumed parametric transition function. What makes this new direct testing approach possible is the subsequent development, described above, of closed-form expansions for $p_X(x|x_0, \Delta, \theta)$, which, as already noted, is rarely known in closed form. A complementary approach to this is the one proposed by Hong and Li [27], who use the fact that under the null hypothesis, the random variables $\{P_X(X_{i\Delta}|X_{(i-1)\Delta}, \Delta, \theta)\}$ are a sequence of independent and identically distributed (i.i.d.) uniform random variables [12, 15]. Using for P_X the closed-form approximations described above, they detect the departure from the null hypothesis by comparing the kernel-estimated bivariate density of $\{(Z_i, Z_{i+\Delta})\}$ with that of the uniform distribution on the unit square, where $Z_i = P(X_{i\Delta}|X_{(i-1)\Delta}, \Delta, \theta)$.

Acknowledgments

Financial support from the NSF under grants DMS-0532370 and SES-0850533 is gratefully acknowledged.

References

- [1] Ait-Sahalia, Y. (1996). Testing continuous-time models of the spot interest rate, *Review of Financial Studies* **9**, 385–426.
- [2] Ait-Sahalia, Y. (1999). Transition densities for interest rate and other nonlinear diffusions, *Journal of Finance* **54**, 1361–1395.
- [3] Ait-Sahalia, Y. (2002). Maximum-likelihood estimation of discretely-sampled diffusions: a closed-form approximation approach, *Econometrica* **70**, 223–262.
- [4] Ait-Sahalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusions, *Annals of Statistics* **36**, 906–937.
- [5] Ait-Sahalia, Y., Fan, J. & Peng, H. (2009). Nonparametric transition-based tests for jump-diffusions, *Journal of the American Statistical Association* forthcoming.
- [6] Ait-Sahalia, Y. & Kimmel, R. (2002). *Estimating Affine Multifactor Term Structure Models Using Closed-Form Likelihood Expansions*. Technical Reports, Princeton University.
- [7] Ait-Sahalia, Y. & Kimmel, R. (2007). Maximum likelihood estimation of stochastic volatility models, *Journal of Financial Economics* **83**, 413–452.
- [8] Bakshi, G., Ju, N. & Ou-Yang, H. (2006). Estimation of continuous-time models with an application to equity volatility dynamics, *Journal of Financial Economics* **82**, 227–249.
- [9] Bibby, B.M. & Sørensen, M.S. (1995). Estimation functions for discretely observed diffusion processes, *Bernoulli* **1**, 17–39.
- [10] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.
- [11] Chan, K.C., Karolyi, G.A., Longstaff, F.A. & Sanders, A.B. (1992). An empirical comparison of alternative models of the short-term interest rate, *Journal of Finance* **48**, 1209–1227.
- [12] Chen, S.X., Jiti Gao, J. & Tang, C. (2008). A test for model specification of diffusion processes, *Annals of Statistics* **36**, 167–198.
- [13] Cheridito, P., Filipović, D. & Kimmel, R.L. (2007). Market price of risk specifications for affine models: theory and evidence, *Journal of Financial Economics* **83**, 123–170.
- [14] Constantinides, G.M. (1992). A theory of the nominal term structure of interest rates, *Review of Financial Studies* **5**, 531–552.
- [15] Corradi, V. & Swanson, N.R. (2005). A bootstrap specification test for diffusion processes, *Journal of Econometrics* **124**, 117–148.

- [16] Courtadon, G. (1982). The pricing of options on default free bonds, *Journal of Financial and Quantitative Analysis* **17**, 75–100.
- [17] Cox, J.C. (1975). The constant elasticity of variance option pricing model, *The Journal of Portfolio Management* 15–17, Special issue published in 1996.
- [18] Cox, J.C., Ingersoll, J.E. & Ross, S.A. (1985). A theory of the term structure of interest rates, *Econometrica* **53**, 385–408.
- [19] Dai, Q. & Singleton, K.J. (2000). Specification analysis of affine term structure models, *Journal of Finance* **55**, 1943–1978.
- [20] Duffie, D. & Kan, R. (1996). A yield-factor model of interest rates, *Mathematical Finance* **6**, 379–406.
- [21] Egorov, A.V., Li, H. & Ng, D. (2008). A tale of two yield curves: modeling the joint term structure of dollar and euro interest rates, *Journal of Econometrics* forthcoming.
- [22] Gallant, A.R. & Tauchen, G.T. (1996). Which moments to match? *Econometric Theory* **12**, 657–681.
- [23] Gouriéroux, C., Monfort, A. & Renault, E. (1993). Indirect inference, *Journal of Applied Econometrics* **8**, S85–S118.
- [24] Hansen, L.P. & Scheinkman, J.A. (1995). Back to the future: generating moment implications for continuous-time Markov processes, *Econometrica* **63**, 767–804.
- [25] Harrison, M. & Kreps, D. (1979). Martingales and arbitrage in multiperiod securities markets, *Journal of Economic Theory* **20**, 381–408.
- [26] Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bonds and currency options, *Review of Financial Studies* **6**, 327–343.
- [27] Hong, Y. & Li, H. (2005). Nonparametric specification testing for continuous-time models with applications to term structure of interest rates, *Review of Financial Studies* **18**, 37–84.
- [28] Hurn, A., Jeisman, J. & Lindsay, K. (2007). Seeing the wood for the trees: a critical evaluation of methods to estimate the parameters of stochastic differential equations, *Journal of Financial Econometrics* **5**, 390–455.
- [29] Jensen, B. & Poulsen, R. (2002). Transition densities of diffusion processes: numerical comparison of approximation techniques, *Journal of Derivatives* **9**, 1–15.
- [30] Kessler, M. & Sørensen, M. (1999). Estimating equations based on eigenfunctions for a discretely observed diffusion, *Bernoulli* **5**, 299–314.
- [31] Marsh, T. & Rosenfeld, E. (1982). Stochastic processes for interest rates and equilibrium bond prices, *Journal of Finance* **38**, 635–646.
- [32] Mosburger, G. & Schneider, P. (2005). *Modelling International Bond Markets with Affine Term Structure Models*. Technical Reports, Vienna University of Economics and Business Administration.
- [33] Phillips, P.C. & Yu, J. (2009). Maximum likelihood and Gaussian estimation of continuous time models in finance, in *Handbook of Financial Time Series*, T. Andersen, R. Davis, J.-P. Kreib & T. Mikosch, eds, Springer, New York.
- [34] Smith, A.A. (1993). Estimating nonlinear time series models using simulated vector autoregressions, *Journal of Applied Econometrics* **8**, S63–S84.
- [35] Tang, C.Y. & Chen, S.X. (2009). Parameter estimation and bias correction for diffusion processes, *Journal of Econometrics* forthcoming.
- [36] Thompson, S. (2008). Identifying term structure volatility from the LIBOR-swap curve, *Review of Financial Studies* **21**, 819–854.
- [37] Vasicek, O. (1977). An equilibrium characterization of the term structure, *Journal of Financial Economics* **5**, 177–188.

Related Articles

Markov Processes; Stochastic Volatility Models.

YACINE AÏT-SAHALIA

Extreme Value Theory

Univariate extreme value theory is primarily concerned with the limiting behavior of *large* (respectively, *small*) values of real-valued random observations $X_1, \dots, X_n$ as the *sample size* n increases indefinitely. We denote this by $n \rightarrow \infty$, bearing in mind that, in applications, n may be as small as 10 or 20. In some models, n may also be random, which brings some added technicalities (see, e.g., § 6.2 in [25]). For example, the number n of *insurance claims* $X_1, \dots, X_n$ observed in a time period is often assumed to follow a *Poisson distribution* independent of the claim sequence [3]. In the following, we let $n \geq 1$ be nonrandom, and denote by

$$X_{1,n} \leq \dots \leq X_{n,n} \quad (1)$$

the *order statistics* of $X_1, \dots, X_n$, obtained by sorting the observations in increasing order. Of interest is the study of the *extreme values* $M_n := X_{n,n} = \max\{X_1, \dots, X_n\}$ and $m_n := X_{1,n} = \min\{X_1, \dots, X_n\}$ as $n \rightarrow \infty$. *Multivariate extreme value theory* concerns observations $\mathbf{X}_1, \dots, \mathbf{X}_n$ in $\mathbb{R}^d$, with *extremal vectors* obtained by taking maxima (or minima) coordinate-wise. We start with the univariate case $d = 1$, before briefly considering extremes in higher dimensions, in the section Multivariate Extremes. We note that the multivariate theory of extreme values relies heavily on the notion of *copula*, which has given rise to intensive research in the last decades [12, 45]. An example of its application to *option pricing* is provided in [34].

The study of limit laws for M_n and m_n is motivated by the quest of *realistic probabilistic models* for *extreme values* of large collections of random variables (rv's). This is inspired by the common practice of statistical science, which often uses asymptotic formulae to approximate finite sample distributions. In the setup of estimation based on *sample means*, rates of convergence to limiting Gaussian laws are of order $O(n^{-1/2})$, which is accurate enough for a number of applications, even when $n = 50$ or less. In extreme value theory, one may obtain convergence rates as slow as $O((\log n)^{-1/2})$, in which case, the limiting theory becomes irrelevant unless the sample size n is huge (see, e.g., [29] and § 2.10 in [25]). Unfortunately, other than the *standard limit laws* discussed in the section Extremes of Independent and Identically

Distributed (IID) Sequences, there are not many alternative models adapted to this situation, especially when not much is known about the data structure. One should therefore always remain *extremely cautious* in the conclusions that may be drawn from a blunt application of extreme value models. This drawback (the *curse of small sample sizes*) turns out here to be one of the major practical difficulties to cope with.

There are two great families of statistical problems dealing with extremes, depending on whether the *sample components* $X_1, \dots, X_n$ are observed or not. In the first case, one infers from $X_1, \dots, X_n$ information concerning the law of M_n (or m_n), and, possibly, on extremes of future observations from analog sequences. This may often be reduced to *tail estimation*, where one evaluates the *tail behavior* of the component distributions out of a selected number of extreme order statistics (see the section Tail Estimation). In the second case, $X_1, \dots, X_n$ are generally unknown, and only extremes generated by these hidden quantities are at hand. Extreme value theory is then used for *model building*, and statistical applications rely on observations of random samples of extreme values. In this case, there are many competing methods of *parametric estimation* for the standard extreme-value distributions (refer to [8, 39], Chapter 19–21 in [35] and Chapter 22 in [36]); however, this problem is not discussed here. We mention that these estimation techniques are always difficult to use in practice, especially for small sample sizes. A major application can be found in *reliability theory* and *lifetime analysis*, where the basic historical references are the books of Barlow and Proschan [1], Crowder *et al.* [10], and Kalbfleisch and Prentice [37].

Univariate extreme values has received considerable attention since the pioneering work of Fréchet [24], Fisher and Tippett [22] and Gnedenko [28], from 1925 to 1945. Somewhat later, between 1935 and 1965, Gumbel [30] illustrated the statistical interest of these methods on a number of natural phenomenon (in particular, *floods*). The probabilistic aspects of the theory made serious progress after 1960. In particular, the contributions of de Haan [13] in the years 1970–1985, through a use of the theory of *slow* and *regular variation*, allowed to fully characterize the domains of attraction of extremal limit laws for independent and identically distributed (i.i.d.) sequences [5, 51]. Other remarkable developments include the point process theory related to *records* and *extremal processes*. These are not considered

here, and we refer to [46] and [49] for details. Multivariate extreme value theory was mostly initiated by Geffroy [27], Sibuya [52], and Tiago de Oliveira [54] during 1958–1960. We refer to [7, 15, 16, 47, 53], and the references therein for more recent contributions. A series of books on the subject are available. We may cite, among others, the monographs, surveys and proceedings volumes [4, 8, 9, 13, 14, 19–21, 23, 25, 26, 39, 40, 48, 55], and [57]. We have not attempted to cite more than a tiny fraction of the thousands of papers of interest, and limit ourselves to a few selected references.

Upper and Lower Extremes

Given an integer $k \geq 1$, the k *upper* (respectively, *lower*) sample *extremes* of $X_1, \dots, X_n$, are denoted by

$$M_n^{(1)} := X_{n,n} \geq \dots \geq M_n^{(k)} := X_{n-k+1,n} \\ \text{(respectively } m_n^{(1)} := X_{1,n} \leq \dots \leq m_n^{(k)} := X_{k,n}) \quad (2)$$

for each $n \geq k$. The (sample) *maximum* and *minimum*, are, respectively given by

$$M_n := M_n^{(1)} = X_{n,n} = \max\{X_1, \dots, X_n\} \quad \text{and} \\ m_n := m_n^{(1)} = X_{1,n} = \min\{X_1, \dots, X_n\} \quad (3)$$

These quantities are of major interest in various fields, especially in economics and finance [3, 14, 19, 21]. For example, when $X_1, \dots, X_n$ denote the successive *claims* registered in an *insurance portfolio*, the upper extremes $M_n^{(1)} \geq M_n^{(2)} \geq \dots$, must be closely monitored to evaluate *reinsurance premiums* [3]. Extreme values are also critical in a number of natural phenomenon, especially when $X_1, \dots, X_n$ are measures of *wind speeds*, *sea levels*, *temperature*, or other environmental indicators of interest [6, 8]. Lower extreme values are essential to *life studies* and *reliability* [1]. In the latter application, $X_1, \dots, X_n$ denote the *life spans* of the n critical components of an equipment. Failure occurs (under the *weakest link principle*, see, e.g., p. 25 in [8]) at time $m_n^{(1)} = \min\{X_1, \dots, X_n\}$. All kind of variants of this basic model are at hand (such as the *k-out-of-n rule*, where the equipment fails when k of its components have stopped functioning (see, e.g., p. 25 in [8]).

In the sequel, we concentrate on *upper extremes*. The description of lower extremes is identical after a scale reversal. Namely, the change of $X_1, \dots, X_n$ into $-X_1, \dots, -X_n$ replaces the upper (respectively, lower) extremes of the first sequence, into (-1) times the lower (respectively, upper) extremes of the second sequence.

Extremes of Independent and Identically Distributed (IID) Sequences

In the *standard* (or i.i.d.) *model* of extreme value theory, $X = X_1, X_2, \dots$ are i.i.d. rv's with distribution function (df) $F(x) = \mathbb{P}(X \leq x)$. Here, $\mathbb{P}(A)$ denotes the probability of the event A . Even though this model may be far from reality (because of departures from the i.i.d. assumption caused by *trends*, *dependence between observations* or *nonidentically distributed components*), it is useful as a *starting point* in a number of real-life applications [8, 39]. The fundamental results of this model are captured into the following facts.

We first seek conditions for the maximum $M_n := M_n^{(1)}$ to have a limiting *nondegenerate* distribution (meaning the law of a nonconstant rv) $G(\cdot)$. This is equivalent to the existence of constants $a_n > 0$ and b_n , $n = 1, 2, \dots$, such that, as $n \rightarrow \infty$

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = F^n(a_n x + b_n) \longrightarrow G(x) \quad (4)$$

for almost all x in $\mathbb{R} := (-\infty, \infty)$. The choice of a_n , b_n , and $G(\cdot)$ in equation (4) being not unique, the limiting $G(\cdot)$, when it exists, is only defined through its *type* (two df's $G_1(\cdot)$ and $G_2(\cdot)$ are of the *same type* iff there are constants $A > 0$ and B fulfilling $G_1(Ax + B) = G_2(x)$ for almost all $x \in \mathbb{R}$).

Fact 1 The limit law (4) holds for some $F(\cdot)$, and appropriate *norming sequences* $a_n > 0$ and b_n iff, up to a *type equivalence* (through the replacement of $G(x)$ by $G((x - \mu)/\sigma)$ for some $\sigma > 0$ and μ), $G(\cdot)$ belongs to the class of *extreme value distributions*, collecting the following df's.

- The *Gumbel* df:

$$G(x) = \Lambda(x) := \exp(-e^{-x}) \quad \text{for } x \in \mathbb{R} \quad (5)$$

- The *Fréchet* (class of) distribution function(s), defined, for $\alpha > 0$ by

$$G(x) = \Phi_\alpha(x) := \begin{cases} \exp(-x^{-\alpha}) & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases} \quad (6)$$

- The (reverse) *Weibull* (class of) distribution function(s), defined, for $\alpha > 0$, by

$$G(x) = \Psi_\alpha(x) := \begin{cases} 1 & \text{for } x \geq 0 \\ \exp(-(-x)^\alpha) & \text{for } x < 0 \end{cases} \quad (7)$$

Remark 1 Whenever equation (4) holds, we say that $F(\cdot)$ belongs to the *domain of attraction* of $G(\cdot)$ and write $F \in \mathcal{D}(G)$. An extreme value df, given by (5), (6), or (7), is always of the form $G(x) = \exp(-H(x))$, where $H(x)$ is either exponential (in the Gumbel case), or a power function (in the Fréchet and Weibull cases). In the section, Beyond the Independent and Identically Distributed (IID) Case, this property is extended to nonidentically distributed sequences. In the present setup, (4) holds iff, for all $x \in \mathbb{R}$, as $n \rightarrow \infty$,

$$n(1 - F(a_n x + b_n)) \longrightarrow H(x) \quad (8)$$

Remark 2 In extreme value theory, Fact 1 plays a role of the same importance as the *central limit theorem* (CLT) in “classical” statistics. The CLT leads experimenters to favor the use of *Gaussian* distributions (within the class of *stable* laws) to describe statistics based on *sums* (or *means*) of large numbers of i.i.d. components. In the same spirit, Fact 1, hints that extreme value laws should provide *reasonable candidates* to model the distributions of extremes of large numbers of rv’s. This point is discussed further in the section, Beyond the Independent and Identically Distributed (IID) Case.

Fact 2 The existence of a limiting df $G(\cdot)$, such that $F \in \mathcal{D}(G)$, is, unfortunately, limited to a *relatively small class of smooth distributions* $F(\cdot)$, and applied scientists should not always take it for granted that this property is fulfilled by the data they consider. The characterizations of $\mathcal{D}(\Phi_\alpha)$ and $\mathcal{D}(\Psi_\alpha)$ are simple in terms of the *survivor function* $1 - F(x) = \mathbb{P}(X > x)$ (see, for example, (Φ.1–2) and (Ψ.1–2) below). On the other hand, the characterization

of $\mathcal{D}(\Lambda)$ is a difficult mathematical question, which was only solved, after decades of research, by De Haan [13], Meijzler [43] and Marcus and Pinsky [41], among others, in terms of the *upper quantile function* $U(t) = F^{\text{inv}}(1 - t) = \inf\{x : F(x) \geq 1 - t\}$ for $0 < t < 1$. These results (see equations (11)–(13) below) rely on the notion of *slow variation*. A positive function $L(t)$ defined for $t \geq t_0$ is *slowly varying* (in the neighborhood of infinity) iff, independently of $\lambda > 0$,

$$\lim_{t \rightarrow \infty} \frac{L(\lambda t)}{L(t)} = 1 \quad (9)$$

For example, for $C > 0$ and $r \in \mathbb{R}$, $L(t) = C(\log t)^r$ is slowly varying. Such functions have been used in various branches of mathematics, such as *Tauberian theory* (see, e.g., Chapter V in [56]). Their structure is governed by *Karamata’s theorem* (see [38], and [51]), showing that $L(\cdot)$ fulfills equation (9) iff there exist functions $c(t) \rightarrow c \in (0, \infty)$ and $\epsilon(t) \rightarrow 0$ as $t \rightarrow \infty$, such that, for all $t \geq t_0$,

$$L(t) = c(t) \exp\left(\int_{t_0}^t \frac{\epsilon(u)}{u} du\right) \quad (10)$$

In view of equation (10), we may characterize the domains of attraction of the extreme value laws as follows:

1. $F \in \mathcal{D}(\Phi_\alpha)$ iff F has an infinite upper endpoint $\omega = \sup\{x : F(x) < 1\} = \infty$, and fulfills one of the following equivalent conditions.

(Φ.1) There exists a slowly varying function $L_1(\cdot)$ such that $1 - F(x) = x^{-\alpha} L_1(x)$.

(Φ.2) There exists a slowly varying function $\ell_1(\cdot)$ such that $U(t) = t^{-1/\alpha} \ell_1(1/t)$ ($0 < t < 1$).

The df’s $F(\cdot)$ fulfilling (Φ.1) compose the class of *Pareto-type* laws, which is of major importance in applications of extreme value theory to distributions with infinite upper endpoints.

2. $F \in \mathcal{D}(\Psi_\alpha)$, iff F has a finite upper endpoint $\omega = \sup\{x : F(x) < 1\} < \infty$, and fulfills one of the following equivalent conditions.

(Ψ.1) There exists a slowly varying function $L_2(\cdot)$ such that $1 - F(\omega - x) = x^\alpha L_2(1/x)$ ($x > 0$).

- (Ψ.2) There exists a slowly varying function $\ell_2(\cdot)$ such that $U(t) = \omega - t^{1/\alpha} \ell_2(1/t)$ ($0 < t < 1$).
3. $F \in \mathcal{D}(\Lambda)$ for distributions with arbitrary upper endpoint, iff, the following condition holds.
- (Λ) There exists a slowly varying function $\ell_3(\cdot)$ such that, for each $r > 0$, as $t \downarrow 0$,

$$U(rt) - U(t) = -(1 + o(1))(\log r) \ell_3(1/t) \quad (11)$$

The conditions (Φ.1)–(Φ.2), (Ψ.1)–(Ψ.2) and (Λ), may be combined as follows [13]. The existence of $G \in \{\Phi_\alpha, \Psi_\alpha, \Lambda\}$ such that $F \in \mathcal{D}(G)$ is equivalent to the existence (in $\mathbb{R}$) of the limit, for all $r > 0$ and $s > 0$ with $s \neq 1$,

$$L(r, s) = \lim_{t \downarrow 0} \frac{U(rt) - U(t)}{U(st) - U(t)} \quad (12)$$

Whenever it exists, the limit in equation (12) can only be of the form $L_\gamma(r, s)$, for some $\gamma \in \mathbb{R}$, where

$$L(r, s) = L_\gamma(r, s) := \frac{r^{-\gamma} - 1}{s^{-\gamma} - 1} \quad \text{when } \gamma \neq 0 \quad (13)$$

or

$$L(r, s) = L_0(r, s) := \frac{\log r}{\log s} \quad (14)$$

The constant $\gamma \in \mathbb{R}$ is usually called the *extremal index* pertaining to $F(\cdot)$ (or $U(\cdot)$). We have the following equivalent statements:

$$\gamma > 0 \iff F \in \mathcal{D}(\Phi_\alpha) \quad \text{with } \alpha = 1/\gamma \quad (15)$$

$$\gamma < 0 \iff F \in \mathcal{D}(\Psi_\alpha) \quad \text{with } \alpha = -1/\gamma \quad (16)$$

$$\gamma = 0 \iff F \in \mathcal{D}(\Lambda) \quad (17)$$

Remark 3 As follows from Fact 2, a *serious amount of smoothness* is needed for a df $F(\cdot)$ to belong to $\mathcal{D}(G)$ for some $G(\cdot)$. This is especially the case for $\mathcal{D}(\Lambda)$ and somewhat less for $\mathcal{D}(\Phi_\alpha)$ and $\mathcal{D}(\Psi_\alpha)$. A number of *standard distributions* of the literature fulfill the appropriate conditions, but there are some notable exceptions (in particular for the discrete df's, as mentioned below). Also, one should observe that the *standard* continuous distributions of

statistics [35, 36] are always *extremely smooth*, and this property may not be shared by *real-life* distributions originating from experimental data. Therefore, one should always keep in mind that *the basic assumptions needed for extreme value models are not likely to be fulfilled by real-life observations*. For example, distributions $F(\cdot)$ with $U(1/t)$ slowly varying as $t \rightarrow \infty$ do not necessarily fulfill (Λ), and may very well not belong to $\mathcal{D}(\Lambda)$. This illustrates the fact that *the use of Gumbel laws to model extremal data is often problematic*. For standard distributions, the situation is more clear cut, as follows [35, 36]:

- All *standard continuous* df's $F(\cdot)$ of interest *do belong to $\mathcal{D}(G)$ for some $G \in \{\Lambda, \Phi_\alpha, \Psi_\alpha\}$* . In particular, upper extremes of smooth *exponential-tailed* distributions, such as the exponential, gamma, normal, lognormal, and Weibull laws, belong to the domain of attraction $\mathcal{D}(\Lambda)$ of the Gumbel law. The Fréchet $\mathcal{D}(\Phi_\alpha)$ domain holds for distributions with polynomial-type tails, such as *exact Pareto laws* of the form $1 - F(x) = C(x - a)^{-\alpha}$ for $x \geq x_0 > a$, which are special cases of (Φ.1). Another important example of distributions in the Fréchet domain $\mathcal{D}(\Phi_\alpha)$ is given by the non-Gaussian *stable laws*. The case of the (reverse) Weibull law Ψ_α is somewhat similar to Φ_α , and is not discussed here.
- Almost all *standard discrete* df's $F(\cdot)$ of interest *do not belong to $\mathcal{D}(G)$ for any possible $G \in \{\Lambda, \Phi_\alpha, \Psi_\alpha\}$* . For example, the upper extremes of Poisson, geometric, or negative binomial distributions are not in the domain of attraction of an extreme value law. This disappointing fact brings some obvious practical difficulties, since all possible data sets are given on discrete scales.
- Whenever the upper endpoint $\omega := \sup\{x : F(x) < 1\}$ of the distribution $F \in \mathcal{D}(G)$ is *finite*, the only possible choice for $G(\cdot)$ is either Weibull Ψ_α or Gumbel Λ .
- Whenever the upper endpoint $\omega := \sup\{x : F(x) < 1\}$ of the distribution $F \in \mathcal{D}(G)$ is *infinite*, the only possible choice for $G(\cdot)$ is either Fréchet Φ_α or Gumbel Λ .

Remark 4 Because of Fact 2, the study of upper extremes for *unbounded* distributions may be limited to df's $F(\cdot)$ belonging to the Fréchet and Gumbel domains of attraction, with a nonnegative extremal index $\gamma \geq 0$. In the sequel, we concentrate on this case, which is the most relevant for financial

models. Our point of view would be different if we had been concerned with *reliability theory*, for which the underlying distributions are bounded below by 0 (since the lifetime of a component cannot be negative). For this reason, the (direct or usual) Weibull distributions (including the exponential law) have become reference models for lifetime distributions. These laws correspond to $\gamma \leq 0$. We note that the standard form of a Weibull survival time T is obtained by changing signs in equation (7), so that (after some proper changes of scale and origin) $\mathbb{P}(T > t) = \exp(-t^\alpha)$ for $t > 0$ and some $\alpha > 0$. The special case $\alpha = 1$ yields the *exponential law*, which appears as the foremost example of extreme value law for minima.

Fact 3 The constants a_n and b_n in equation (4) can be chosen as follows. Recall the definition $U(t) = F^{\text{inv}}(1 - t) := \inf\{x : F(x) \geq 1 - t\}$ for $0 < t < 1$ of the *upper quantile function* of F .

- When $F \in \mathcal{D}(\Phi_\alpha)$, we may set $a_n = U(1/n)$ and $b_n = 0$ in equation (4), so that, as $n \rightarrow \infty$,

$$\mathbb{P}\left(\frac{M_n}{a_n} \leq x\right) \longrightarrow \Phi_\alpha(x) = \exp(-x^{-\alpha}) \quad \text{for } x > 0 \quad (18)$$

- When $F \in \mathcal{D}(\Lambda)$, we may set $a_n = U(1/(ne)) - U(1/n)$ and $b_n = U(1/n)$ in equation (4).

Remark 5 As a consequence of the results above, when $F \in \mathcal{D}(\Phi_\alpha)$, we have, for some slowly varying function $L_0(\cdot)$, the convergence in distribution, as $n \rightarrow \infty$

$$\frac{M_n}{n^{1/\alpha} L_0(n)} \xrightarrow{d} \Phi_\alpha \quad (19)$$

This implies a *very fast rate of increase* for M_n as $n \rightarrow \infty$, especially when $0 < \alpha < 1$, in which case, it may be shown (see, e.g., §4.5 in [25]) that the sample sum $X_1 + \dots + X_n$ and maximum $\max\{X_1, \dots, X_n\}$ have the same order of magnitude. On the other hand, when $F \in \mathcal{D}(\Lambda)$, we have, for any choice of $\alpha > 0$ and $L_0(\cdot)$, the convergence in probability, as $n \rightarrow \infty$

$$\frac{M_n}{n^{1/\alpha} L_0(n)} \xrightarrow{\mathbb{P}} 0 \quad (20)$$

This has important practical consequences. In many real-life examples, the data sets are *mixtures*

of rv's with different distributions. For example, in an insurance portfolio, it is often the case that the central part of the claim distribution is well fitted with a *lognormal distribution*. The exceptions, when they occur, are mostly in the upper tails, where one often detects a small proportion of *outliers* [2], which cannot fit into the lognormal model. When such a phenomenon occurs, it should be interpreted as the presence in the data set of a *Pareto-type* component (fulfilling $(\Phi.1)$), by definition, in the domain of attraction of a Fréchet law [50]. What one should do then is to perform some tail-estimation procedure on the outlying data. This, however, is always problematic, especially if the number of outliers (or identified as such) is relatively small. On the other hand, the presence of Pareto-type claims in a portfolio may generate huge losses if they are not properly taken into account.

Fact 4 In the standard model, the limiting behavior of the maximum $M_n = M_n^{(1)}$ governs the joint limiting behavior of any fixed number $k \geq 1$ of upper extremes $M_n^{(1)}, \dots, M_n^{(k)}$. More specifically, for each fixed $k \geq 1$, the existence of a nondegenerate limiting distribution $G(\cdot)$ for the maximum $M_n^{(1)}$ fulfilling equation (4) is equivalent to the existence of a limiting joint distribution for any (nonempty) subset of the $k \geq 1$ upper extremes $M_n^{(1)}, \dots, M_n^{(k)}$. In particular, equation (4) holds if and only if, for each $x_1, \dots, x_k \in \mathbb{R}$, as $n \rightarrow \infty$,

$$\mathbb{P}\left(\frac{M_n^{(1)} - b_n}{a_n} \leq x_1, \dots, \frac{M_n^{(k)} - b_n}{a_n} \leq x_k\right) \longrightarrow G_k(x_1, \dots, x_k) \quad (21)$$

where G_k is a nondegenerate multivariate distribution depending only upon G and $k \geq 1$. Its structure is of major interest, and is detailed below. First, we consider the k th maximum, for an arbitrary (but fixed) $k \geq 1$. Let $G(x) = \exp(-H(x))$ in equation (4). This statement is equivalent to having, for each specified $k \geq 1$ and all $x \in \mathbb{R}$, as $n \rightarrow \infty$,

$$\mathbb{P}\left(\frac{M_n^{(k)} - b_n}{a_n} \leq x\right) \longrightarrow \left\{ \sum_{j=0}^{k-1} \frac{H(x)^j}{j!} \right\} \times \exp(-H(x)) \quad (22)$$

It is convenient to define the inverse function of $H(x) = -\log G(x)$ by $H^{\text{inv}}(t) = \inf\{x : H(x) \geq t\}$

Table 1 Scaling functions for extreme value laws

Extreme value law	$G(x)$	$H(x)$	$H^{\text{inv}}(t)$
Gumbel	$\Lambda(x)$	e^{-x} for $x \in \mathbb{R}$	$-\log t$ for $t > 0$
Fréchet	$\Phi_\alpha(x)$	$x^{-\alpha}$ for $x > 0$	$t^{-1/\alpha}$ for $t > 0$
Weibull	$\Psi_\alpha(x)$	$(-x)^{-\alpha}$ for $x < 0$	$-t^{-1/\alpha}$ for $t > 0$

for $t > 0$. We obtain the particular cases given in Table 1.

In view of Table 1, we may give an explicit form for the limit law in equation (21). Denote by $\omega_1, \omega_2, \dots$ i.i.d. exponential rv's with mean 1. In other words, $\mathbb{P}(\omega_j > y) = e^{-y}$ for $y \geq 0$. Then, for each specified $k \geq 1$, we have the convergence in distribution, as $n \rightarrow \infty$,

$$\left\{ \frac{M_n^{(1)} - b_n}{a_n}, \dots, \frac{M_n^{(k)} - b_n}{a_n} \right\} \xrightarrow{d} \{H^{\text{inv}}(\omega_1), \dots, H^{\text{inv}}(\omega_1 + \dots + \omega_k)\} \quad (23)$$

Remark 6 This result has the important consequence that (in the standard model, subject to $F \in \mathcal{D}(G)$), for each fixed $k \geq 1$, the limiting structure of the k upper extremes depends only upon $G(x) = \exp(-H(x))$ through the nonlinear change of scale $t \rightarrow H^{\text{inv}}(t)$.

Remark 7 The fact that the Gumbel class is a boundary case between the Fréchet and Weibull laws becomes straightforward when $\Lambda, \Phi_\alpha, \Psi_\alpha$ are united into the class of *generalized extreme value* (GEV) laws. The latter, in standard form, depend only upon the *extreme value index* γ (or *shape parameter*) through

$$G_\gamma(x) = \begin{cases} \exp(-(1 + \gamma x)^{1/\gamma}) & \text{for } 1 + \gamma x > 0 \text{ and } \gamma \neq 0 \\ \exp(-e^{-x}) & \text{for } x \in \mathbb{R} \text{ and } \gamma = 0 \end{cases} \quad (24)$$

In the GEV setup, G_γ is of Weibull $\Psi_{-1/\gamma}$ type when $\gamma < 0$, of Fréchet $\Phi_{1/\gamma}$ type when $\gamma > 0$, and Gumbel Λ type when $\gamma = 0$. The representation of extreme value laws through Λ, Φ_α , and Ψ_α with $\alpha > 0$, or through G_γ with $\gamma \in \mathbb{R}$ is a matter of pure convenience, and corresponds to identical

models. In each case, an extreme value distribution is characterized by the extremal index (also called the *shape parameter*) γ , and by centering and scale factors μ and $\sigma > 0$. The general form of $G(\cdot)$ in equation (4) is therefore given by

$$G(x) = G_\gamma \left(\frac{x - \mu}{\sigma} \right) \quad (25)$$

where G_γ is as in equation (24). The GEV model has the advantage of characterizing an extreme value distribution through the triplet of parameters $\gamma \in \mathbb{R}$, $\mu \in \mathbb{R}$, and $\sigma > 0$.

Beyond the Independent and Identically Distributed (IID) Case

The simplest variant of the standard model of the section, Extremes of Independent and Identically Distributed (IID) Sequences, is when $X_1, X_2, \dots$ are independent but nonidentically distributed. Setting $F_j(x) = \mathbb{P}(X_j \leq x)$ for $j = 1, 2, \dots$, the corresponding version of equation (4) is given by

$$\mathbb{P} \left(\frac{M_n - b_n}{a_n} \leq x \right) = \prod_{j=1}^n F_j(a_n x + b_n) \longrightarrow G(x) \quad (26)$$

If no further assumption is made, then the asymptotic theory based upon equation (26) is rather disappointing, since any given df $G(\cdot)$ may be the limit law of $(M_n - b_n)/a_n$ for some suitable b_n and $a_n > 0$, subject to the proper choice of $F_1, F_2, \dots$ [33]. It is natural to impose some additional *uniformity assumptions* (see, e.g., § 3.10 in [25]) implying, in particular, that the limiting law in equation (26) is unaffected by deleting one component of the sequence. If so, setting $G(x) = \exp(-H(x))$, in view of equation (8), we see that equation (26) (subject to the uniformity assumptions) is equivalent to

$$\sum_{j=1}^n (1 - F_j(a_n x + b_n)) \longrightarrow H(x) \quad (27)$$

In this case (see, e.g., § 3.9 in [25]), the class of $H(\cdot)$ on the right-hand side of equation (27) is limited to the functions fulfilling one among the following conditions.

1. $H(\cdot) = -\log G(\cdot)$ is convex.
2. The upper endpoint $\omega_G := \sup\{x : G(x) < \infty\}$ of $G(\cdot)$ is finite and $H(\omega_G - e^{-x})$ is convex.
3. The lower endpoint $a_G := \inf\{x : G(x) > 0\}$ of $G(\cdot)$ is finite and $H(a_G + e^x)$ is convex.

It follows that (1) holds when $G(x) = \Lambda(x)$, since $H(x) = e^{-x}$ is convex. When $G(x) = \Phi_\alpha(x)$, we have $a_G = 0$ and $H(x) = x^\alpha$. In this case, (3) holds, since $H(e^x) = e^{-\alpha x}$ is convex. A similar result holds for $G(x) = \Psi_\alpha(x)$, which is covered by (2).

At this point, we reach the unpleasant conclusion that *some slight variations in the assumption that the sample components are identically distributed may result in extremes with distributions far away from the classical triplet Φ_α , Ψ_α , and Λ of the section Extremes of Independent and Identically Distributed (IID) Sequences*. Recalling Remark (2), we may compare this with the *erroneous belief* (which is quite common among nonexpert statisticians) that most distributions generated by sums of random components should be close to Gaussian laws. Any serious applied scientist knows quite well that this is very often untrue. Moreover, there is no such thing as a *perfect Gaussian sample*, even when it comes from simulated data. In the same spirit, there is no such thing as a *perfect extreme-value sample*, in the sense that, on real-life data sets, Φ_α , Ψ_α , and Λ provide, at best, some rough approximations of some more complex unknown distributions. There is no general recipe available in the scientific literature to *go beyond* extreme value laws, and one must, therefore, proceed with care in the interpretation of data by extreme value laws, in particular, by using *nonparametric methods* when the sample size allows their application.

There is a very important literature on extremes generated by *dependent sequences*. We refer to [25] for details on exchangeable and related models, and to [40] for a general treatment of extremes generated by various types of stochastic processes, such as stationary sequences. As far as financial time series are concerned, the recent review [44] is most useful. It turns out that, under rather weak *tail-mixing conditions*, measuring the asymptotic independence of random components, subject to the fact that they

are simultaneously large and with indices far from each other, most of the results of the standard case still apply. This is even true for *tail estimators* [32]. Extremes generated by *continuous processes* tend to be much more complex than those that we have been considering up to now for integer-indexed sequences. For this reason, this question is not discussed here. The case of *Gaussian processes* has been extensively explored, and we refer to [40] for an advanced initiation to this problem.

Tail Estimation

We limit ourselves to the case where the upper endpoint $\omega = \sup\{x : F(x) < 1\}$ of the df F is infinite ($\omega = \infty$), and where F is in the domain of attraction of an (upper) extreme value distribution. As discussed in the section Upper and Lower Extremes we have, either $F \in \mathcal{D}(\Lambda)$, or $F \in \mathcal{D}(\Phi_\alpha)$ for some $\alpha > 0$. Since the Gumbel law Λ is the limit of Φ_α when $\alpha \rightarrow \infty$, we set, for convenience, $\Lambda = \Phi_\infty$, and assume that $F \in \mathcal{D}(\Phi_\alpha)$ for some $\alpha \in (0, \infty]$. The problem is then to estimate the *regular variation index* $\alpha > 0$, or equivalently, the *tail index* $\gamma = 1/\alpha \in [0, \infty)$, out of $X_1, \dots, X_n$. In the following, we restrict our study to $\gamma \in (0, \infty)$, for Pareto-type laws, fulfilling, as $x \rightarrow \infty$,

$$1 - F(x) = x^{-1/\gamma} L_1(x) = x^{-1/\gamma} L_1(x) \quad (28)$$

where $L_0(\cdot)$ is a slowly varying function. The modern approach to this problem was initiated in 1975 by Hill [31], who introduced the estimator

$$\begin{aligned} \hat{\gamma}_n &= \frac{1}{k} \sum_{j=1}^k \left\{ \log X_{n-j+1,n} - \log X_{n-k,n} \right\} \\ &= \sum_{j=1}^k \frac{j}{k} \left\{ \log X_{n-j+1,n} - \log X_{n-j,n} \right\} \end{aligned} \quad (29)$$

where $1 \leq k = k_n < n$ is an integer sequence varying with n . The use of $\hat{\gamma}_n$ is motivated by its *optimal character* as an estimator of γ based upon the k upper extremes $\{X_{n-j+1,n} : 1 \leq j \leq k+1\}$, in the particular case where $F(x) = 1 - x^{-\alpha}$ for $x \geq 1$, follows an *exact* Pareto law. Moreover, the *Hill estimator* remains consistent under equation (28) subject to minimal conditions imposed upon k_n . This was proved in 1982 by Mason [42]

(see also Deheuvels *et al.* [17]), who showed that the conditions $k = k_n \rightarrow \infty$ and $k_n/n \rightarrow 0$ are *necessary and sufficient* for the convergence (in probability) of γ_n to γ as $n \rightarrow \infty$ for all $F \in \mathcal{D}(\Phi_{1/\gamma})$. For an *exact Pareto law*, these conditions also imply the asymptotic normality of $\hat{\gamma}_n$, which fulfills the convergence in distribution (with $N(0, 1)$ denoting the standard normal law), as $n \rightarrow \infty$,

$$\sqrt{k_n} \left\{ \frac{\hat{\gamma}_n}{\gamma} - 1 \right\} \xrightarrow{d} N(0, 1) \quad (30)$$

Unfortunately, with the exception of the special case of exact Pareto laws, the Hill estimator is always *biased*, and problematic to use. Besides the choice of the origin (to allow the definition of the logarithms in equation 29), the selection of $k = k_n$ is quite delicate. This was illustrated in a number of papers where the performance of the Hill estimator was investigated on simulated and real data sets [3, 50], with different choices of k . At times, the results are so poor that these graphs have been called *Hill horror plots*. Hill's estimator as well as a number of alternative estimators of γ fall into the general class of *Kernel estimators* introduced in 1985 by Csörgő *et al.* [11]. These require a nonincreasing kernel $K(t) \geq 0$ on $[0, 1]$ and are defined by

$$\begin{aligned} \tilde{\gamma}_n = & \left(\sum_{j=1}^k \frac{j}{k} K\left(\frac{j}{k}\right) \{\log X_{n-j+1,n} - \log X_{n-j,n}\} \right) \\ & \times \left(\sum_{j=1}^k \frac{1}{k} K\left(\frac{j}{k}\right) \right)^{-1} \end{aligned} \quad (31)$$

Hill's estimator is the special case of equation (31) obtained by choosing $K(t) = 1$ for $0 \leq t \leq 1$. As seen from the results of [11], the optimality of Hill's estimator is restricted to exact Pareto laws. An interesting application is given when $\tilde{\gamma}_n$ is applied to non-Gaussian *stable laws*. It turns out that in this case, the optimal choice of $K(\cdot)$ is given by $K(t) = 1 - t$ for $0 \leq t \leq 1$, and the optimal choice of k_n is of the form $Rn^{1/3}$ for some appropriate constant R . This illustrates again the *curse of small sample sizes*. For example, when $n = 1000$, which is a rather large sample, we get $n^{1/3} = 10$. It is then rather difficult to admit that such a small value should fall within the range of $k_n \rightarrow \infty$.

In applied science, tail estimation is a mix between know-how and advanced mathematics. The latter provides nice limit laws, with assumptions which are, more or less, never really fulfilled by the data of interest. At this point, an expert opinion becomes closer to a fancy guess than a seriously established scientific fact. We omit a detailed discussion of the various methods in use [3, 8, 14, 18, 19, 21, 39].

Multivariate Extremes

For convenience, we limit our exposition to an i.i.d. sequence $(X_1, Y_1), \dots, (X_n, Y_n)$ of random vectors in $\mathbb{R}^2$. We are concerned with the limiting behavior of $(U_n, V_n) := (\max\{X_1, \dots, X_n\}, \max\{Y_1, \dots, Y_n\})$ as $n \rightarrow \infty$, assuming the existence of sequences of constants $a'_n > 0$, $a''_n > 0$, b'_n and b''_n such that, as $n \rightarrow \infty$,

$$\mathbb{P}\left(\frac{U_n - b'_n}{a'_n} \leq x, \frac{V_n - b''_n}{a''_n} \leq y\right) \longrightarrow G(x, y) \quad (32)$$

Here, $G(\cdot, \cdot)$ denotes a bivariate df with nondegenerate marginal df's $K_1(x) = G(x, \infty)$ and $K_2(y) = G(\infty, y)$. This implies that K_1 and K_2 belong to the class $\{\Phi_\alpha, \Psi_\alpha, \Lambda : \alpha > 0\}$ of extreme value laws. This, however, is not enough. The *copula function* $C(\cdot, \cdot)$ of $G(\cdot, \cdot)$ defined through the identity

$$G(x, y) = C(K_1(x), K_2(y)) \quad (33)$$

must belong to the class of *extreme value copulas*. As shown by Pickands [47], the characterization of a bivariate extreme value copula reduces to a convex function $\{D(x) : x \in [0, 1]\}$ fulfilling the inequalities $\max\{x, 1 - x\} \leq D(x) \leq 1$. Several nonparametric estimation procedures have been proposed ([7, 16, 53, 57], and, for example, Chapter 3 [39]) to estimate $D(\cdot)$, which is often called the *dependence function* (even though this denomination is ambiguous, having been used with other meanings) of $G(\cdot, \cdot)$. This is a very active research field at the present time. The situation in higher dimensions ($d \geq 3$) is even more complex, and has not yet led to very practical solutions.

Conclusion

Extreme value theory is a fascinating mathematical domain, which is often rather difficult to apply to real-life experiments, mostly because it relies on *tail properties* of the underlying distributions, for which very few observations are available. One should therefore always keep in mind, in this domain, the general principle that *statistical conclusions are always limited by the information carried by the data itself*. When this information is sparse or not available, scientists should remain careful about its interpretation.

References

- [1] Barlow, R.E. & Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing: Probability Models*, Holt, Rinehart and Winston, New York.
- [2] Barnett, V. & Lewis, T. (1995). *Outliers in Statistical Data*, 3rd Edition, Wiley, New York.
- [3] Beirlant, J., Teugels, J.L. & Vynckier, P. (1994). Extremes in non-life insurance, in *Extreme Value Theory and Applications*, J. Galambos, J. Lechner, E. Simiu & N. Macri, eds, Kluwer, Dordrecht, pp. 489–510.
- [4] Beirlant, J., Teugels, J.L. & Vynckier, P. (1996). *Practical Analysis of Extreme Values*, Leuven University Press, Leuven.
- [5] Bingham, N.H., Goldie, C. & Teugels, J. (1987). *Regular Variation*, Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, Vol. 27.
- [6] Buishand, T.A. (1989). Statistics of extremes in climatology, *Statistica Neerlandica* **43**, 1–30.
- [7] Capérea, P. & Fougères, A.-L. (2000). Estimation of a bivariate extreme value distribution, *Extremes* **3**, 311–329.
- [8] Castillo, E. (1988). *Extreme Value Theory in Engineering*, Academic Press, New York.
- [9] Coles, S.G. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer, New York.
- [10] Crowder, M.J., Kimber, A.C., Smith, R.L. & Sweeting, T.J. (1991). *Statistical Analysis of Reliability Data*, Chapman & Hall, London.
- [11] Csörgő, S., Deheuvels, P. & Mason, D. (1985). Kernel estimates of the tail index of a distribution, *Annals of Statistics* **13**, 1050–1077.
- [12] Dall'Aglio, G. (1991). *Advances in Probability Distributions with given Marginals*, Kluwer, Dordrecht.
- [13] De Haan, L. (1970). *On Regular Variation and Its Application to the Weak Convergence of Sample Extremes*, Mathematical Centre Tracts, Mathematisch Centrum, Amsterdam, Vol. 32.
- [14] De Haan, L., De Vries, C.G., Hüsler, J. & Smith, W.B. (1996). Multivariate extreme value estimation with applications to economics and finance, *Communications in Statistics – Theory and Methods* **25**(4), 685–908.
- [15] Deheuvels, P. (1978). Caractérisation complète des lois extrêmes multivariées et de la convergence des types extrêmes. *Publications de l'Institut de Statistique de l'Université de Paris* **23**, 1–36.
- [16] Deheuvels, P. (1991). On the limiting behavior of the Pickands estimator for bivariate extreme value distributions, *Statistics and Probability Letters* **12**, 429–439.
- [17] Deheuvels, P., Haeusler, E. & Mason, D.M. (1988). Almost sure convergence of the Hill estimator, *Mathematical Proceedings of the Cambridge Philosophical Society* **104**, 371–381.
- [18] Dekkers, A.L.M., Einmahl, J.H.J. & De Haan, L. (1989). A moment estimator for the index of an extreme value distribution, *Annals of Statistics* **17**, 1833–1855.
- [19] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.
- [20] Epstein, N. (1960). Elements of the theory of extreme values, *Technometrics* **2**, 27–41.
- [21] Finkenstädt, B. & Rootzén, H. (2004). *Extreme Values in Finance, Telecommunications, and the Environment*, Chapman & Hall/CRC, Boca Raton.
- [22] Fisher, R.A. & Tippett, L.H.C. (1928). Limiting forms of the frequency distributions of largest or smallest member of a sample, *Mathematical Proceedings of the Cambridge Philosophical Society* **24**, 180–190.
- [23] Fougères, A.-L. (2004). Multivariate extremes, in *Extreme Values in Finance, Telecommunications, and the Environment*, B. Finkenstädt & H. Rootzén, eds, Chapman & Hall/CRC, Boca Raton, pp. 373–388.
- [24] Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Annales de la Société Mathématique Polonaise de Cracovie (Kraków)*, **6**, 93–116.
- [25] Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*, 2nd Edition, Krieger, Dordrecht.
- [26] Galambos, J., Lechner, J. & Simiu, E. (1994). *Extreme Value Theory and Applications*, Kluwer, Dordrecht.
- [27] Geffroy, J. (1958–59). Contributions à la théorie des valeurs extrêmes. I,II, *Publications de l'Institut de Statistique de l'Université de Paris* **7,8**, 37–121, 124–184.
- [28] Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire, *Annals of Mathematics* **44**, 423–453.
- [29] Gomes, M.I. (1984). Penultimate limiting forms in extreme value theory, *Annals of the Institute of Statistical Mathematics* **36**, 71–85.
- [30] Gumbel, E.J. (1958). *Statistics of Extremes*, Columbia University Press, New York.
- [31] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics* **3**, 1163–1174.
- [32] Hsing, T. (1991). On tail index estimation using dependent data, *Annals of Statistics* **19**, 1547–1569.
- [33] Hüsler, J. (1994). Extremes: limit results for univariate and multivariate nonstationary sequences, in *Extreme*

- Value Theory and Applications*, J. Galambos, J. Lechner, E. Simiu & N. Macri, eds, Kluwer, Dordrecht, pp. 283–304.
- [34] Hüsler, J. (1996). Multivariate option pricing and extremes, in *Multivariate Extreme Value Estimation with Applications to Economics and Finance*, W.B. Smith, J. Husler, L. DeHaan & C.G. DeVries, eds, Communications in Statistics, Special Issue, pp. 853–870, Vol. 25.
- [35] Johnson, N.L., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions*, 2nd Edition, Wiley, New York, Vol. 1.
- [36] Johnson, N.L., Kotz, S. & Balakrishnan, N. (1995). *Continuous Univariate Distributions*, 2nd Edition, Wiley, New York, Vol. 2.
- [37] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.
- [38] Karamata, J. (1933). Sur un mode de croissance régulière. Théorèmes fondamentaux, *Bulletin de la Société Mathématique de France* **61**, 55–62.
- [39] Kotz, S. & Nadarajah, S. (2000). *Extreme Value Distributions – Theory and Applications*, Imperial College Press, London.
- [40] Leadbetter, M.R., Lindgren, G. & Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer, New York.
- [41] Marcus, M.B. & Pinsky, M.A. (1969). On the domain of attraction of $\exp(-e^{-x})$, *Journal of Mathematical Analysis and Applications* **28**, 440–449.
- [42] Mason, D.M. (1982). Laws of large numbers for sums of extreme values, *Annals of Probability* **10**, 754–764.
- [43] Meizler, D.G. (1949). On a theorem of B.V. Gnedenko (Russian). *Sbornik Nauchnik Trudov Institut Akademiyi Nauk Ukrainskoyi SSR*, **12**, 31–35.
- [44] Mikosch, T. (2004). Modeling dependence and tails of financial time series, in *Extreme Values in Finance, Telecommunications, and the Environment*, B. Finkenstadt & H. Rootzen, eds, Chapman & Hall/CRC, Boca Raton, pp. 187–278.
- [45] Nelsen, R.B. (1998). *An Introduction to Copulas*, Springer, New York.
- [46] Nevzorov, V.B. (2001). *Records: Mathematical Theory*. Translation of Mathematical Monographs, American Mathematical Society, Providence. Vol. 194.
- [47] Pickands, J. (1981). Multivariate extreme value distributions. *Proceedings of the 43d ISI Session, Buenos Aires*, 859–878.
- [48] Reiss, R.D. & Thomas, M. (1997). *Statistical Analysis of Extreme Values*, Birkhäuser, Basel.
- [49] Resnick, S. (1987). *Extreme Values, Regular Variation and Point Processes*, Springer, New York.
- [50] Resnick, S. (2004). Modeling data networks, in *Extreme Values in Finance, Telecommunications, and the Environment*, B. Finkenstadt & H. Rootzen, eds, Chapman & Hall/CRC, Boca Raton, pp. 289–361.
- [51] Seneta, E. (1976). *Regularly Varying Functions*, Lecture Notes in Mathematics, Springer, Berlin, Vol. 508.
- [52] Sibuya, M. (1960). Bivariate extremal statistics, *Ann. Inst. Statist. Math.* **11**, 195–210.
- [53] Tawn, J.A. (1988). Bivariate extreme value theory: models and estimation, *Biometrika* **75**, 397–415.
- [54] Tiago de Oliveira, J. (1958). Extremal distributions, *Revista da Faculdade Ciências. Lisboa* **2**, Ser. A Math., 7, 215–227.
- [55] Tiago de Oliveira, J. (1984). *Statistical Extremes and Applications*, Reidel, Dordrecht.
- [56] Widder, D.V. (1972). *The Laplace Transform*, Princeton University Press, Princeton.
- [57] Xin, H. (1992). *Statistics of Bivariate Extreme Values*. Tinbergen Institute Research Series, Tinbergen.

Related Articles

Copulas in Econometrics; Heavy Tails; Moment Explosions; Risk Measures; Statistical Estimation; Stylized Properties of Asset Returns.

PAUL DEHEUVELS

GARCH Models

Univariate Models of Conditional Heteroskedasticity

The ARCH Model

Financial economists are concerned with modeling and forecasting volatility in asset returns. This is the case since volatility is considered a measure of risk, and investors want a premium for investing in risky assets. Marginal distributions of typical return series have heavy tails, and the series show clustering of volatility. Furthermore, these series have little or no autocorrelation, but they do display higher order dependence. For example, the squared returns are positively autocorrelated. The autoregressive conditional heteroskedasticity (ARCH) model and its variants have been found useful in describing data with these properties.

The original ARCH model, introduced by Engle [21], can be defined as follows. Let ε_t be a random variable that has a mean and a variance conditionally on the information set $\mathcal{F}_{t-1}$ that contains all past information up until $t - 1$ and available at time t . The ARCH model of ε_t has the following properties. First, the conditional mean $\mathbf{E}\{\varepsilon_t|\mathcal{F}_{t-1}\} = 0$ and, second, the conditional variance $h_t = \mathbf{E}\{\varepsilon_t^2|\mathcal{F}_{t-1}\}$ is a nontrivial positive-valued parametric function of $\mathcal{F}_{t-1}$. The sequence $\{\varepsilon_t\}$ may be observed directly. In that case it may be, for example, a sequence of daily returns of an asset or an index. It may also be an unobserved error or innovation sequence of an econometric model. In the latter case,

$$\varepsilon_t = y_t - \mathbf{E}\{y_t|\mathcal{F}_{t-1}\} \quad (1)$$

where y_t is an observable random variable and $\mathbf{E}\{y_t|\mathcal{F}_{t-1}\}$ is the conditional mean of y_t given $\mathcal{F}_{t-1}$. Here, the focus is on h_t and it is assumed that $\mathbf{E}\{y_t|\mathcal{F}_{t-1}\} = 0$.

Engle assumed that ε_t can be decomposed as follows:

$$\varepsilon_t = z_t h_t^{1/2} \quad (2)$$

where $\{z_t\}$ is a sequence of independent, identically distributed (iid) random variables with zero mean and unit variance. This implies $\varepsilon_t|\mathcal{F}_{t-1} \sim \mathcal{D}(0, h_t)$ where $\mathcal{D}$ stands for a distribution, assumed to be

either a normal distribution or a leptokurtic one, such as the t -distribution. The information set $\mathcal{F}_{t-1} = \{\varepsilon_{t-j} : j \geq 1\}$. The conditional variance of an ARCH model of order q has the following parametric form:

$$h_t = \alpha_0 + \sum_{j=1}^q \alpha_j \varepsilon_{t-j}^2 \quad (3)$$

where $\alpha_0 > 0$, $\alpha_j \geq 0$, $j = 1, \dots, q - 1$, and $\alpha_q > 0$.

The parameter restrictions in equation (3) form a necessary and sufficient condition for positivity of the conditional variance. Suppose the unconditional variance $\mathbf{E}\varepsilon_t^2 = \sigma^2 < \infty$. The definition of ε_t through the decomposition (2) involving z_t then guarantees the white noise property of the sequence $\{\varepsilon_t\}$, since $\{z_t\}$ is a sequence of iid variables. In fact, the ARCH model can be viewed as a special case of the family of models with uncorrelated observations but higher order dependence, such that the squared observations are correlated.

The ARCH model and its generalizations are applied to modeling, among other things, high-frequency data on interest rates, exchange rates and stock and stock index returns. “High frequency” in this context means weekly, daily, or intradaily observations. Bollerslev *et al.* [12] already listed a large variety of applications in their survey of these models, and their number has increased enormously since then. Forecasting volatility of these series is different from forecasting the conditional mean of a process because volatility, the object to be forecast, is not observed. The question then is how volatility should be measured. Using ε_t^2 is an obvious solution, but it is not necessarily a good one if data of higher frequency are available. Andersen and Bollerslev [3] and others recommend *realized volatility* (see **Realized Volatility and Multipower Variation**) for the purpose; for a review see [2].

The Generalized ARCH Model and Extensions

Generalized ARCH. The ARCH model has since been extended by adding lagged conditional variances to equation (3). This extension, called the *generalized ARCH (GARCH) model*, proposed independently by Bollerslev [10] and Taylor [67], has since become very popular. The GARCH(p, q) model has the

following form:

$$h_t = \alpha_0 + \sum_{j=1}^q \alpha_j \varepsilon_{t-j}^2 + \sum_{j=1}^p \beta_j h_{t-j} \quad (4)$$

A sufficient condition for the conditional variance h_t in equation (4) to be positive is $\alpha_0 > 0$, $\alpha_j \geq 0$, $j = 1, \dots, q$; $\beta_j \geq 0$, $j = 1, \dots, p$. This condition is not necessary if either p or q , or both, exceed 1; see [56]. For the GARCH model to be identified if at least one $\beta_j > 0$ (the model is a genuine GARCH model), it is required that also at least one $\alpha_j > 0$. If $\alpha_1 = \dots = \alpha_q = 0$, the conditional and unconditional variances of ε_t are equal and $\beta_1, \dots, \beta_p$ are unidentified nuisance parameters. The GARCH(p, q) process is weakly stationary if and only if $\sum_{j=1}^q \alpha_j + \sum_{j=1}^p \beta_j < 1$.

The overwhelmingly most frequently applied GARCH model has been the GARCH(1,1) model. The weakly stationary GARCH(1,1) model only contains three parameters, but one more parameter can be saved by “variance targeting”. The idea is to replace the intercept α_0 in equation (4) by $(1 - \alpha_1 - \beta_1)\sigma^2$ where $\sigma^2 = E\varepsilon_t^2$. The estimate $\hat{\sigma}^2 = T^{-1} \sum_{t=1}^T \varepsilon_t^2$ is substituted for σ^2 before estimating the other two parameters. As a result, the conditional variance converges toward the “long-run” unconditional variance, which may be considered a desirable property of the model.

Since its introduction, the GARCH model has been generalized and extended in various directions. This has been done in order to increase the flexibility of the original model. Because it only contains squared lags of ε_t , the response of the conditional variance to a shock ε_t is independent of the sign of the shock and just a function of its size. Several extensions of the GARCH model aim at accommodating the asymmetry in the response, caused by the *leverage effect* present in stock returns. The most frequently applied asymmetric GARCH model is the GJR-GARCH model of [32]. It is defined as follows:

$$h_t = \alpha_0 + \sum_{j=1}^q \{\alpha_j + \delta_j I(\varepsilon_{t-j} > 0)\} \varepsilon_{t-j}^2 + \sum_{j=1}^p \beta_j h_{t-j} \quad (5)$$

where $I(\varepsilon_{t-j} > 0)$ is an indicator function obtaining value 1 when the argument is true and 0 otherwise. The asymmetry in the data is modeled through the terms $\delta_j I(\varepsilon_{t-j} > 0) \varepsilon_{t-j}^2$, $j = 1, \dots, q$, in equation (5). Typically, in applications, $p = q = 1$. Other asymmetric GARCH models include the asymmetric GARCH (AGARCH) of [27] and the quadratic GARCH of [59]. Furthermore, the smooth transition GARCH (STGARCH) model, introduced by Hagerud [38] and Gonzalez-Rivera [33], can be used for describing asymmetries in return series. This model contains a continuum of regimes, whereas in the GJR-GARCH(1,1) model there are only two regimes with the switch point at $\varepsilon_t = 0$. The GJR-GARCH model is still linear in parameters, which makes it easier to estimate than, say, the STGARCH model that is genuinely nonlinear in parameters.

There are other nonlinear GARCH models that have been applied in practice. Perhaps the most prominent example is the Markov-switching GARCH model. It contains a finite number (usually two) of regimes, and the GARCH process switches between them according to a latent random variable. It was originally introduced as an ARCH model (see [16] and [39]) and is in that form defined as follows:

$$h_t = \sum_{i=1}^k \left(\alpha_0^{(i)} + \sum_{j=1}^q \alpha_j^{(i)} \varepsilon_{t-j}^2 \right) I(s_t = i) \quad (6)$$

where s_t is the discrete latent random variable obtaining values from the set $S = \{1, \dots, k\}$ of regime indicators. It follows a (usually first-order) homogeneous Markov chain:

$$\Pr\{s_t = j | s_t = i\} = p_{ij}, \quad i, j = 1, \dots, k \quad (7)$$

hence the name Markov-switching GARCH. The transition probabilities p_{ij} are parameters to be estimated. An argument put forward to motivate the use of these models is that they can be applied in situations where the conditional variance process contains breaks, that is, sudden changes in parameters. Augmenting the ARCH form (6) by lags of the conditional variance causes estimation problems; see [35, 42] and [37] for discussion and solutions.

Modeling the Conditional Standard Deviation. Models for the conditional variance constitute the most popular way of modeling conditional heteroskedasticity. Similar models, however, have also

been constructed for the conditional standard deviation. By replacing h_t in equation (4) by $h_t^{1/2}$, h_{t-j} by $h_{t-j}^{1/2}$, $j = 1, \dots, p$, and ε_{t-j}^2 by $|\varepsilon_{t-j}|$, $j = 1, \dots, q$, one obtains the absolute-valued GARCH model. Doing this in the GJR-GARCH model yields the threshold GARCH (TGARCH) model that [71] considered. The TGARCH(p, q) model is usually parameterized somewhat differently from the GJR-GARCH model:

$$h_t^{1/2} = \alpha_0 + \sum_{j=1}^q (\alpha_j^+ \varepsilon_{t-j}^+ - \alpha_j^- \varepsilon_{t-j}^-) + \sum_{j=1}^q \beta_j h_{t-j}^{1/2} \quad (8)$$

where $\varepsilon_{t-j}^+ = \max(\varepsilon_{t-j}, 0)$, $\varepsilon_{t-j}^- = \min(\varepsilon_{t-j}, 0)$, and $\alpha_j^+, \alpha_j^-, j = 1, \dots, q$, are the corresponding parameters. Like the GJR-GARCH model, even the TGARCH model is linear in parameters. It can be generalized by assuming the switch point to be an unknown parameter, which makes the model non-linear in parameters. This type of threshold ARCH model was introduced by Li and Li [46] who also considered a threshold autoregressive conditional mean, but it has not enjoyed the same popularity as the TGARCH or the GJR-GARCH model.

Integrated and Fractionally Integrated GARCH.

In applications it often occurs, if the time series is sufficiently long, that the estimated sum of the parameters α_1 and β_1 in the GARCH(1,1) model (equation (4) with $p = q = 1$) is close to unity. Engle and Bollerslev [24], who first paid attention to this phenomenon, suggested imposing the restriction $\alpha_1 + \beta_1 = 1$ and called the ensuing model an *integrated GARCH (IGARCH) model*. The IGARCH process is not weakly stationary as $E\varepsilon_t^2$ is not finite but it is, however, strongly stationary.

The empirical fact that the estimate of the sum $\alpha_1 + \beta_1$ often lies close to unity has prompted discussion of its causes. First Diebold [18] and later Lamoureux and Lastrapes [44] suggested that this often happens if the intercept of the GARCH process does not remain constant during the estimation period. This may not be surprising as it means that the underlying GARCH process is not stationary. This has aroused interest in testing parameter constancy and finding breaks in the GARCH process; see, for example [43, 52] and [8].

The integrated GARCH model has also served as a starting point for so-called fractionally integrated GARCH (FIGARCH) models. The GARCH(1,1) equation (4) can be written in the “ARMA(1,1) form” as follows:

$$\varepsilon_t^2 = \alpha_0 + (\alpha_1 + \beta_1)\varepsilon_{t-1}^2 + v_t - \beta_1 v_{t-1} \quad (9)$$

where $\{v_t\} = \{\varepsilon_t^2 - h_t\}$ is a martingale difference sequence with respect to h_t . For the IGARCH process, equation (9) has the “ARIMA(0,1,1) form”

$$(1 - L)\varepsilon_t^2 = \alpha_0 + v_t - \beta_1 v_{t-1} \quad (10)$$

where L is the lag operator, $Lx_t = x_{t-1}$. The FIGARCH(1, d ,0) model is obtained from equation (10) simply by assuming that the difference in equation (10) is *fractional* (see **Long Range Dependence**) instead of the conventional one:

$$(1 - L)^d \varepsilon_t^2 = \alpha_0 + v_t - \beta_1 v_{t-1} \quad (11)$$

where d is the fractional order of integration. The FIGARCH equation (11) can be written as an infinite-order ARCH model by applying the definition $v_t = \varepsilon_t^2 - h_t$ to it. This yields

$$h_t = \alpha_0(1 - \beta_1)^{-1} + \lambda(L)\varepsilon_t^2 \quad (12)$$

where $\lambda(L) = \{1 - (1 - L)^d(1 - \beta_1 L)^{-1}\}\varepsilon_t^2 = \sum_{j=1}^{\infty} \lambda_j L^j \varepsilon_t^2$, and $\lambda_j \geq 0$ for all j . Expanding the fractional difference operator into an infinite sum yields the result that for long lags j ,

$$\lambda_j = \{(1 - \beta_1)\Gamma(d)^{-1}\}j^{-(1-d)} = cj^{-(1-d)}, \quad c > 0 \quad (13)$$

where $d \in (0, 1)$ and $\Gamma(d)$ is the gamma function. When $d > 0.5$, the FIGARCH model is not weakly stationary. From equation (13) it is seen that the effect of the lag ε_{t-j}^2 on the conditional variance decays hyperbolically as a function of the lag length j . A very slow decay of the autocorrelation function of $\{\varepsilon_t^2\}$ is a frequently encountered empirical fact in return series, and this is the reason for developing the FIGARCH model (see [5]). But then, it has been argued (see for example [54]) that the observed slow decay is due to breaks in the variance of the return process. The FIGARCH model thus offers a competing explanation of this empirical phenomenon.

The GARCH-in-mean Model. Since volatility is viewed as a measure of risk, GARCH models are used for assessing the risk of a portfolio. From this it follows that the GARCH type conditional variance could be useful as a representation of the time-varying risk premium in explaining excess returns, that is, returns compared to the return of a riskless asset. This extends the considerations to involve the conditional mean of the return process as well. An excess return would be a combination of the unforecastable difference ε_t between the *ex ante* and *ex post* rates of return and a function of the conditional variance of the portfolio. Thus, if y_t is the excess return at time t ,

$$y_t = \beta + g(h_t) + \varepsilon_t \quad (14)$$

where h_t is defined as a GARCH process (4) and $g(h_t)$ is a well-defined function. Engle *et al.* [26] originally defined $g(h_t) = \delta h_t^{1/2}$, $\delta > 0$, which corresponds to the assumption that changes in the conditional variance appear less than proportionally in the mean. The alternative $g(h_t) = \delta h_t$ has also appeared in the literature. Equations (4) and (14) form the GARCH-in-mean or GARCH-M model. It has been quite frequently applied in the applied econometrics and finance literature. It may be mentioned that Glosten *et al.* [32] developed the GJR-GARCH model in the framework of the GARCH-M model.

Family of Exponential GARCH Models

The standard GARCH model has some disadvantages. Parameter restrictions are required to ensure positivity of the conditional variance at every point of time. Furthermore, the model does not allow an asymmetric response to shocks. This disadvantage was overcome by the appearance of asymmetric GARCH models but this was an important reason for the introduction of the exponential GARCH (EGARCH) model by Nelson [55]. A family of EGARCH(p, q) models may be defined as in equation (2) with

$$\ln h_t = \alpha_0 + \sum_{j=1}^q g_j(z_{t-j}) + \sum_{j=1}^p \beta_j \ln h_{t-j} \quad (15)$$

When $g_j(z_{t-j}) = \alpha_j z_{t-j} + \psi_j(|z_{t-j}| - \mathbb{E}|z_{t-j}|)$, $j = 1, \dots, q$, model (15) becomes the EGARCH model of [55]. It is seen from equation (15) that no parameter restrictions are necessary to ensure

positivity of h_t . Parameters α_j , $j = 1, \dots, q$, make an asymmetric response to shocks possible.

As in the standard GARCH case, the first-order model ($p = q = 1$) is the most popular EGARCH model in practice. Nelson [55] derived existence conditions for moments of the general infinite-order exponential ARCH model. For the EGARCH model with $g_j(z_{t-j})$ defined as above, these conditions imply that if the error process $\{z_t\}$ has all moments and $\sum_{j=1}^p \beta_j^2 < 1$ in equation (15), then all moments for the EGARCH process $\{\varepsilon_t\}$ exist. For example, if $\{z_t\}$ is a sequence of independent standard normal variables, then the restrictions on β_j , $j = 1, \dots, p$, are necessary and sufficient for the existence of all moments simultaneously. For GARCH models of the previous section, the moment conditions become more and more stringent for higher and higher even moments; see [40].

There is some similarity between the EGARCH model and the autoregressive *stochastic volatility* model (see **Stochastic Volatility Models**). Introducing a sequence $\{s_t\}$ of continuous unobservable independent random variables mutually independent of $\{z_t\}$ and substituting $g_j(s_{t-j})$ for $g_j(z_{t-j})$ in equation (15) leads to another class of models. Typically, in applications, $p = q = 1$ and $g_1(s_{t-1}) = \delta s_{t-1}$ where δ is a parameter to be estimated. This is called the *autoregressive stochastic volatility model*. Strictly speaking, it is not a model of conditional heteroskedasticity, because its information set at time t also contains $\{s_{t-j}, j \geq 1\}$.

Nonparametric and Semiparametric GARCH Models

Nonparametric GARCH models have been developed to remedy some of the disadvantages of the standard GARCH model mentioned earlier. A straightforward example is the model (2) in which $h_t = h(\varepsilon_{t-1}, \dots, \varepsilon_{t-q})$ is a smooth positive-valued function estimated using nonparametric techniques. When q is large, however, this model suffers from the curse of dimensionality. This may be partly remedied by an additive function, $h_t = \sum_{j=1}^q h_j(\varepsilon_{t-j})$, where $h_j(\varepsilon_{t-j})$, $j = 1, \dots, q$, are positive-valued functions that are estimated separately. Both of these models are ARCH models. Engle and Ng [27] introduced the following semiparametric GARCH model:

$$h_t = h(\varepsilon_{t-j}) + \beta h_{t-1} \quad (16)$$

where $h(\cdot)$ is a smooth unknown function to be estimated. With $j = 1$ this functional form is more flexible than the standard GARCH(1,1) model and allows modeling possible asymmetry in returns. For more discussion of nonparametric and semiparametric ARCH and GARCH models, see [50].

Estimation of GARCH Models and Properties of Estimators

The parameters of the GARCH(p, q) model (4) and its extensions may be estimated jointly by maximum likelihood. Asymptotic theory for the estimators has been developed over the years. Weiss [70] already proved consistency and asymptotic normality for maximum likelihood estimators in the ARCH(q) case under the assumption $E\varepsilon_t^4 < \infty$, which implies $Ez_t^4 < \infty$. General results are provided by Berkes *et al.* [9] who showed the asymptotic normality of the quasi-maximum likelihood (QML) estimators of the parameters of the GARCH(p, q) model where only the moment restriction $E|z_t^2|^{2+\delta} < \infty$, $\delta > 0$, is required. Francq and Zakoian [30] proved asymptotic normality of QML estimators for yet another set of conditions, and their results included the ARMA-GARCH family of models. Straumann and Mikosch [66] derived conditions for asymptotic normality of the QML estimators for a class of GARCH models that also includes some nonlinear GARCH models. In particular, they verified these conditions for the estimators of the AGARCH model. The corresponding conditions for the GARCH-in-mean model have not been verified, although consistency of the maximum likelihood estimators has been proven; see [53].

Rather often, the parameters of GARCH models are estimated assuming a t -distribution for the errors. This choice is based on empirical observation suggesting that the residuals of the estimated GARCH models assuming normality tend to be leptokurtic. If the t -distribution is applied, its degrees-of-freedom parameter or its inverse is generally estimated as well.

The parameters of GARCH models have to be estimated numerically because the conditional variances are not observed and have to be estimated separately for each iteration, that is, conditionally on parameter estimates obtained from the previous iteration. The estimated unconditional variance is used as the starting value for $h_0, h_{-1}, \dots, h_{-(p-1)}$, whereas $\hat{\varepsilon}_0^2 = \dots = \hat{\varepsilon}_{-q+1}^2 = 0$. In practice, the most popular algorithm for estimating GARCH models

by maximum likelihood has been the so-called Berndt–Hall–Hall–Hausman (BHHH) algorithm. Its popularity is partly based on the fact that it does not require the second derivatives of the log-likelihood function. Those may be considered difficult to compute given their recursive structure. Nevertheless, Fiorentini *et al.* [29] worked out the analytical second derivatives of the log-likelihood function for the GARCH model. This makes iterative estimation of parameters by the Newton–Raphson algorithm possible.

Several econometric software packages contain a module for estimating GARCH models. Brooks *et al.* [15] studied a number of them, focusing on numerical accuracy of the estimates. Their conclusion was that the algorithms based on analytical derivatives are preferable to the other alternatives. The best performing package used both analytical first and second derivatives. The authors reported substantial differences in numerical accuracy between the packages considered.

The parameters of the EGARCH model are also estimated by maximum likelihood. Nelson [55] gives the log-likelihood function based on the assumption that the errors follow a generalized error distribution (GED) with mean zero. The normal distribution is a special case of the GED. Nelson also discusses the role of the starting values in the estimation. In his simulations, the use of other starting values rather than the unconditional mean of $\ln h_t$ very rapidly leads to values of h_t obtained by starting from $E \ln h_t$. As in the case of GARCH models, the precision of the estimates may be expected to depend heavily on whether or not analytical derivatives of the log-likelihood are used in iterative estimation.

Properties of the QML estimators of the EGARCH(1,1) model have been considered by Straumann and Mikosch [66] who verified conditions for consistency of these estimators. The authors pointed out that verifying the corresponding conditions for asymptotic normality of the QML estimators is still an open question.

Multivariate Models of Conditional Heteroskedasticity

General Multivariate GARCH Model

The univariate GARCH model has been generalized to the vector case, which makes it possible to model

and forecast not only conditional variances of asset returns but covariances between returns as well. This has relevance in portfolio management. In the most general multivariate GARCH model, the so-called VEC-GARCH model of [14], every conditional variance and covariance is a linear function of all previous conditional variances and covariances. Assume a vector of returns $\mathbf{e}_t = (\varepsilon_{1t}, \dots, \varepsilon_{mt})'$ such that $\mathbf{E}\mathbf{e}_t = \mathbf{0}$ and $\mathbf{E}(\mathbf{e}_t \mathbf{e}_t' | \mathcal{F}_{t-1}) = \mathbf{H}_t = [h_{ijt}]$ where $\mathbf{H}_t$ is positive definite almost surely. It follows that

$$\mathbf{e}_t = \mathbf{H}_t^{1/2} \mathbf{z}_t \quad (17)$$

where $\mathbf{z}_t \sim \text{iid}(\mathbf{0}, \mathbf{I}_m)$. The VEC-GARCH(p, q) model is defined as follows:

$$\begin{aligned} \text{vech}(\mathbf{H}_t) &= \boldsymbol{\omega} + \sum_{k=1}^q \mathbf{A}_k \text{vech}(\mathbf{e}_{t-k} \mathbf{e}_{t-k}') \\ &+ \sum_{k=1}^p \mathbf{B}_k \text{vech}(\mathbf{H}_{t-k}) \end{aligned} \quad (18)$$

where $\mathbf{A}_k = [\alpha_{kij}]$ and $\mathbf{B}_k = [\beta_{kij}]$ are coefficient matrices with $m(m+1)/2$ rows and columns, and $\boldsymbol{\omega}$ is an $(m(m+1)/2) \times 1$ intercept vector with positive elements.

In equation (18), the number of parameters increases rapidly with the dimension of $\mathbf{e}_t$. Besides, general conditions for positive definiteness of $\mathbf{H}_t$ do not seem to exist. A simplified version in which the coefficient matrices $\mathbf{A}_k$ and $\mathbf{B}_k$ are diagonal has sometimes been applied. For this process, it is possible to work out the conditions for $\mathbf{H}_t$ to be positive definite, at least in the first-order case. In many applications, however, this model is already too restrictive, because it excludes any interaction between the conditional variances and covariances.

Bollerslev *et al.* [14] considered estimation of the general model and concluded that under sufficient regularity conditions the maximum likelihood estimator of the parameter vector is consistent and asymptotically normal. They used the BHHH optimization algorithm for obtaining the maximum likelihood estimates. Estimation of the VEC-GARCH model is, however, computationally demanding. The number of parameters can be large. Furthermore, the matrix $\mathbf{H}_t$ has to be inverted for each t in every iteration. It is likely that numerical considerations are at least as important here as they are in the univariate case. The

use of analytical first- and second-order derivatives would thus be preferable if they are available.

Constant Conditional Correlation Multivariate GARCH

There exist multivariate GARCH models with fewer parameters than the VEC-GARCH model. The constant conditional correlation GARCH (CCC-GARCH) model by Bollerslev [11] is built on the assumption that the conditional correlations between variables remain constant over time. This assumption is more parsimonious than the VEC-GARCH one, because the conditional covariances are determined through conditional variances and these correlations. Suppose again that $\mathbf{e}_t = (\varepsilon_{1t}, \dots, \varepsilon_{mt})'$ such that $\mathbf{E}\mathbf{e}_t = \mathbf{0}$ and $\mathbf{E}(\mathbf{e}_t \mathbf{e}_t' | \mathcal{F}_{t-1}) = \mathbf{H}_t = [h_{ijt}]$ where $\mathbf{H}_t$ is positive definite almost surely. The CCC-GARCH model has the following structure:

$$\mathbf{H}_t = \mathbf{D}_t \mathbf{P} \mathbf{D}_t \quad (19)$$

where $\mathbf{D}_t = \text{diag}(h_{11t}^{1/2}, \dots, h_{mmt}^{1/2})$ and $\mathbf{P} = [\rho_{ij}]$ is the positive definite correlation matrix, so that $h_{ijt} = \rho_{ij}(h_{iit}h_{jjt})^{1/2}$ for $i, j = 1, \dots, m$. Each conditional variance is assumed to follow a standard univariate GARCH model (4). The most common choice is the first-order GARCH model: $p = q = 1$. In this case, a bivariate first-order CCC-GARCH model has 7 parameters compared to 21 in the corresponding VEC-GARCH model.

One of the attractions of the CCC-GARCH model is that the parameter estimation is not complicated. The (conditional) quasi log-likelihood function of any GARCH model with $\mathbf{e}_t | \mathcal{F}_{t-1} \sim \text{iid } \mathcal{N}(\mathbf{0}, \mathbf{H}_t)$ has the form

$$L(\boldsymbol{\theta}) = c - (1/2) \sum_{t=1}^T (\ln |\mathbf{H}_t| - \mathbf{e}_t' \mathbf{H}_t^{-1} \mathbf{e}_t) \quad (20)$$

where $\boldsymbol{\theta}$ is the parameter vector. As already mentioned, maximizing equation (20) can be computationally quite costly when the model is the VEC-GARCH model, because an $m \times m$ matrix has to be inverted T times during a single iteration. The assumption of constant conditional correlations considerably simplifies things. Following Bollerslev [11], rewrite the conditional variances as $h_{jjt} = k_j \sigma_{jt}^2$, where $k_j > 0$, $j = 1, \dots, m$. Then one may partition the covariance matrix $\mathbf{H}_t$ as follows:

$$\mathbf{H}_t = \mathbf{D}_t^* \boldsymbol{\Gamma} \mathbf{D}_t^* \quad (21)$$

where $\mathbf{D}_t^* = \text{diag}(\sigma_{1t}, \dots, \sigma_{mt})$ and $\mathbf{\Gamma} = [\gamma_{ij}] = [\rho_{ij}(k_i k_j)^{1/2}]$. Applying equation (21) to equation (20) and keeping in mind that $\mathbf{D}_t^*$ is a diagonal matrix yields

$$L(\theta) = c - (T/2) \ln |\mathbf{\Gamma}| - \sum_{t=1}^T \sum_{j=1}^m \ln \sigma_{jt} - (1/2) \sum_{t=1}^T \bar{\mathbf{e}}_t' \mathbf{\Gamma}^{-1} \bar{\mathbf{e}}_t \quad (22)$$

where $\bar{\mathbf{e}}_t = \mathbf{D}_t^{*-1} \mathbf{e}_t$. In the maximization of equation (22), only a single matrix inversion is needed for each iteration. Even here, the BHHH optimization algorithm has been recommended for maximizing the log-likelihood. The calculations can be made even simpler by concentrating the log-likelihood, but the derivatives of the original log-likelihood (22) are required for obtaining an estimate of the asymptotic covariance matrix of the estimators of the parameters. The correlations are recovered from the relationship $\rho_{ij} = \gamma_{ij}(k_i k_j)^{-1/2}$.

Results on statistical properties of the QML estimators of a stationary CCC-GARCH model are available. Under appropriate regularity conditions, Jeantheau [41] proved consistency of these estimators and Ling and McAleer [49] verified conditions for their asymptotic normality.

The Dynamic Conditional Correlation GARCH Model

The assumption of constant correlations, however, does not always hold in practice. Tse and Tsui [68] and Engle [23] have extended the CCC-GARCH model by introducing time variation in the correlation matrix, and the two models are quite similar. The dynamic conditional correlation (DCC-) GARCH model of [23] has the following GARCH(1,1) form:

$$\begin{aligned} q_{ijt} &= \rho_{ij}(z) + \alpha^*(z_{i,t-1} z_{j,t-1} - \rho_{ij}(z)) \\ &\quad + \beta^*(q_{ij,t-1} - \rho_{ij}(z)) \\ &= (1 - \alpha^* - \beta^*) \rho_{ij}(z) + \alpha^* z_{i,t-1} z_{j,t-1} \\ &\quad + \beta^* q_{ij,t-1}, \quad i \neq j \end{aligned} \quad (23)$$

where $\rho_{ij}(z)$ is the unconditional correlation between z_{it} and z_{jt} , and $\alpha^* + \beta^* < 1$. If $\mathbb{E}q_{ijt} < \infty$, it

follows that $\mathbb{E}q_{ijt} = \rho_{ij}(z)$. The intercept is defined by variance targeting. The two ‘‘GARCH parameters’’ α^* and β^* are the same for all correlations. This makes it possible to model a large number of assets at the same time. The estimated conditional correlations are obtained *ex post* by making use of the relationship

$$\rho_{ijt} = \frac{q_{ijt}}{\sqrt{q_{iit} q_{jjt}}} \quad (24)$$

When $\alpha^* + \beta^* = 1$ in (23), ρ_{ijt} is determined by exponential smoothing.

Estimation of parameters of the DCC-GARCH model can be carried out as follows. The parameters of the individual GARCH equations contained in $\mathbf{D}_t$ (see equation (19)) can be estimated equation by equation assuming that $\mathbf{P} = \mathbf{I}_m$. Conditionally upon the results, one can then estimate the correlation parameters using the estimated standardized residuals z_{it} in equation (23). For example, assuming the parameters in $\mathbf{D}_t$ are known, the remaining parameters can be estimated by direct maximization of

$$\begin{aligned} L_T(\theta_{\text{corr}}) &= c - (1/2) \sum_{t=1}^T \ln |\mathbf{P}_t| \\ &\quad - (1/2) \sum_{t=1}^T \mathbf{z}_t' \mathbf{P}_t^{-1} \mathbf{z}_t \end{aligned} \quad (25)$$

where the elements of $\mathbf{P}_t$ are defined as in equation (23). Under regularity conditions, the estimators based on this two-step method are consistent but inefficient. Engle [23] also discusses simplified ways of obtaining estimates for the correlation parameters.

The CCC-GARCH model has other extensions that are not considered here. A few of them are based on the idea that the time-varying correlation matrix is a time-varying convex combination of two or more positive definite correlation matrices; see [7, 58] and [62, 63] for details.

The BEKK-GARCH Model

The BEKK-GARCH model (BEKK stands for Baba, Engle, Kraft and Kroner) is discussed in detail in [25]. It is another parsimonious GARCH model and has the property that the conditions of positive definiteness of the conditional covariance matrix are known. The

model is defined by completing equation (17) with

$$\begin{aligned} \mathbf{H}_t = & (\mathbf{C}_0^*)' \mathbf{C}_0^* + \sum_{k=1}^K \sum_{j=1}^q (\mathbf{A}_{jk}^*)' \boldsymbol{\varepsilon}_{t-j} \boldsymbol{\varepsilon}_{t-j}' \mathbf{A}_{jk}^* \\ & + \sum_{k=1}^K \sum_{j=1}^p (\mathbf{B}_{jk}^*)' \mathbf{H}_{t-j} \mathbf{B}_{jk}^* \end{aligned} \quad (26)$$

where $\mathbf{C}_0^* = (\mathbf{c}_{01}', \dots, \mathbf{c}_{0m}')'$ is an $m \times m$ triangular matrix, and $\mathbf{A}_{jk}^* = [\alpha_{jk,il}^*]$ and $\mathbf{B}_{jk}^* = [\beta_{jk,il}^*]$ are $m \times m$ coefficient matrices. When $\mathbf{C}_0^*$ is of full rank so that $(\mathbf{C}_0^*)' \mathbf{C}_0^*$ is positive definite, it follows that $\mathbf{H}_t$ is positive almost surely if the starting values $\mathbf{H}_0, \mathbf{H}_{-1}, \dots, \mathbf{H}_{-p+1}$ are positive semidefinite matrices, and $\boldsymbol{\varepsilon}_0, \boldsymbol{\varepsilon}_{-1}, \dots, \boldsymbol{\varepsilon}_{-q+1} = \mathbf{0}$. This is the main attraction of the BEKK-GARCH model. In applications, it is typically assumed that $K = p = q = 1$.

The log-likelihood function of the BEKK model is equation (20), and its partial derivatives are defined through equation (26). Engle and Kroner [25] remark that in the GARCH case (as opposed to ARCH) calculating analytical derivatives is complicated because it requires recursive computation of certain derivatives. They therefore recommend numerical derivatives and the BHHH algorithm. Nevertheless, Luchetti [51] has worked out the analytical score for the first-order BEKK-GARCH model when $K = 1$. Given the results reported in [15] in the univariate case, it appears that making use of the analytical score in the estimation of the parameters of this model would be very useful. Conditions for the asymptotic normality of the QML estimators of the parameters of the BEKK-GARCH model can be found in [17].

Factor GARCH Models

Observable Factor GARCH Models. The BEKK-GARCH model may be used as a starting point in a search for even more parsimonious models. Assume that, for each k , the matrices $\mathbf{A}_{jk}^*$ and $\mathbf{B}_{jk}^*$ in equation (26) have rank one and the same $m \times 1$ eigenvectors, $\mathbf{g}_k$ for $\mathbf{A}_{jk}^*$ and $\mathbf{f}_k$ for $\mathbf{B}_{jk}^*$, respectively, with $\mathbf{g}_k' \mathbf{f}_i = 0$, $i \neq k$, and $\mathbf{g}_k' \mathbf{f}_k = 1$. Then equation (26) collapses into

$$\mathbf{H}_t = \boldsymbol{\Omega} + \sum_{k=1}^K \mathbf{g}_k \mathbf{g}_k' \left\{ \sum_{j=1}^q \alpha_{kj}^2 \gamma_{k,t-j}^2 + \sum_{j=1}^p \beta_{kj}^2 h_{k,t-j}^{(f)} \right\} \quad (27)$$

where $\boldsymbol{\Omega} = (\mathbf{C}_0^*)' \mathbf{C}_0^*$, $\gamma_{kt} = \mathbf{f}_k' \boldsymbol{\varepsilon}_t$ and $h_{kt}^{(f)} = \mathbf{f}_k' \mathbf{H}_t \mathbf{f}_k$. This is the K -factor GARCH model of [28]. Each factor γ_{kt} has a univariate GARCH(p, q) structure: $\gamma_{kt} | \mathcal{F}_{t-1} \sim \mathcal{D}(0, h_{kt}^{(f)})$ where $\mathcal{D}$ denotes a distribution, such that the conditional variance

$$h_{kt}^{(f)} = c_k + \sum_{j=1}^q \alpha_{kj}^2 \gamma_{k,t-j}^2 + \sum_{j=1}^p \beta_{kj}^2 h_{k,t-j}^{(f)} \quad (28)$$

where $c_k = \mathbf{f}_k' \boldsymbol{\Omega} \mathbf{f}_k$. Writing $\gamma_{kt} = z_{kt} (\mathbf{f}_k' \mathbf{H}_t \mathbf{f}_k)^{1/2}$, $k = 1, \dots, K$, where $\{z_{kt}\} \sim \text{iid}(0, 1)$ such that $\mathbf{E} z_{kt} z_{jt} = \omega_{kj}$, it follows that γ_{kt} and γ_{jt} have a constant conditional correlation ω_{kj} . The model consisting of equations (17) and (27) and characterizing $\boldsymbol{\varepsilon}_t$ is just a linear combination of the K conditional variances of the univariate common factors γ_{kt} . It is important to note that the factors γ_{kt} , $k = 1, \dots, K$, are observable and not latent variables because $\mathbf{f}_k$, $k = 1, \dots, K$, are assumed known.

This model has an economic interpretation. Theories of asset pricing suggest that the risk premium of an asset depends on its covariance with other assets as well as its own variance. The factor GARCH model represents those covariances in a parsimonious way and makes it possible to simultaneously consider a larger number of asset returns than many other vector GARCH models would allow in practice. The K factors define as many portfolios of these assets. It is then reasonable to assume that the elements of $\mathbf{f}_k$ are nonnegative and sum up to one.

The predetermined factors in this model can be correlated. There are also factor models with uncorrelated factors. The factors are obtained through a one-to-one transformation of $\boldsymbol{\varepsilon}_t$:

$$\boldsymbol{\varepsilon}_t = \mathbf{W} \mathbf{f}_t \quad (29)$$

where $\mathbf{W}$ is a nonsingular $m \times m$ matrix. Each f_{jt} in $\mathbf{f}_t = (f_{1t}, \dots, f_{mt})'$ is thus a function of the elements of $\boldsymbol{\varepsilon}_t$ and is typically assumed to follow a GARCH process. The specification of $\mathbf{W}$ varies from one model to the next. Examples include [1, 69] and [45].

The parameters of the factor GARCH model (17) and (27) can be estimated by QML with restrictions $\mathbf{g}_k' \mathbf{f}_i = 0$, $i \neq k$, and $\mathbf{g}_k' \mathbf{f}_k = 1$, but this can be numerically difficult. The fact that the common conditional variances can be viewed as generated by univariate GARCH processes suggests two-step estimation of parameters. First, c_k , α_{kj}^2 and β_{kj}^2 and the common conditional variances $h_{kt}^{(f)}$ are estimated from the K

univariate GARCH equations (28). Equation (27) is then rewritten as

$$\mathbf{H}_t = [h_{ijt}] = \left\{ \boldsymbol{\Omega} - \sum_{k=1}^K c_k \mathbf{g}_k \mathbf{g}_k' \right\} + \sum_{k=1}^K \mathbf{g}_k \mathbf{g}_k' h_{kt}^{(f)} \quad (30)$$

At the second stage, the remaining parameters are estimated making use of equation (30) and the estimates of c_k and $h_{kt}^{(f)}$. Lin [48] considers the estimation of factor GARCH models, including these two methods, the QML and the two-step one. More details and references can be found in that paper.

Unobserved Factor Models. In the factor model (27), the factors at time t , as can be seen from equation (30), are functions of observed variables $\boldsymbol{\varepsilon}_{t-j}$, $j \geq 1$. Another strand of factor GARCH models has drawn inspiration from the factor analysis and builds on genuinely unobserved factors. In a general form, one may write

$$\boldsymbol{\varepsilon}_t = \mathbf{C}\mathbf{f}_t + \boldsymbol{\eta}_t \quad (31)$$

where $\mathbf{f}_t$ is now a $k \times 1$ vector of unknown factors such that $\mathbf{E}\mathbf{f}_t = 0$ and $\text{cov}(\mathbf{f}_t | \mathcal{F}_{t-1}) = \boldsymbol{\Lambda}_t$ where $\boldsymbol{\Lambda}_t$ is a time-varying covariance matrix, often assumed diagonal, and $\mathbf{C}$ is an $m \times k$ matrix of factor loadings with rank $k \leq m$. Furthermore, for the disturbance vector, $\mathbf{E}\boldsymbol{\eta}_t = 0$ and $\text{cov}(\boldsymbol{\eta}_t | \mathcal{F}_{t-1}) = \boldsymbol{\Sigma}_\eta$, and $\mathbf{E}\{\mathbf{f}_t \boldsymbol{\eta}_t' | \mathcal{F}_{t-1}\} = \mathbf{0}$. The information set $\mathcal{F}_{t-1}$ contains the unobservable variables $\mathbf{f}_{t-j}$, $j \geq 1$, unlike the econometrician's information set $\mathcal{F}_{t-1}$ that only contains the observable variables $\boldsymbol{\varepsilon}_{t-j}$, $j \geq 1$. In [60], it is shown that there always exists a model of type (31) such that the observable factor model (27) is observationally equivalent to it.

Diebold and Nerlove [20] introduced a model of form (31), in which $k = 1$ and $\mathbf{f}_t$ has an ARCH structure. They use it to model the volatility of a set of weekly spot exchange rates. In their model,

$$\boldsymbol{\varepsilon}_t = \boldsymbol{\gamma} f_t + \boldsymbol{\eta}_t$$

$$\mathbf{E}(f_t | \mathcal{F}_{t-1}) = h_t^{(f)} = \alpha_0 + \sum_{j=1}^q \alpha_j f_{t-j}^2 \quad (32)$$

From equation (32) it follows that

$$\mathbf{H}_t = \mathbf{E}(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t' | \mathcal{F}_{t-1}) = h_t^{(f)} \boldsymbol{\gamma} \boldsymbol{\gamma}' + \boldsymbol{\Sigma}_\eta \quad (33)$$

The conditional variances and covariances are thus determined by a single factor. Diebold and Nerlove [20] estimated the parameters of (32) using the Kalman filter.

Final Remarks

The literature on GARCH models is quite voluminous. Modern econometrics texts contain accounts of conditional heteroskedasticity. A number of surveys of GARCH models exist as well. Bollerslev *et al.* [13], Diebold and Lopez [19], Palm [57], and Guégan [36] survey developments till the early 1990s; see [31] for a very recent survey. Shephard [61] considers both univariate GARCH and stochastic volatility models. Gouriéroux [34] concentrates on ARCH models, both univariate and multivariate. The survey by Bollerslev *et al.* [12] also reviews applications to financial series. The focus in [65] is on estimation in models of conditional heteroskedasticity. Theoretical results on time series models with conditional heteroskedasticity are also reviewed in [47]. Engle [22] provides a selection of the most important articles on ARCH and GARCH models up until 1993. Recent surveys of multivariate GARCH models include [6] and [64]. Several articles in [4] provide detailed descriptions of various aspects of GARCH models and their use in applications.

Acknowledgments

This work has been supported by the Danish National Research Foundation.

References

- [1] Alexander, C. & Chibumba, A. (1996). Multivariate orthogonal factor GARCH, *Discussion Papers in Mathematics*, University of Sussex.
- [2] Andersen, T.G. (2009). Realized volatility, in *Handbook of Financial Time Series*, T.G. Andersen, R.A. Davis, J.-P. Kreiss & T. Mikosch, eds, Springer, New York, pp. 555–575.
- [3] Andersen, T.G. & Bollerslev, T. (1998). Answering the skeptics: yes, standard volatility models provide accurate forecasts, *International Economic Review* **39**, 885–905.
- [4] Andersen, T.G., Davis, R.A., Kreiss, J.-P. & Mikosch, T. (eds) (2009). *Handbook of Financial Time Series*, Springer, New York.

- [5] Baillie, R.T., Bollerslev, T. & Mikkelsen, H.O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **74**, 3–30.
- [6] Bauwens, L., Laurent, S. & Rombouts, J.V.K. (2006). Multivariate GARCH models: a survey, *Journal of Applied Econometrics* **21**, 79–109.
- [7] Berben, R.-P. & Jansen, W.J. (2005). Comovement in international equity markets: a sectoral view, *Journal of International Money and Finance* **24**, 832–857.
- [8] Berkes, I., Gombay, E., Horváth, L. & Kokoszka, P. (2004). Sequential change-point detection in GARCH (p,q) models, *Econometric Theory* **20**, 1140–1167.
- [9] Berkes, I., Horváth, L. & Kokoszka, P. (2003). GARCH processes: structure and estimation, *Bernoulli* **9**, 201–227.
- [10] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**, 307–327.
- [11] Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model, *Review of Economics and Statistics* **72**, 498–505.
- [12] Bollerslev, T., Chou, R.Y. & Kroner, K.F. (1992). ARCH modeling in finance. A review of the theory and empirical evidence, *Journal of Econometrics* **52**, 5–59.
- [13] Bollerslev, T., Engle, R.F. & Nelson, D.B. (1994). ARCH models, in *Handbook of Econometrics*, R.F. Engle & D.L. McFadden, eds, North-Holland, Amsterdam, Vol. 4, pp. 2959–3038.
- [14] Bollerslev, T., Engle, R.F. & Wooldridge, J. (1988). A capital asset-pricing model with time-varying covariances, *Journal of Political Economy* **96**, 116–131.
- [15] Brooks, C., Burke, S.P. & Persaud, G. (2001). Benchmarks and the accuracy of GARCH model estimation, *International Journal of Forecasting* **17**, 45–56.
- [16] Cai, J. (1994). A Markov model of switching-regime ARCH, *Journal of Business and Economic Statistics* **12**, 309–316.
- [17] Comte, F. & Lieberman, O. (2003). Asymptotic theory for multivariate GARCH processes, *Journal of Multivariate Analysis* **84**, 61–84.
- [18] Diebold, F.X. (1986). Modeling the persistence of conditional variances: a comment, *Econometric Reviews* **5**, 51–56.
- [19] Diebold, F.X. & Lopez, J.A. (1995). Modeling volatility dynamics, in *Macroeconometrics: Developments, Tensions, and Prospects*, K.D. Hoover, ed, Kluwer Academic Publishers, Boston, pp. 427–472.
- [20] Diebold, F.X. & Nerlove, M. (1989). The dynamics of exchange rate volatility: a multivariate latent factor ARCH model, *Journal of Applied Econometrics* **4**, 1–21.
- [21] Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica* **50**, 987–1007.
- [22] Engle, R.F. (ed) (1995). *ARCH Selected Readings*, Oxford University Press, Oxford.
- [23] Engle, R.F. (2002). Dynamic conditional correlation: a simple class of multivariate generalized autoregressive conditional heteroscedasticity models, *Journal of Business and Economic Statistics* **20**, 339–350.
- [24] Engle, R.F. & Bollerslev, T. (1986). Modeling the persistence of conditional variances, *Econometric Reviews* **5**, 1–50.
- [25] Engle, R.F. & Kroner, K.F. (1995). Multivariate simultaneous generalized ARCH, *Econometric Theory* **11**, 122–150.
- [26] Engle, R.F., Lilien, D.M. & Robins, R.P. (1987). Estimating time-varying risk premia in the term structure: the ARCH-M model, *Econometrica* **55**, 391–407.
- [27] Engle, R.F. & Ng, V.K. (1993). Measuring and testing the impact of news on volatility, *Journal of Finance* **48**, 1749–1777.
- [28] Engle, R.F., Ng, V.K. & Rotschild, M. (1990). FACTOR-ARCH covariance structure: empirical estimates for treasury bills, *Journal of Econometrics* **45**, 213–237.
- [29] Fiorentini, G., Calzolari, G. & Panattoni, L. (1996). Analytic derivatives and the computation of GARCH estimates, *Journal of Applied Econometrics* **11**, 399–417.
- [30] Francq, C. & Zakoian, J.-M. (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes, *Bernoulli* **10**, 605–637.
- [31] Giraitis, L., Leipus, R. & Surgailis, D. (2006). Recent advances in ARCH modelling, in *Long Memory in Economics*, A. Kirman & G. Teyssiere, eds, Springer, Berlin, pp. 3–38.
- [32] Glosten, L., Jagannathan, R. & Runkle, D. (1993). On the relation between expected value and the volatility of the nominal excess return on stocks, *Journal of Finance* **48**, 1779–1801.
- [33] Gonzalez-Rivera, G. (1998). Smooth transition GARCH models, *Studies in Nonlinear Dynamics and Econometrics* **3**, 161–178.
- [34] Gouriéroux, C. (1997). *ARCH Models and Financial Applications*, Springer, Berlin.
- [35] Gray, S.F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process, *Journal of Financial Economics* **42**, 27–62.
- [36] Guégan, D. (1994). *Séries chronologiques non-linéaires à temps discret*, Economica, Paris.
- [37] Haas, M., Mittnik, S. & Paoletta, M.S. (2004). A new approach to Markov-switching GARCH models, *Journal of Financial Econometrics* **4**, 493–530.
- [38] Hagerud, G. (1997). *A New Non-Linear GARCH Model*, EFI Economic Research Institute, Stockholm.
- [39] Hamilton, J.D. & Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime, *Journal of Econometrics* **64**, 307–333.
- [40] He, C. & Teräsvirta, T. (1999). Properties of moments of a family of GARCH processes, *Journal of Econometrics* **92**, 173–192.
- [41] Jeantheau, T. (1998). Strong consistency of estimators for multivariate ARCH models, *Econometric Theory* **14**, 70–86.

- [42] Klaassen, F. (2002). Improving GARCH volatility forecasts with regime-switching GARCH model, *Empirical Economics* **27**, 363–394.
- [43] Kokoszka, P. & Leipus, R. (1999). Testing for parameter changes in ARCH models, *Lithuanian Mathematical Journal* **39**, 182–195.
- [44] Lamoureux, C.G. & Lastrapes, W.G. (1990). Persistence in variance, structural change and the GARCH model, *Journal of Business and Economic Statistics* **8**, 225–234.
- [45] Lanne, M. & Saikkonen, P. (2007). A multivariate generalized orthogonal factor GARCH model, *Journal of Business and Economic Statistics* **25**, 61–75.
- [46] Li, C.W. & Li, W.K. (1996). On a double threshold autoregressive heteroskedasticity time series model, *Journal of Applied Econometrics* **11**, 253–274.
- [47] Li, W.K., Ling, S. & McAleer, M. (2002). Recent theoretical results for time series models with GARCH errors, *Journal of Economic Surveys* **16**, 245–269.
- [48] Lin, W.-L. (1992). Alternative estimators for factor GARCH models—a Monte Carlo comparison, *Journal of Applied Econometrics* **7**, 259–279.
- [49] Ling, S. & McAleer, M. (2003). Asymptotic theory for a vector ARMA-GARCH model, *Econometric Theory* **19**, 280–310.
- [50] Linton, O.B. (2009). Semiparametric and nonparametric ARCH modelling, in *Handbook of Financial Econometrics*, T.G. Andersen, R.A. Davis, J.-P. Kreiss & T. Mikosch, eds, Springer, New York, pp. 157–167.
- [51] Lucchetti, R. (2002). Analytical scores for multivariate GARCH models, *Computational Economics* **19**, 133–143.
- [52] Lundbergh, S. & Teräsvirta, T. (2002). Evaluating GARCH models, *Journal of Econometrics* **110**, 417–435.
- [53] Meitz, M. & Saikkonen, P. (2008). Ergodicity, mixing, and existence of moments of a class of Markov models with applications to GARCH and ACD models, *Econometric Theory* **24**, 1291–1320.
- [54] Mikosch, T. & Starica, C. (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects, *Review of Economics and Statistics* **86**, 378–390.
- [55] Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: a new approach, *Econometrica* **59**, 347–370.
- [56] Nelson, D.B. & Cao, C.Q. (1992). Inequality constraints in the univariate GARCH model, *Journal of Business and Economic Statistics* **10**, 229–235.
- [57] Palm, F.C. (1996). GARCH models of volatility, in *Handbook of Statistics 14: Statistical Methods in Finance*, G. Maddala & C. Rao, eds, Elsevier, Amsterdam, pp. 209–240.
- [58] Pelletier, D. (2006). Regime switching for dynamic correlations, *Journal of Econometrics* **131**, 445–473.
- [59] Sentana, E. (1995). Quadratic ARCH models, *Review of Economic Studies* **62**, 639–661.
- [60] Sentana, E. (1998). The relation between conditionally heteroskedastic factor models and factor GARCH models, *Econometrics Journal* **1**, 1–9.
- [61] Shephard, N.G. (1996). Statistical aspects of ARCH and stochastic volatility, in *Time Series Models. In Econometrics, Finance and Other Fields*, D.R. Cox, D.V. Hinkley & O.E. Barndorff-Nielsen, eds, Chapman & Hall, London, pp. 1–67.
- [62] Silvennoinen, A. & Teräsvirta, T. (2005). Multivariate autoregressive conditional heteroskedasticity with smooth transitions in conditional correlations, *SSE/EFI Working Papers in Economics and Finance* 577, Stockholm School of Economics.
- [63] Silvennoinen, A. & Teräsvirta, T. (2007). Modelling multivariate autoregressive conditional heteroskedasticity with the double smooth transition conditional correlation GARCH model, *SSE/EFI Working Paper Series in Economics and Finance* 652, Stockholm School of Economics.
- [64] Silvennoinen, A. & Teräsvirta, T. (2009). Multivariate GARCH models, in *Handbook of Financial Time Series*, T.G. Andersen, R.A. Davis, J.-P. Kreiss & T. Mikosch, eds, Springer, New York, pp. 201–229.
- [65] Straumann, D. (2004). *Estimation in Conditionally Heteroscedastic Time Series Models*, Springer, New York.
- [66] Straumann, D. & Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: a stochastic recurrence equations approach, *Annals of Statistics* **34**, 2449–2495.
- [67] Taylor, S.J. (1986). *Modelling Financial Time Series*, Wiley, Chichester.
- [68] Tse, Y.K. & Tsui, K.C. (2002). A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations, *Journal of Business and Economic Statistics* **20**, 351–362.
- [69] van der Weide, R. (2002). GO-GARCH: a multivariate generalized orthogonal GARCH model, *Journal of Applied Econometrics* **17**, 549–564.
- [70] Weiss, A.A. (1986). Asymptotic theory for ARCH models: estimation and testing, *Econometric Theory* **2**, 107–131.
- [71] Zakoian, J.-M. (1994). Threshold heteroskedastic models, *Journal of Economic Dynamics and Control* **18**, 931–955.

Related Articles

Long Range Dependence; Realized Volatility and Multipower Variation; Volatility.

TIMO TERÄSVIRTA

Volatility

Financial volatility is, loosely speaking, a measure of the variability of asset returns (*see Stochastic Volatility Models*). It has been known for a long time that volatility changes through time. Both *ex post* and *ex ante* volatility measures are in common use. The most well-known *ex post* measure is realized volatility (Rvol), while *ex ante* measures include those generated by the ARCH-type models and option-based numbers such as implied volatility (*see Implied Volatility Surface*) and the VIX. Reviews of this literature include, among others, [4, 12, 23, 36].

We focus here on high frequency *ex post* measures of volatility and models of volatility driven by Lévy processes. Throughout our discussion, we refer to a vector of price processes

$$Y_t = Y_0 + \int_0^t \mu_u du + \int_0^t \sigma_u dW_u, \quad t \geq 0 \quad (1)$$

where W_t is Brownian motion and μ_t, σ_t are spot drift and volatility. We write $\mathcal{F}_t$ as the continuous time record of the past Y_t . For simplicity of exposition, we ignore jumps in the price process.

Ex post Measures

Itô's formula (*see Itô's Formula*) implies

$$\begin{aligned} \text{Var} \left\{ \left(Y_t - Y_0 - \int_0^t \mu_u du \right) | \mathcal{F}_0 \right\} \\ = E \left\{ \int_0^t \sigma_u \sigma_u' du | \mathcal{F}_0 \right\} \\ = E \{ [Y, Y]_t | \mathcal{F}_0 \} \end{aligned} \quad (2)$$

linking together *ex ante* variance forecasts with *ex ante* forecasts of the quadratic variation $[Y, Y]$ of Y . However, we can use high-frequency data to estimate the $[Y, Y]_t$ process, which allows us to project this quantity into the future without an explicit model for the spot volatility matrix σ_t . This method of constructing forecasts of volatility through QV was developed by Andersen *et al.* [2] (*see Realized Volatility and Multipower Variation*).

There is active literature on estimating the $[Y, Y]_t$ process. Most of it deals with estimating discrete

increments of the process over specified time periods, such as a day, $IV_i = [Y, Y]_i - [Y, Y]_{i-1}$. The most familiar estimator is the realized variance (RV), for the i th day, which is

$$RV_i = \sum_{i-1 < t_j \leq i} (Y_{t_j} - Y_{t_{j-1}})(Y_{t_j} - Y_{t_{j-1}})' \quad (3)$$

where t_j are the times at which data is available—which could represent the times of trades and quotes (or a subset of such trades and quotes). In the univariate case, the square root of RV is the realized volatility. Many option contracts are directly written on Rvol and RV, but usually this is where volatility is measured over the month and daily returns are used in equation (3).

The econometric theory associated with RV uses an in-fill asymptotic analysis, imagining $t_j - t_{j-1} \downarrow 0$. This is somewhat problematic as equation (1) ignores market microstructure effects (*see Market Microstructure Effects*) which bite in the limit. However, for the moment we ignore this. Then, $RV_i \xrightarrow{P} IV_i$ by the definition of QV. The central limit theory extension of this, in the univariate case, is

$$\widehat{V}_i^{-1/2} \sqrt{n} (RV_i - IV_i) \xrightarrow{d} N(0, 1) \quad (4)$$

and was developed by Barndorff-Nielsen and Shephard [11] (who suggested a variety of $\widehat{V}_i$ quantities). See also [28, 31, 32]. Figure 1 illustrates this (together with results from another estimator discussed below), showing 95% confidence intervals using 20-min returns for General Electric, based on trades made on the stock exchange in November 2004.

Zhou [41] was the first to formally study the properties of RV-type statistics in the presence of market microstructure noise—which causes returns to be autocorrelated in time. Here, we think of noise as $U = X - Y$, where X is the observed trade or quote process and Y continues to be the underlying efficient price. Modern research has developed three methods for being robust to noise: preaveraging, [29] multiscale [40] and [39], and realized kernels [6, 7].

We focus on realized kernels, which generalize RV to

$$K(X) = \sum_{h=-H}^H k\left(\frac{h}{H+1}\right) \gamma_h, \quad \gamma_h = \sum_{j=|h|+1}^n r_j r_{j-|h|} \quad (5)$$

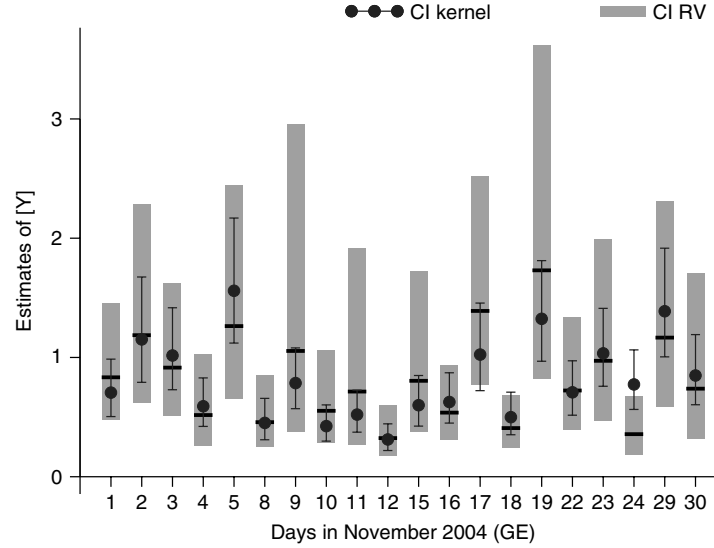


Figure 1 Daily estimates of *ex post* variability: RV and realized kernel, together with 95% confidence intervals

where $\{r_j\}$ are high-frequency data and k is a Parzen weight function:

$$k(x) = \begin{cases} 1 - 6x^2 + 6x^3, & 0 \leq x \leq 1/2 \\ 2(1 - x)^3, & 1/2 \leq x \leq 1 \\ 0, & x > 1 \end{cases} \quad (6)$$

This is like an HAC estimator, which is popular in econometrics, but there is no scaling by sample size. It is robust to (i) irregularly spaced data, (ii) endogenous noise, and (iii) autocorrelation in the noise. A key feature of realized kernels is the choice of the bandwidth H . The value that approximately minimizes the asymptotic MSE of the estimator is $H = 0.97\xi^{4/5}n^{3/5}$, $\xi^2 = \text{Var}(X - Y)/IV$. Figure 1 illustrates the use of this method, which shows the improvement from using the high-frequency data. More details on how to implement this method are given in [8]. The multivariate implementation is discussed in [7].

These high-frequency-data-based volatility measures can be combined to make volatility forecasts in a variety of ways. Andersen *et al.* [1–3] have pioneered the use of simple linear time-series methods. An alternative is to use lagged RV measures as an explanatory variable in the volatility dynamics of an ARCH model [24].

A wide variety of diffusion-based models of volatility have been proposed (*see Stochastic Volatility Models; Local Volatility Model; Heston Model*). The leading examples of these are the models proposed by Heston [25, 27, 35], while an interesting analysis of such models is found in [22, 30].

Pure Jump-based Volatility Models

There is now quite a lot of evidence that there are jumps in the volatility process [38] and that jumps play a key role in generating statistical leverage effects [9]. An article that is a distinct break from the traditional diffusion-based stochastic volatility models associated with [27] is that of Barndorff-Nielsen and Shephard [10]. In their simplest model, these authors wrote the univariate spot variance as an OU-type process (*see Ornstein–Uhlenbeck Processes*)

$$d\sigma_t^2 = -\lambda\sigma_t^2 dt + dZ_{\lambda t}, \quad \lambda > 0 \quad (7)$$

where Z is a subordinator. It may be recalled that a subordinator is a Lévy process with nonnegative increments, which means σ_t^2 has no Brownian component at all (*see Barndorff-Nielsen and Shephard (BNS) Models*). They have the feature that if $\text{Var}(Z_1) < \infty$, then $\text{Corr}(\sigma_t^2, \sigma_{t+s}^2) = \exp(-\lambda s)$.

A main advantage of this OU process is that

$$\sigma_t^2 = e^{-\lambda t} \sigma_0^2 + \int_0^t e^{-\lambda(t-s)} dZ_{\lambda s} \quad (8)$$

and

$$\begin{aligned} \int_0^t \sigma_u^2 du &= \lambda^{-1} (1 - e^{-\lambda t}) \sigma_0^2 \\ &+ \lambda^{-1} \int_0^t \{1 - e^{-\lambda(t-s)}\} dZ_{\lambda s} \end{aligned} \quad (9)$$

Hence they are easy to analyze and simulate from.

More sophisticated dynamics can be achieved by adding together independent OU processes, each with a different decay parameter λ . Such superposition processes can be extended to the case of an infinite number of components, as analyzed by Barndorff-Nielsen [5].

A different route is to generate ARMA-type dynamics. Such processes have been advocated by Brockwell [19, 20, 37].

The OU process has been extended to the multivariate case by Barndorff-Nielsen and Stelzer [13] and [34]. They set up an OU process for the positive semidefinite covariance matrix $\Sigma_t = \sigma_t \sigma_t'$ and analyzed the properties of the resulting implied multivariate return process.

These BNS models have been used extensively in applications. Derivative pricing on the basis of this type of model was first discussed by Nicolato and Venardos [33]. Subsequent research includes [14]. Benth *et al.* [15] studied the pricing of the variance and volatility swaps based on the BNS model. Minimal martingale measures for BNS models are analyzed by Benth and Meyer-Brandis [18, 26]. Benth *et al.* [16] studied the calculation of Greeks for BNS-type models.

Benth *et al.* and Delong and Klüppelberg [17, 21] have used BNS models to provide analytic solutions to the portfolio allocation problem.

These models and variants based on them have been used as the models of prices traded in energy markets, where extreme movements are quite common (see **Electricity Markets**).

References

- [1] Andersen, T.G., Bollerslev, T., Diebold, F.X. & Ebens, H. (2001). The distribution of realized stock return volatility, *Journal of Financial Economics* **61**, 43–76.
- [2] Andersen, T.G., Bollerslev, T., Diebold, F.X. & Labys, P. (2001). The distribution of exchange rate volatility, *Journal of the American Statistical Association* **96**, 42–55. Correction published in 2003, **98**, 501.
- [3] Andersen, T.G., Bollerslev, T., Diebold, F.X. & Labys, P. (2003). Modeling and forecasting realized volatility, *Econometrica* **71**, 579–625.
- [4] Andersen, T.G., Bollerslev, T. & Diebold, F.X. (2009). Parametric and nonparametric measurement of volatility, in *Handbook of Financial Econometrics*, Y. Ait-Sahalia & L.P. Hansen, eds, North Holland, Amsterdam, forthcoming.
- [5] Barndorff-Nielsen, O.E. (2001). Superposition of Ornstein-Uhlenbeck type processes, *Theory of Probability and its Applications* **46**, 175–194.
- [6] Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. & Shephard, N. (2008). Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise, *Econometrica* **76**, 1481–1536.
- [7] Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. & Shephard, N. (2008). *Multivariate Realised Kernels: Consistent Positive Semi-definite Estimators of the Covariation of Equity Prices with Noise and Non-synchronous Trading*, Oxford-Man Institute, University of Oxford, forthcoming.
- [8] Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. & Shephard, N. (2009). Realised kernels in practice: trades and quotes, *Econometrics Journal* forthcoming.
- [9] Barndorff-Nielsen, O.E., Kinnebrock, S. & Shephard, N. (2009). Measuring downside risk: realised semivariance, in *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, T. Bollerslev, J. Russell & M. Watson, eds, Oxford University Press, pp. 117–136.
- [10] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics (with discussion), *Journal of the Royal Statistical Society, Series B* **63**, 167–241.
- [11] Barndorff-Nielsen, O.E. & Shephard, N. (2002). Econometric analysis of realised volatility and its use in estimating stochastic volatility models, *Journal of the Royal Statistical Society, Series B* **64**, 253–280.
- [12] Barndorff-Nielsen, O.E. & Shephard, N. (2007). Variation, jumps and high frequency data in financial econometrics, in *Advances in Economics and Econometrics. Theory and Applications, Ninth World Congress*, R. Blundell, T. Persson & W.K. Newey, eds, Econometric Society Monographs, Cambridge University Press, pp. 328–372.
- [13] Barndorff-Nielsen, O.E. & Stelzer, R. (2007). Positive-definite matrix processes, *Probability and Mathematical Statistics* **27**, 3–43.
- [14] Benth, F.E. (2004). *Option Theory with Stochastic Analysis*, Springer.

- [15] Benth, F.E., Groth, M. & Kufakunesu, R. (2007). Valuing volatility and variance swaps for a non-Gaussian Ornstein-Uhlenbeck stochastic volatility model, *Applied Mathematical Finance* **14**, 347–363.
- [16] Benth, F.E., Groth, M. & Wallin, O. (2007). *Derivative-free Greeks for the Barndorff-Nielsen and Shephard Stochastic Volatility Model*, University of Oslo, E-print, No. 3 March 2007.
- [17] Benth, F.E., Karlsen, K.H. & Reikvam, K. (2003). Merton's portfolio optimization problem in a Black and Scholes market with non-Gaussian stochastic volatility of Ornstein-Uhlenbeck type, *Mathematical Finance* **13**, 215–244.
- [18] Benth, F.E. & Meyer-Brandis, T. (2006). The density process of the minimal entropy martingale measure in a stochastic volatility model with jumps, *Finance and Stochastics* **9**, 563–575.
- [19] Brockwell, P.J. (2001). Lévy driven CARMA processes, *Annals of the Institute of Statistical Mathematics* **52**, 113–124.
- [20] Brockwell, P.J. & Marquardt, T. (2005). Lévy-driven and fractionally integrated ARMA processes with continuous time parameter, *Statistica Sinica* **15**, 477–494.
- [21] Delong, L. & Klüppelberg, C. (2007). Optimal Investment and Consumption in a Black-Scholes Market with Stochastic Coefficients Levy driven, *Annals of Applied Probability*, **18**, 879–908.
- [22] Drost, F.C. & Werker, B. (1996). Closing the GARCH gap: continuous time GARCH modelling, *Journal of Econometrics* **74**, 31–57.
- [23] Engle, R.F. (2009). *Anticipating Correlations*, Princeton University Press.
- [24] Engle, R.F. & Gallo, J.P. (2006). A multiple indicator model for volatility using intra daily data, *Journal of Econometrics* **131**, 3–27.
- [25] Heston, S.L. (1993). A closed-form solution for options with stochastic volatility, with applications to bond and currency options, *Review of Financial Studies* **6**, 327–343.
- [26] Hubalek, F. & Sgarra, C. (2007). *On the Esscher Transforms and the other Equivalent Martingale Measures for the Barndorff-Nielsen and Shephard Stochastic Volatility Models with Jumps*. Thiele Research Report 2007 - 14.
- [27] Hull, J. & White, A. (1987). The pricing of options on assets with stochastic volatilities, *Journal of Finance* **42**, 281–300.
- [28] Jacod, J. (1994). *Limit of Random Measures Associated with the Increments of a Brownian Semimartingale*, Université Pierre et Marie Curie, Paris. Preprint number 120, Laboratoire de Probabilités.
- [29] Jacod, J., Li, Y., Mykland, P.A., Podolskij, M. & Vetter, M. (2009). Microstructure Noise in the Continuous Case: The Pre-averaging Approach, *Stochastic Processes and Their Applications* **119**, 2249–2276.
- [30] Meddahi, N. & Renault, E. (2004). Temporal aggregation of volatility models, *Journal of Econometrics* **119**, 355–379.
- [31] Mykland, P.A. & Zhang, L. (2006). ANOVA for diffusions and Ito processes, *Annals of Statistics* **34**, 1931–1963.
- [32] Mykland, P.A. & Zhang, L. (2009). Inference for continuous semimartingales observed at high frequency: a general approach, *Econometrica* forthcoming.
- [33] Nicolato, E. & Venardos, E. (2003). Option pricing in stochastic volatility models of the Ornstein-Uhlenbeck type, *Mathematical Finance* **13**, 445–466.
- [34] Pigorsch, C. & Stelzer, R. (2007). *A Multivariate Generalisation of the Ornstein-Uhlenbeck Stochastic Volatility Model*, forthcoming.
- [35] Stein, E.M. & Stein, J. (1991). Stock price distributions with stochastic volatility: an analytic approach, *Review of Financial Studies* **4**, 727–752.
- [36] Taylor, S.J. (2005). *Asset Price Dynamics, Volatility, and Prediction*, Princeton University Press, Princeton, New Jersey.
- [37] Todorov, V. & Tauchen, G. (2006). Simulation methods for Lévy-driven CARMA stochastic volatility models, *Journal of Business and Economic Statistics* **24**, 450–469.
- [38] Todorov, V. & Tauchen, G. (2008). *Volatility Jumps*, Department of Economics, Duke University, forthcoming.
- [39] Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach, *Bernoulli* **12**, 1019–1043.
- [40] Zhang, L., Mykland, P.A. & Aït-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high-frequency data, *Journal of the American Statistical Association* **100**, 1394–1411.
- [41] Zhou, B. (1998). Parametric and nonparametric volatility measurement, in *Nonlinear Modelling of High Frequency Financial Time Series*, C.L. Dunis & B. Zhou, eds, John Wiley & Sons, New York, Chapter 6, pp. 109–123.

Related Articles

Barndorff-Nielsen and Shephard (BNS) Models; Corridor Variance Swap; Econometrics of Diffusion Models; Exponential Lévy Models; GARCH Models; Itô's Formula; Lévy Processes; Market Microstructure Effects; Measurements Errors; Realized Volatility and Multipower Variation; Semimartingale; Stochastic Volatility Models; Stylized Properties of Asset Returns; Variance Swap; Volatility Swaps.

OLE E. BARNDORFF-NIELSEN & NEIL
SHEPHARD

Realized Volatility and Multipower Variation

Many financial markets effectively operate in continuous time with multiple transaction prices and quotes recorded each second. Although, such “ultra high-frequency” data are not useful for assessing the expected mean return of the underlying asset, it is highly informative regarding the strength of another key financial characteristic, namely, return volatility. In particular, it is feasible to estimate the realized return volatility over a fixed time period directly from high-frequency data with good precision without imposing a specific parametric structure on the return dynamics. Hence, as access to tick-by-tick data became more commonplace and the first empirical links between cumulative absolute return measures and the underlying volatility were established in the mid-1990s, the literature on extracting time-varying return variation measures from high-frequency data has grown dramatically. The initial developments focused on measures reflecting the actual squared return variation, loosely denoted realized volatility (RV), over daily and weekly frequencies. However, extended measures capturing different return moments and/or providing more robust inference regarding the continuous *versus* jump components of the return variation process are now an integral part of this literature. The potential applications to areas such as risk management, derivatives pricing, portfolio choice, market microstructure, and general asset pricing are basically unlimited. This article briefly reviews the main developments in this literature.

A general no-arbitrage setting for a continuously evolving logarithmic asset price is given by a jump-diffusive process. Formally, on an appropriate probability space, the stochastic logarithmic price process X_t is defined *via*

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s + \sum_{0 \leq s \leq t} \Delta X_s \quad (1)$$

where the predictable drift component, b_s , signifies the instantaneous (continuously compounded) mean return, the diffusive coefficient, σ_s , reflects the strength of the diffusive volatility, W_s denotes a standard Brownian motion, and $\Delta X_s := X_s - X_{s-}$

indicates the jump size and is nonzero only if the price jumps exactly at time s . In the definition of X_t , we implicitly assume certain (weak) regularity conditions and that jumps are of finite variation; some results below extend to the case of infinite variation jumps, [29]. Equation (1) is a generic representation of the jump-diffusive stochastic volatility model that serves as a workhorse for continuous-time finance.

We are interested in inference regarding the actual volatility displayed by X over a given interval of time $[0, T]$. In the (infeasible) scenario, where we have access to a continuous record of X_t , we can directly determine the actual realization of the so-called quadratic variation (QV) of X ,

$$QV_{[0,T]} = \int_0^T \sigma_s^2 ds + \sum_{0 \leq s \leq T} |\Delta X_s|^2 \quad (2)$$

Importantly, in the continuous-record case, we may also perfectly identify the jumps in the sample path and we thus “observe” the continuous and discontinuous parts of the QV and obtain the decomposition,

$$QV_{[0,T]}^c = \int_0^T \sigma_s^2 ds \quad \text{and} \quad QV_{[0,T]}^d = \sum_{0 \leq s \leq T} |\Delta X_s|^2 \quad (3)$$

The QV is of a particular interest as it is closely related to the (variance) risk of the asset. This is most transparent in a simplified setting. For a pure diffusion, if the (stochastic) volatility process is independent of the return innovations, and ignoring the (negligible) return variation induced by innovations to the mean drift, the returns are conditionally Gaussian with variance governed by $QV_{[0,T]}^c$. If the return interval is small, such as a day, we may safely ignore the mean drift altogether and for simplicity equate it to zero. We then have the distributional result,

$$(X_T - X_0) \Big| \{\sigma_s\}_{0 \leq s \leq T} \sim N\left(0, \int_0^T \sigma_s^2 ds\right) \quad (4)$$

Notice that the Gaussian distributional result is conditional on the realization of the so-called integrated variance which *a priori* is stochastic. In other words, the returns follow a Gaussian mixture distribution with the realization of the (diffusive) QV

governing the actual return variation. If this integrated variance process is persistent, that is, displays pronounced positive serial correlation, then the return distribution is symmetric, but displays both unconditional and conditional leptokurtosis along with volatility clustering. The main features, which may modify this characterization and induce both asymmetries and more extreme outliers in the conditional return distribution are correlation between the return innovations and the volatility, the so-called leverage effect, and the presence of jumps. Nonetheless, the interpretation of $QV_{[0,T]}$ as the relevant return variation measure remains intact. Moreover, the multivariate version of equation (4) applies as stated in [7].

In practice, we do not observe financial series continuously but rather obtain transaction price and quote data referring to specific points in time. Hence, the challenge is to design estimators from the discrete observations of X that can estimate continuous-time quantities like the QV and its decomposition into continuous and jump components. A closely related issue is the development of tests for the presence of jumps in the (partially) observed path of X . Suppose, we observe X at $0, \dots, i\Delta_n, \dots, [T/\Delta_n]\Delta_n$, where Δ_n is a small positive sampling interval and $[y]$ denotes the integer part of y . Since the time interval, T , is fixed, we cannot exploit standard “large (sample length) T ” asymptotics for developing estimation and testing procedures but instead appeal to continuous-record asymptotics, that is, $\Delta_n \rightarrow 0$. For example, we may think of $[0, T]$ as one trading day and assume that we observe the price process every 10 min, 5 min, 1 min, and so on. The exposition is split in three parts. We first define RV as a feasible estimator of QV and discuss its relation to the underlying return distribution. Next, the notion of multipower variation is introduced and we outline its use for robust inference regarding critical aspects of the quadratic return variation. In the following section, we discuss the application of realized multipower variation in formal testing for the presence of jumps. Finally, we briefly touch on procedures designed to mitigate the impact of market microstructure noise, which potentially allow for more efficient inference, as well as some multivariate generalizations of the results discussed in this article.

Realized Volatility

Realized volatility, also interchangeably referred to as *realized quadratic variation*, is defined as

$$RV_T = \sum_{i=1}^{[T/\Delta_n]} |\Delta_i^n X|^2 \quad (5)$$

where $\Delta_i^n X := X_{i\Delta_n} - X_{(i-1)\Delta_n}$. Some authors label RV_T the realized variance, and refer to its square root as the Realized volatility. However, in this article, we follow the early literature and term RV_T Realized volatility. RV_T is simply a cumulative sum of squared high-frequency observations. Its usefulness stems from the fact that it consistently estimates the (unobservable) QV as $\Delta_n \rightarrow 0$, that is,

$$RV_T \xrightarrow{\mathbb{P}} QV_T \quad (6)$$

A few comments are in order. First, the estimator is model-free and applies very broadly independent of parametric model assumptions. Moreover, it remains valid in the multivariate case as well. Second, the measure estimates the return variation over an interval and not the underlying spot volatility at any given point. In fact, inference regarding spot volatility is problematic without restrictive assumptions owing to the lack of a continuous record of price observations and the impact of microstructure frictions at ultra high frequencies. Third, RV_T resembles a rolling sample variance estimator. This statistic, labeled “historical volatility” has precedents in the literature. Fourth, Merton [36] notes that, in theory, high-frequency data can provide perfect inference whenever volatility is locally constant. Direct linkage of an empirical RV measure based on intraday returns to an underlying (stochastic) quadratic return variation appears first in [3]. Fifth, one trading day is the natural measurement period owing to the pronounced intraday volatility patterns and significant news announcement effects. These features render interpretation of realized return variation over intraday periods much more complex, [4]. Finally, it is important to recognize that RV_T provides an *ex post* estimate of the return variation, which is conceptually distinct from the *ex ante* conditional return variance.

In order to assess the accuracy with which RV_T measures QV_T , a central limit theorem (CLT)—characterizing its behavior as the sampling frequency

increases—is useful. In the simplest case without jumps it follows from [14, 16, 17] that

$$\frac{1}{\sqrt{\Delta_n}} (RV_T - QV_T) \xrightarrow{\mathcal{L}-s} \epsilon \times \sqrt{2 \int_0^T \sigma_s^4 ds} \quad (7)$$

where $\xrightarrow{\mathcal{L}-s}$ means stable convergence in law and ϵ is a standard normal variable, defined on an extension of the original probability space. Stable convergence in law strengthens the usual convergence in law by guaranteeing that the convergence in equation (7) is joint with any random variable defined on the original probability space. This property is critical for feasible asymptotic inference as well as in the application of the delta method. We further note that for two nonoverlapping intervals, the measurement errors for the QV are serially uncorrelated. Evidently, the precision of RV_T can be gauged formally only if we can obtain a measure of the so-called integrated quarticity, $IQ_T = \int_0^T \sigma_s^4 ds$, from a given set of discrete price observations. This is, in fact, feasible through the general realized multipower variation statistics introduced below. Finally, we note that bootstrapping may improve finite sample inference, [26].

The result in equation (7) can be extended to the case when X contains jumps, [29]. The realized variation then provides a consistent estimate of the overall QV. If we want to further decompose it into the continuous and jump parts, we need to exploit results pertaining to the theory for general multipower variation statistics.

Multipower Variation

The initial example of a multipower variation process is the so-called bipower variation (BV) proposed by Barndorff-Nielsen and Shephard [18, 19] and defined as

$$BV_T = \frac{\pi}{2} \sum_{i=2}^{[T/\Delta_n]} |\Delta_{i-1}^n X| |\Delta_i^n X| \quad (8)$$

The above papers prove that, even in the presence of a jump process,

$$BV_T \xrightarrow{\mathbb{P}} QV_T^c \quad (9)$$

Hence, it provides a consistent estimator for the integrated variance and we may thus further consistently estimate the jump contribution to QV *via* the difference between RV_T and BV_T . However, we note that the associated CLT for BV_T in the continuous case will not hold when jumps are present.

More generally, we may define realized multipower variation by raising m successive absolute price increments to arbitrary powers, indexed by the vector $\mathbf{p} = (p_1, \dots, p_m)$, where $p_j \geq 0$, $j = 1, \dots, m$

$$V(X; \mathbf{p}; \Delta_n)_T = (\mu_{p_1} \dots \mu_{p_m})^{-1} \times \sum_{i=m}^{[T/\Delta_n]} |\Delta_{i-m+1}^n X|^{p_m} \dots |\Delta_i^n X|^{p_1} \quad (10)$$

for μ_a denoting the a th absolute moment of a standard normal. By appropriate scaling of the realized multipower variation, *in the absence of jumps*, we obtain the corresponding convergence result

$$\Delta_n^{1-p/2} V(X; \mathbf{p}; \Delta_n)_T \xrightarrow{\mathbb{P}} \int_0^T \sigma_s^p ds \quad (11)$$

where $p = p_1 + \dots + p_m$. In analogy to the bipower variation case, this result is robust to the presence of jumps provided $\max\{p_i\}_{i=1, \dots, m} < 2$. Further generalizations and asymptotic properties are explored in [14] for X continuous. Some of these properties are extended to the case when X contains jumps in [21].

A prominent special case of multipower variation arises when $\mathbf{p}$ is a scalar. Then $V(X; p; \Delta_n)_T$ is labeled realized power variation of order p . Obviously, $V(X; 2; \Delta_n)_T$ coincides with the RV estimator. In addition, a natural estimator of integrated quarticity under the null hypothesis of no jumps may be based on the realized fourth power variation, $V(X; 4; \Delta_n)_T$. While this statistic is inconsistent under the jump alternative, many realized multipower variation statistics are consistent in the presence of jumps, including the so-called realized tripower variation with $\mathbf{p} = (4/3, 4/3, 4/3)$ and the realized quadpower variation with $\mathbf{p} = (1, 1, 1, 1)$.

An alternative way of estimating the continuous part of the QV is to use the truncated squared power variation proposed by Mancini [33, 34] and also

analyzed by Jacod [29]. It is formally defined as

$$\tilde{V}(X, 2, \Delta_n)_T = \sum_{i=1}^{\lfloor T/\Delta_n \rfloor} |\Delta_i^n X|^2 1_{\{|\Delta_i^n X| \leq \alpha(\Delta_n)^\gamma\}} \quad (12)$$

for arbitrary $\alpha > 0$ and $\gamma \in (0, 1/2)$. The intuition is simple—we discard increments higher than a given threshold in the summation of the squared returns—thus effectively discarding the impact of jumps.

Given their consistency for QV_T and QV_T^c and the absence of serial correlation in the associated measurement errors, RV_T and BV_T serve as a natural basis for measuring and forecasting return volatility. The literature is too extensive for a thorough review and so we only provide an account of a few established findings. First, Andersen *et al.* [7], demonstrate that reduced-form time-series models for a realized volatility within the ARFIMA class generate forecasts which dominate traditional stochastic volatility and GARCH models estimated from daily return data. The long memory dependence incorporated in the ARFIMA setting is invariably highly significant and an important factor in forecast performance. Another key feature in the superior forecast performance is the ability of RV_T to adapt quickly to shifts in the underlying level of volatility. This allows forecasts to be conditioned on a more accurate assessment of the current volatility state compared to models based on daily returns. Second, Andersen *et al.* [5] find that forecast performance may be boosted further through a decomposition of RV_T into the continuous and jump parts. The diffusive volatility is the main source of long-range dependence, so separating out the jumps allows for a more refined measure of the relevant volatility state and improves forecast precision correspondingly. An alternative approach exploiting a mix of different realized power variation statistics for forecasting is the so-called MIDAS regressions proposed by Ghysels *et al.* [25]. Third, Andersen *et al.* [10] provide an analytic framework for gauging the performance of reduced form a realized volatility forecasts across a broad class of popular diffusive stochastic volatility specifications, finding that the loss of predictive ability is minor relative to using (infeasible) forecasts based on the true underlying model. The usefulness of reduced form a realized volatility and (in the multivariate extension) covariation forecast models from practical perspectives is

documented, for example, in [24] and [13] for portfolio allocation, in [8] for dynamic estimation of systematic (beta) market risk, in [2] for specification testing of term structure models, and in [12] for option pricing and associated trading strategies.

Realized multipower variation also facilitates estimating general parametric continuous-time models, which are challenging because of the latency of the stochastic volatility and the presence of price jumps. Barndorff-Nielsen and Shephard [16, 20] consider estimation by computing second-order moments of realized volatility. Meddahi [35] provides moments of realized volatility across a broad class of models. Closed-form expressions for moments of general realized multipower variation statistics (e.g., bipower variation), however, are typically not known, even when the moments of their continuous-time counterparts can be computed. Bollerslev and Zhou [22] estimate affine jump-diffusion models by treating the realized volatility as the unobservable quadratic variation in a method-of-moments procedure. They show *via* simulations that the impact of the measurement error on the parameter estimates is small. Corradi and Distaso [23] provide formal justification for the estimator of [22]. They consider joint asymptotics, both long time span and continuous record, and show that (in the general case) when the intraperiod sampling frequency increases faster than sample length, the measurement error of realized volatility (and in some cases bipower variation) is asymptotically negligible. Todorov [38] further generalizes these results. He proves the uniform integrability of realized multipower variation statistics under general specifications for the price jump process and demonstrates that the realized multipower variation can be used effectively in making semiparametric inference about general classes of continuous-time models containing jumps.

Jump Testing

Another important application of realized multipower variation is to determine whether, *on a given path*, the process X contains jumps or not. This may again be accomplished in entirely model-independent fashion, that is, without imposing additional structure on X besides specification (1). Barndorff-Nielsen and Shephard [18, 19] propose testing for jumps by determining whether $QV_{[0,T]}^d$ is statistically different from zero. For this purpose, they derive the joint

asymptotic distribution of RV and BV conditional on the path of X not containing jumps.

Huang and Tauchen [28] discuss various transformations of these tests which often lead to significant improvements in finite sample test performance. They also document that these procedures generally provide reliable tests. Alternative tests for the *null hypothesis of no jumps* can be derived by replacing the bipower variation with other realized multipower variation estimates of QV_T^c or with the truncated squared power variation (see also [31] for an alternative but related jump test).

Recently, Ait-Sahalia and Jacod [1] proposed an attractive alternative to these tests. They constructed a test statistic as a ratio of power variation of order four computed over two different frequencies

$$\Phi_n^{(j)} = \frac{V(X; 4; k\Delta_n)_T}{V(X; 4; \Delta_n)_T} \quad (13)$$

where $k \geq 2$ is an integer (typically 2 or 3). This test statistic behaves very differently depending on whether X contains jumps on the (partially) observed path or not, since,

$$\Phi_n^{(j)} \xrightarrow{\mathbb{P}} \begin{cases} 1 & \text{if } X \text{ contains jumps} \\ k & \text{if } X \text{ does not contain jumps} \end{cases} \quad (14)$$

Ait-Sahalia and Jacod [1] derive the CLT for $\Phi_n^{(j)}$ both when the given path of X contains jumps and when it does not. This allows for tests of both the null of no jumps (like above) and the null of jumps.

Finally, Lee and Mykland [32] pursue a different strategy as they test for jumps at each single high-frequency observation. The motivation is that, under the null hypothesis of no jumps, the high-frequency increments, standardized by the estimated local volatility, are asymptotically normally distributed. It has the convenient feature that the exact timing of significant jumps within the trading period is identified. Andersen *et al.* [9] explore a similar strategy determining the exact location of, potentially multiple, jumps from uniform test statistics across all high-frequency returns over each trading day.

Some Important Extensions

An important feature ignored in our exposition is the impact of so-called market microstructure noise.

This refers to the patterns induced in high-frequency returns owing to features such as a discrete price grid, the absence of continuous trading, and the presence of a bid–ask spread. Hence, the observed price may be seen as a noisy indicator of an underlying ideal price contaminated by a (small) noise process. This noise component induces a bias in the empirical RV measures which tends to grow with increasing sampling frequency. As such, when a realized volatility is computed from ultra high-frequency returns, it is critical to adjust for the impact of microstructure noise. Andersen *et al.* [6] suggest gauging the highest sampling frequency at which the systematic bias induced by the noise is negligible, *via* a so-called volatility signature plot, and use that frequency for the realized volatility computation. However, it is potentially more efficient to sample more frequently and adjust for the noise component. This approach is pursued in a number of papers in the recent years; see, for example, [11, 15, 27, 39, 40]. While the asymptotic theory has developed for the case of a pure diffusive price process, the properties of the various procedures in the presence of jumps remain largely unknown. Recently, Podolskij and Vetter [37] propose to modify the bipower variation in a way which is robust both to the microstructure noise and price jumps (of finite activity).

In addition, our exposition has focused on the univariate case. If X is continuous, the multivariate extension is conceptually straightforward and has been done in [14, 17] (although the practical application can be challenging). If X contains jumps, matters are much more complicated, as realized multipower variation can behave very differently depending on whether the jumps in individual series arrive jointly or not. This is exploited by Jacod and Todorov [30], who propose tests for both whether the jumps in the individual components arrive together or not without any restriction on the possible jump dependence.

In summary, the realized volatility and related realized multipower variation literature is vibrant. Given the large efficiency gains obtained from volatility measurements exploiting high-frequency data relative to daily data and the increasing availability of tick-by-tick data, both the theoretical and the empirical research within this area will surely continue to grow in the future.

Acknowledgments

Andersen's work was supported by a grant from the NSF to the NBER and CREATES funded by the Danish National Research Foundation.

References

- [1] Ait-Sahalia, Y. & Jacod, J. (2009). Testing for jumps in a discretely observed process, *Annals of Statistics* **37**, 184–222.
- [2] Andersen, T.G. & Benzoni, L. (2009). Do Bonds span volatility risk in the U.S. Treasury Market? A specification test for affine term structure models, *Journal of Finance* forthcoming.
- [3] Andersen, T.G. & Bollerslev, T. (1998). Answering the critics: yes, standard volatility models do provide accurate forecasts, *International Economic Review* **39**, 885–905.
- [4] Andersen, T.G. & Bollerslev, T. (1998). Deutsche mark-dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies, *Journal of Finance* **53**, 219–265.
- [5] Andersen, T., Bollerslev, T. & Diebold, F.X. (2007). Roughing it up: including jump components in the measurement, modeling, and forecasting of return volatility, *Review of Economics and Statistics* **89**, 701–720.
- [6] Andersen, T.G., Bollerslev, T., Diebold, F.X. & Labys, P. (2000). Great realizations, *Risk* **13**, 105–108.
- [7] Andersen, T.G., Bollerslev, T., Diebold, F.X. & Labys, P. (2003). Modeling and forecasting realized volatility, *Econometrica* **71**, 579–625.
- [8] Andersen, T.G., Bollerslev, T., Diebold, F.X. & Wu, G. (2006). Realized beta: persistence and predictability, in *Advances in Econometrics: Econometric Analysis of Economic and Financial Time Series*, T. Fomby & D. Terrell, eds, Elsevier Science, Vol. 20.
- [9] Andersen, T.G., Bollerslev, T. & Dobrev, D. (2007). No-arbitrage semimartingale restrictions for continuous-time volatility models subject to leverage effects, jumps and i.i.d. noise: theory and testable distributional assumptions, *Journal of Econometrics* **138**, 125–180.
- [10] Andersen, T.G., Bollerslev, T. & Meddahi, N. (2004). Analytic evaluation of volatility forecasts, *International Economic Review* **45**, 1079–1110.
- [11] Bandi, F. & Russell, J. (2006). Separating microstructure noise from volatility, *Journal of Financial Economics* **79**, 655–692.
- [12] Bandi, F.M., Russell, J.R. & Yang, C. (2008). Realized volatility forecasting and option pricing, *Journal of Econometrics* **71**, 34–46.
- [13] Bandi, F., Russell, J. & Zhu, J. (2008). Using high-frequency data in dynamic portfolio choice, *Econometric Reviews* **27**, 163–198.
- [14] Barndorff-Nielsen, O.E., Graversen, S., Jacod, J., Podolskij, M. & Shephard, N. (2005). A central limit theorem for realised power and bipower variations of continuous semimartingales, in *From Stochastic Analysis to Mathematical Finance, Festschrift for Albert Shiryaev*, Y. Kabanov & R. Lipster, eds, Springer.
- [15] Barndorff-Nielsen, O.E., Hansen, P., Lunde, A. & Shephard, N. (2008). Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise, *Econometrica* **76**, 1481–1536.
- [16] Barndorff-Nielsen, O.E. & Shephard, N. (2002). Econometric analysis of realised volatility and its use in estimating stochastic volatility models, *Journal of the Royal Statistical Society: Series B* **64**, 253–280.
- [17] Barndorff-Nielsen, O.E. & Shephard, N. (2004). Econometric analysis of realised covariation: high-frequency covariance, regression and correlation in financial economics, *Econometrica* **72**, 885–925.
- [18] Barndorff-Nielsen, O.E. & Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps, *Journal of Financial Econometrics* **2**, 1–37.
- [19] Barndorff-Nielsen, O.E. & Shephard, N. (2006). Econometrics of testing for jumps in financial economics using bipower variation, *Journal of Financial Econometrics* **4**, 1–30.
- [20] Barndorff-Nielsen, O.E. & Shephard, N. (2006). Impact of jumps on returns and realised variances: econometric analysis of time-deformed Lévy processes, *Journal of Econometrics* **131**, 217–252.
- [21] Barndorff-Nielsen, O.E., Shephard, N. & Winkel, M. (2006). Limit theorems for multipower variation in the presence of jumps in financial econometrics, *Stochastic Processes and their Applications* **116**, 796–806.
- [22] Bollerslev, T. & Zhou, H. (2002). Estimating stochastic volatility diffusion using conditional moments of integrated volatility, *Journal of Econometrics* **109**, 33–65.
- [23] Corradi, V. & Distaso, W. (2006). Semiparametric comparison of stochastic volatility models using realized measures, *Review of Economic Studies* **73**, 635–667.
- [24] Fleming, J.R., Kirby, C.W. & Ostdiek, B.F. (2003). The economic value of volatility timing using realized volatility, *Journal of Financial Economics* **67**, 473–509.
- [25] Ghysels, E., Santa-Clara, P. & Valkanav, R. (2006). Predicting volatility: how to get most out of returns data sampled at different frequencies, *Journal of Econometrics* **131**, 59–95.
- [26] Goncalves, S. & Meddahi, N. (2009). Bootstrapping realized volatility, *Econometrica* **77**, 283–306.
- [27] Hansen, P. & Lunde, A. (2006). Realized variance and market microstructure noise, *Journal of Business and Economic Statistics* **24**, 127–161.
- [28] Huang, X. & Tauchen, G. (2005). The relative contributions of jumps to total variance, *Journal of Financial Econometrics* **3**, 456–499.
- [29] Jacod, J. (2008). Asymptotic properties of power variations and associated functionals of semimartingales, *Stochastic Processes and their Applications* **118**, 517–559.

-
- [30] Jacod, J. & Todorov, V. (2009). Testing for common arrivals of jumps for discretely observed multidimensional processes, *Annals of Statistics* **37**, 1792–1838.
- [31] Jiang, G. & Oomen, R. (2008). A new test for jumps in asset prices, *Journal of Econometrics* **144**, 352–370.
- [32] Lee, S. & Mykland, P. (2008). Jumps in financial markets: a new nonparametric test and jump dynamics, *Review of Financial Studies* **21**, 2535–2563.
- [33] Mancini, C. (2001). Disentangling the jumps of the diffusion in a geometric jumping Brownian motion, *Giornale dell'Istituto Italiano degli Attuari* **LXIV**, 19–47.
- [34] Mancini, C. (2009). Nonparametric threshold estimation for models with stochastic diffusion coefficient and jumps, *Scandinavian Journal of Statistics* **36**, 270–296.
- [35] Meddahi, N. (2002). Theoretical comparison between integrated and realized volatility, *Journal of Applied Econometrics* **17**, 479–508.
- [36] Merton, R.C. (1980). On estimating the expected return on the market, *Journal of Financial Economics* **8**, 323–361.
- [37] Podolskij, M. & Vetter, M. (2009). Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps, *Bernoulli* **15**, 634–658.
- [38] Todorov, V. (2009). Estimation of continuous-time stochastic volatility models with jumps using high-frequency data, *Journal of Econometrics* **148**, 131–148.
- [39] Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach, *Bernoulli* **19**, 1019–1043.
- [40] Zhang, L., Mykland, P. & Ait-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high frequency data, *Journal of the American Statistical Association* **100**, 1394–1411.

TORBEN G. ANDERSEN & VIKTOR TODOROV

Jump Processes

Processes with Jumps in General

Although, historically, models in mathematical finance were based on Brownian motion and thus are models with continuous price paths, jump processes now play a key role across all areas of finance (see, e.g., [5]). One reason for this move into a new class of processes is that because of their distributional properties, diffusions in many cases cannot provide a realistic picture of empirically observed facts. Another reason is the enormous progress made in understanding and handling jump processes due to the development of semimartingale theory on one side and of computational power on the other side.

The simplest jump process is a process with just one jump. Let T be a random time—actually a stopping time with respect to an information structure given by a filtration $(\mathcal{F}_t)_{t \geq 0}$ —then

$$X_t = \mathbb{1}_{\{T \leq t\}} \quad (t \geq 0) \quad (1)$$

has the value 0 until a certain event occurs and then 1. Although this process looks simple, it is important in modeling credit risk, namely as the process that describes the time of default of a company. The next step leads to processes that have integer values with positive jumps of size 1 only, the so-called *counting processes* $(X_t)_{t \geq 0}$. X_t describes the number of events that have occurred between time 0 and t . This could be the number of defaults in a large credit portfolio or the number of claims customers report to an insurance company. The standard case is a *Poisson process* $(N_t)_{t \geq 0}$, where the distribution of X_t is given by a Poisson distribution with parameter λt . Equivalently, one can describe this process by requiring that the waiting times between successive jumps are independent, exponentially distributed random variables with parameter λ .

The natural extension is a *compound Poisson process* $(X_t)_{t \geq 0}$, that is, a process with stationary independent increments where the jump size is no longer 1, but given by a probability law μ . Let $(Y_k)_{k \geq 1}$ be a sequence of independent random variables with distribution $\mathcal{L}(Y_k) = \mu$ for all $k \geq 1$. Denote by $(N_t)_{t \geq 0}$ a standard Poisson process with parameter $\lambda > 0$ as described above, which is independent of $(Y_k)_{k \geq 1}$.

Then we can represent $(X_t)_{t \geq 0}$ in the following form:

$$X_t = \sum_{k=1}^{N_t} Y_k \quad (2)$$

A typical application of compound Poisson processes is to model the cumulative claim size up to time t in a portfolio of insurance contracts where the individual claim size is distributed according to μ . For the sake of analytical tractability, it is often useful to *compensate* this process, that is, to subtract the average claim size $E[X_t]$. Assuming that μ has a finite expectation and using stationarity and independence, we conclude $E[X_t] = tE[X_1]$ and therefore get the following representation:

$$X_t = tE[X_1] + (X_t - E[X_t]) \quad (3)$$

The compensated process $(X_t - E[X_t])_{t \geq 0}$ is a martingale and, therefore, equation (3) is a decomposition of the process in a linear drift $E[X_1] \cdot t$ and a martingale. Representation (3) motivates the definition of a general *semimartingale* as a process that is adapted to a filtration $(\mathcal{F}_t)_{t \geq 0}$, has paths that are right-continuous and have left limits (càdlàg paths), and allows a decomposition

$$X_t = X_0 + V_t + M_t \quad (t \geq 0) \quad (4)$$

where $V = (V_t)_{t \geq 0}$ is an adapted càdlàg process of finite variation and $M = (M_t)_{t \geq 0}$ is a local martingale. There exist processes that are not semimartingales. An important class of examples is fractional Brownian motions with the exception of usual Brownian motion, which is a semimartingale. We do not go beyond semimartingales in this discussion mainly because for semimartingales there is a well-developed theory of stochastic integration, a fact which is crucial for modeling in finance.

In general, the representation (4) is not unique. It becomes unique with a *predictable* process V if we consider *special semimartingales*. A semimartingale can be made special by taking the big jumps away, for example, jumps with absolute jump size larger than 1. This follows from the well-known fact that a semimartingale with bounded jumps is special [11, I.4.24]. Denote by $\Delta X_t = X_t - X_{t-}$ the jump at time t if there is any. Then

$$X_t - \sum_{s \leq t} \Delta X_s \mathbb{1}_{\{|\Delta X_s| > 1\}} \quad (5)$$

has bounded jumps. Further, let us note that any local martingale M (with $M_0 = 0$) admits a unique (orthogonal) decomposition into a local martingale with continuous paths M^c and a purely discontinuous, local martingale M^d ([11, I.4.18]). Assuming $X_0 = 0$, we got the following unique representation for semimartingales:

$$X_t = V_t + M^c + M^d + \sum_{s \leq t} \Delta X_s \mathbb{1}_{\{|\Delta X_s| > 1\}} \quad (6)$$

To analyze M^d in more detail, we introduce the *random measure of jumps*

$$\mu^X(\omega; dt, dx) = \sum_{s > 0} \mathbb{1}_{\{\Delta X_s(\omega) \neq 0\}} \varepsilon_{(s, \Delta X_s(\omega))}(dt, dx) \quad (7)$$

where ε_a denotes as usual the unit mass in a . Thus, μ^X is a random measure which, for fixed ω , places a point mass of size 1 on each pair $(s, \Delta X_s(\omega)) \in \mathbb{R}_+ \times \mathbb{R}$ whenever for this ω the process jumps by $\Delta X_s(\omega)$ at time s . Expressed differently for any Borel subset $B \subset \mathbb{R}$, $\mu^X(\omega; [0, t] \times B)$ counts the number of jumps with size in B which can be observed along the path $(X_s(\omega))_{0 \leq s \leq t}$. With this notation, equation (5) can be written as

$$X_t - \int_0^t \int_{\mathbb{R}} x \mathbb{1}_{\{|x| > 1\}} \mu^X(ds, dx) \quad (8)$$

The purely discontinuous local martingale M^d , that is, the process of *compensated jumps* of absolute size less than 1, has then the following form:

$$M_t^d = \int_0^t \int_{\mathbb{R}} x \mathbb{1}_{\{|x| \leq 1\}} (\mu^X - \nu^X)(ds, dx) \quad (9)$$

where ν^X is another random measure, the (*pre-*dictable) *compensator* of μ^X . Whereas μ^X counts the exact number of jumps, ν^X roughly stands for the expected, that is, the average number of jumps. The integral with respect to $\mu^X - \nu^X$ in equation (9), in general, cannot be separated in an integral with respect to μ^X and another one with respect to ν^X . This is because the sum of the small jumps

$$\sum_{s \leq t} \Delta X_s \mathbb{1}_{\{|\Delta X_s| \leq 1\}} = \int_0^t \int_{\mathbb{R}} x \mathbb{1}_{\{|x| \leq 1\}} \mu^X(ds, dx) \quad (10)$$

does not converge, in general.

Lévy Processes

For many applications, it is sufficient to reduce generality and to consider the subclass of *Lévy processes*, that is, processes with stationary and independent increments. The components in equations (6) and (9) are then

$$\begin{aligned} V_t &= bt \quad (t \geq 0) \\ M_t^c &= \sqrt{c} W_t \quad (t \geq 0) \\ \nu^X([0, t] \times B) &= tK(B) \end{aligned} \quad (11)$$

where b and c are real numbers, $c \geq 0$, $(W_t)_{t \geq 0}$ is a standard Brownian motion, and K , the *Lévy measure*, is a (possibly infinite) measure on the real line that satisfies $\int (1 \wedge x^2) K(dx) < \infty$. The law of X is completely determined by the *triplet of local characteristics* (b, c, K) since these are the parameters that appear in the classical Lévy–Khinchine formula. This formula expresses the *characteristic function* $\varphi_{X_t}(u) = E[\exp(iuX_t)]$ in the form

$$\varphi_{X_t}(u) = \exp(t\psi(u)) \quad (12)$$

with the characteristic exponent

$$\begin{aligned} \psi(u) &= iub - \frac{1}{2}u^2c \\ &+ \int (\mathrm{e}^{iux} - 1 - iux \mathbb{1}_{\{|x| \leq 1\}}) K(dx) \end{aligned} \quad (13)$$

The *truncation function* $h(x) = x \mathbb{1}_{\{|x| \leq 1\}}$ could be replaced by other versions of truncation functions, for example, smooth functions that are identical to the identity in a neighborhood of the origin and go to 0 outside this neighborhood. Changing h affects the drift parameter b , but not c or K . All the information on the jump behavior of the process $(X_t)_{t \geq 0}$ is contained in K . The frequency of large jumps, expressed by the weight K puts on the tails, determines finiteness of the moments of the process as the following result states (for proofs of the propositions see [15]).

Proposition 1 *Let $X = (X_t)_{t \geq 0}$ be a Lévy process with Lévy measure K . Then $E[|X_t|^p]$ is finite for any $p \in \mathbb{R}_+$ if and only if $\int_{\{|x| > 1\}} |x|^p K(dx) < \infty$.*

We note that if X_1 and consequently any X_t has finite expectation, then one does not have to

truncate in equation (13), that is, $h(x) = x \mathbf{1}_{\{|x| \leq 1\}}$ can be replaced by $h(x) = x$.

The sum of the big jumps which is subtracted in equation (5) is finite since there are only finitely many of them from 0 to t for every path. The *fine structure* of the paths is determined by the frequency of the small jumps. A process is said to have *finite activity* if almost all paths have only a finite number of jumps along finite time intervals. The simplest examples are Poisson and compound Poisson processes. A process is said to have *infinite activity* if almost all paths have infinitely many jumps along any time interval of finite length.

Proposition 2 *Let $X = (X_t)_{t \geq 0}$ be a Lévy process with Lévy measure K . Then X has finite activity if $K(R) < \infty$ and has infinite activity if $K(R) = \infty$.*

Since a Lévy measure has *a priori* finite mass in the tails, that is, $\int_{\{|x| > 1\}} K(dx) < \infty$, the finiteness of $K(R)$ means finiteness of $\int_{\{|x| \leq 1\}} K(dx)$. Consequently, having a finite or an infinite number of jumps along finite time intervals is determined by the mass of K around the origin. From the distribution of mass around the origin, one can also see if the sum of (infinitely many) small jumps converges or not. First, let us recall that a standard Brownian motion has paths of infinite variation. Therefore, a Lévy process has *infinite variation* as soon as it has a continuous martingale component, that is, $c > 0$ in equation (11), but infinite variation can also come from the jumps.

Proposition 3 *Let $X = (X_t)_{t \geq 0}$ be a Lévy process with triplet (b, c, K) . Then almost all paths of X have finite variation if $c = 0$ and $\int_{\{|x| \leq 1\}} |x| K(dx) < \infty$. If this integral is infinite or $c > 0$, then almost all paths of X have infinite variation.*

Figure 1 shows a simulated path of a purely discontinuous, infinite activity process with finite variation, whereas Figure 2 shows a corresponding path with infinite variation.

Important Examples

Now we discuss some of the standard examples. The Poisson process with intensity parameter λ that we considered at the beginning has a finite number of jumps in any finite time interval and is constant between successive jumps. In terms of equation (11),

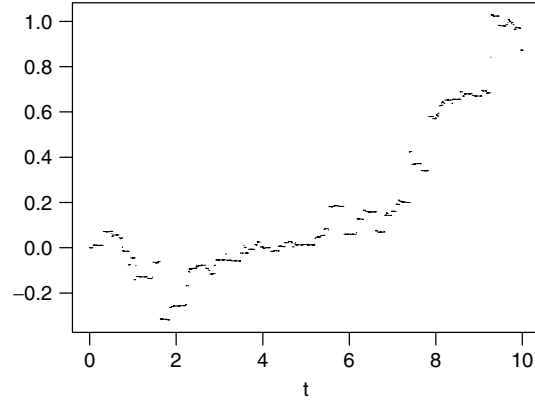


Figure 1 Simulation of a purely discontinuous Lévy process with infinite activity and finite variation

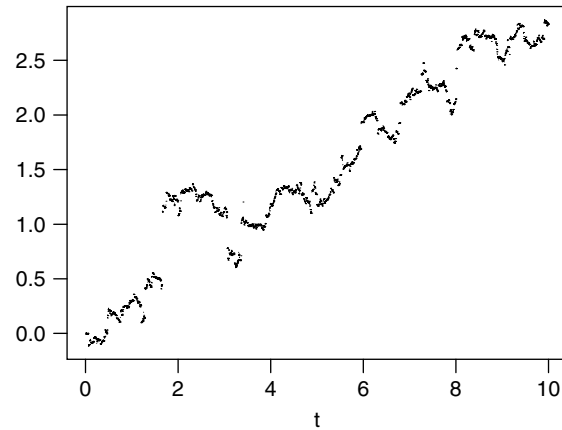


Figure 2 Simulation of a purely discontinuous Lévy process with infinite activity and infinite variation

it is characterized by $b = E[X_1] = \lambda$, $c = 0$, and $K = \lambda \varepsilon_1$. For the compound Poisson process (2), the unit mass ε_1 in K is replaced by a probability measure μ , the law of Y_1 , that is, $K = \lambda \mu$. For the drift parameter b , one gets $\lambda E[Y_1]$. One gets a *Lévy jump diffusion* by adding a general drift term bt and a scaled Brownian motion to equation (2),

$$X_t = bt + \sqrt{c}W_t + \sum_{k=1}^{N_t} Y_k \quad (14)$$

This is the model introduced by Merton [14] to describe asset returns. Merton chose normally distributed variables Y_k . In an article, Kou [12] used

double-exponentially distributed jump sizes Y_k . If one replaces in equation (14) the Brownian motion with drift by a general diffusion process one gets a *jump diffusion*. The key property of jump diffusions is that one adds only a finite number of jumps in any finite time interval to a process with continuous paths. In other words, the jump times can be given by successive stopping times $T_1 < T_2 < T_3 < \dots$. We also note that the distribution of X_t is not known for diffusions, in general. The same holds for jump diffusions. This reduces their applicability in mathematical finance. A key advantage of the pure jump Lévy processes that we discuss now is that they are distributionally very flexible and the distributions are known explicitly.

Generalized hyperbolic (GH) Lévy motions $(X_t)_{t \geq 0}$ (see *Generalized hyperbolic models* in this encyclopedia or in [6, 9]) represent a very large class of Lévy processes which are generated by GH distributions [1], that is, the distribution of X_1 , $\mathcal{L}(X_1)$, is GH. Using equation (12) all other distributions $\mathcal{L}(X_t)$ are determined. GH distributions have an explicit Lebesgue density as has the corresponding Lévy measure. GH distributions can be represented as normal mean–variance mixtures, where the mixing distribution is a generalized inverse Gaussian (GIG) distribution. Moments of any order exist. Since $c = 0$ in equation (13) GH Lévy motions have purely discontinuous paths. They are infinite activity processes. Important subclasses are hyperbolic Lévy motions ([7]) and normal inverse Gaussian (NIG) Lévy motions ([2]). Many well-known distributions can be obtained as limiting cases of GH distributions, which generate the corresponding processes [10]. Among those are the variance gamma distributions (see [13]), scaled and shifted Cauchy distributions, shifted Student- t distributions, GIG distributions, and the gamma, as well as the normal distributions.

The CGMY process introduced in [4] is another purely discontinuous Lévy process, which can be defined by the Lévy density of $\mathcal{L}(X_1)$:

$$g_{\text{CGMY}}(x) = \begin{cases} C \frac{\exp(-G|x|)}{|x|^{1+Y}} & x < 0 \\ C \frac{\exp(-Mx)}{x^{1+Y}} & x > 0 \end{cases} \quad (15)$$

where $Y \in (-\infty, 2)$. The process has infinite activity iff $Y \in [0, 2)$ and it has infinite variation iff $Y \in [1, 2)$. For $Y = 0$, it reduces to the variance gamma process.

A very classical class is given by α -stable Lévy processes where $0 < \alpha \leq 2$. For $\alpha = 2$ one gets Brownian motion, whereas for $\alpha < 2$ one gets purely discontinuous processes. Only for three special cases explicit densities are known: the Gaussian, the Cauchy, and the Lévy distribution.

A tractable extension of Lévy processes are time-inhomogeneous, Lévy processes that is, processes with independent increments and absolutely continuous characteristics, called *PIIAC* in [11]. For any fixed t , the triplet of $\mathcal{L}(X_t)$ for these processes is given in the form $b = \int_0^t b_s ds$, $c = \int_0^t c_s ds$, and $K(dx) = \int_0^t K_s(dx) ds$. This class of processes has been used extensively in the context of interest rate models (see e.g., [8]).

Jump processes with paths that are rather different from those discussed so far were introduced by Barndorff-Nielsen and Shephard [3] in the context of stochastic volatility models. Let $(Z_t)_{t \geq 0}$ be a *subordinator*, that is, a Lévy process, starting at 0 with increasing paths and consequently without a Gaussian component. The volatility process $(\sigma_t^2)_{t \geq 0}$ is modeled by an Ornstein–Uhlenbeck-type stochastic differential equation

$$d\sigma_t^2 = -\lambda\sigma_t^2 dt + dZ_{\lambda t} \quad (16)$$

for some $\lambda > 0$. The solution $(\sigma_t^2)_{t \geq 0}$ moves up entirely by jumps and then tails off exponentially. σ_t is fed into a Brownian semimartingale that then represents the price process.

References

- [1] Barndorff-Nielsen, O.E. (1978). Hyperbolic distributions and distributions on hyperbolae, *Scandinavian Journal of Statistics* **5**, 151–157.
- [2] Barndorff-Nielsen, O.E. (1998). Processes of normal inverse Gaussian type, *Finance and Stochastics* **2**(1), 41–68.
- [3] Barndorff-Nielsen, O.E. & Shephard, O.E. (2001). Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics, *Journal of the Royal Statistical Society, Series B* **63**, 167–207.
- [4] Carr, P., Geman, H., Madan, D. & Yor, M. (2002). The fine structure of asset returns: an empirical investigation, *Journal of Business* **75**, 305–332.
- [5] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, Chapman & Hall/CRC.
- [6] Eberlein, E. (2001). Application of generalized hyperbolic Lévy motions to finance, in *Lévy Processes. Theory and Applications*, O.E. Barndorff-Nielsen, T. Mikosch & S.I. Resnick, eds, Birkhäuser, 319–336.

-
- [7] Eberlein, E. & Keller, U. (1995). Hyperbolic distributions in finance, *Bernoulli* **1**(3), 281–299.
- [8] Eberlein, E. & Kluge, W. (2006). Exact pricing formulae for caps and swaptions in a Lévy term structure model, *Journal of Computational Finance* **9**, 99–125.
- [9] Eberlein, E., Prause, K. (2002). The generalized hyperbolic model: financial derivatives and risk measures, in *Mathematical Finance—Bachelier Congress*, Springer, Paris, pp. 245–267.
- [10] Eberlein, E. & von Hammerstein, E.A. (2004). Generalized hyperbolic and inverse Gaussian distributions: limiting cases and approximation of processes, in R.C. Dalang, M. Dozzi & F. Russo, eds, *Seminar on Stochastic Analysis, Random Fields and Applications IV*, *Progress in Probability* 58, Birkhäuser, pp. 221–264.
- [11] Jacod, J. & Shiryaev, A.N. (1987). *Limit Theorems for Stochastic Processes*, Springer.
- [12] Kou, S.G. (2002). A jump diffusion model for option pricing, *Management Science* **48**, 1086–1101.
- [13] Madan, D. & Seneta, E. (1990). The variance gamma (V.G.) model for share market returns, *Journal of Business* **63**, 511–524.
- [14] Merton, R.C. (1976). Option pricing when underlying stock returns are discontinuous, *Journal of Financial Economics* **3**, 125–144.
- [15] Sato, K.-I. (1999). *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press.

ERNST EBERLEIN

Autoregressive Moving Average (ARMA) Processes

Autoregressive moving average (ARMA) processes $\{Y_t, t = 0, \pm 1, \dots\}$ have played a key role as stationary models for time series observed at regularly spaced times. They constitute a very convenient parametric family, capable of exhibiting a broad range of dependence structures and marginal distributions. Well-established prediction, model-selection, and estimation techniques are also available and the asymptotic properties of the customary estimators are well understood. Continuous-time ARMA (denoted CARMA) processes play an analogous role in continuous time. Although continuous-time data is very rarely available, it is often the case that observations of a time series at discrete times can be fruitfully regarded as sampled values of an underlying continuous-time process. The continuous-time framework is very convenient for the modeling of high-frequency data and for dealing with irregularly spaced discrete-time data. Many of the problems of mathematical finance are formulated most conveniently in a continuous time setting, and in constructing more complex continuous-time models for financial data, CARMA processes also play a role. Although ARMA and CARMA processes with infinite second moments can be defined (see [3] and [7]) we restrict our attention here to second-order processes, for which these moments are finite.

ARMA Processes

The process $\{Y_n, n = 0, \pm 1, \dots\}$ is said to be an ARMA(p, q) process with (real-valued) parameters $\{\phi_1, \dots, \phi_p; \theta_1, \dots, \theta_q; \sigma\}$ if it is a stationary solution of the equations

$$Y_n - \phi_1 Y_{n-1} - \dots - \phi_p Y_{n-p} = \sigma(Z_n + \theta_1 Z_{n-1} + \dots + \theta_q Z_{n-q}) \quad (1)$$

where $\sigma > 0$, $\phi_p \neq 0$, $\theta_q \neq 0$, $\{Z_n\}$ is standard white noise, i.e., a sequence of uncorrelated random variables with mean 0 and variance 1 (written $\{Z_n\} \sim \text{WN}(0, 1)$) and the polynomials $\phi(z) := (1 - \phi_1 z -$

$\dots - \phi_p z^p)$ and $\theta(z) := (1 + \theta_1 z + \dots + \theta_q z^q)$ have no common zeros. Stationarity here means EY_t is independent of t and $E(Y_{t+h}Y_t)$ is independent of t for all h . It can be shown (see [5]) that there is a unique stationary solution of equation (1) for $\{Y_n\}$ in terms of the white noise sequence $\{Z_n\}$ if and only if $\phi(z)$ is nonzero for all complex z such that $|z| = 1$. The solution is

$$Y_n = \sum_{j=-\infty}^{\infty} \sigma \psi_j Z_{n-j} \quad (2)$$

where $\psi(z) := \sum_{j=-\infty}^{\infty} \psi_j z^j$ is the Laurent expansion of $\theta(z)/\phi(z)$, valid in an annulus of the form $|z - 1| < \delta$. If, in addition, the reciprocals $\xi_1, \dots, \xi_p$ of the zeros of the polynomial $\phi(z)$ satisfy

$$|\xi_r| < 1, \quad r = 1, \dots, p \quad (3)$$

then $\psi_j = 0$ for $j < 0$ and $\{Y_n\}$ is said to be a *causal function* of $\{Z_n\}$. Since, from a second-order point of view, i.e., taking into account the first- and second-order moments of $\{Y_n\}$ only, every noncausal ARMA(p, q) process can be written as a causal function of a different standard white noise process, attention is usually restricted to processes defined by equation (1) with condition (3) satisfied. We can then write

$$Y_n = \sum_{j=-\infty}^n g_{n-j} Z_j \quad (4)$$

where $\psi(z) := \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z)$, $|z| \leq 1$, and

$$g_j = \sigma \psi_j, \quad j = 0, 1, 2, \dots \quad (5)$$

The sequence $\{g_j\}$ is called the *kernel* or *absolutely summable linear filter* relating the ARMA process $\{Y_n\}$ to $\{Z_n\}$. In all that follows, we shall assume causality so that the representation (4) holds with g_j given by equation (5).

If we suppose, in addition, that the sequence $\{Z_t\}$ is an independent and identically distributed (iid) sequence, then $\{Y_n\}$ is said to be a *strict* ARMA process. In this case, $\{Y_n\}$ is a *strictly stationary* process, i.e., for each positive integer n and for each $(t_1, \dots, t_n)$, the joint distribution of $(Y_{t_1+h}, \dots, Y_{t_n+h})$ is independent of h .

The ARMA process defined by equation (1) has an equivalent and illuminating *state-space representation*, specified by the equations,

$$Y_n = \sigma \theta' \mathbf{X}_n \quad (6)$$

and

$$\mathbf{X}_{n+1} - \Phi \mathbf{X}_n = \mathbf{e} Z_{n+1} \quad (7)$$

where

$$\Phi = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \phi_r & \phi_{r-1} & \phi_{r-2} & \cdots & \phi_1 \end{bmatrix},$$

$$\mathbf{e} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_{r-1} \\ \theta_{r-2} \\ \vdots \\ \theta_1 \\ \theta_0 \end{bmatrix}$$

$r := \max(p, q + 1)$, $\theta_0 := 1$, $\theta_j := 0$ for $j > q$ and $\phi_j := 0$ for $j > p$. The causality condition (3) implies that the eigenvalues of the matrix Φ are all less than 1 in absolute value so that the state-vector $\mathbf{X}_t$ has the representation,

$$\mathbf{X}_n = \sum_{j=-\infty}^n \Phi^{n-j} \mathbf{e} Z_j \quad (8)$$

Equations (6) and (8) show that the kernel sequence $\{g_j\}$ in the representation (4) can also be written as

$$g_j = \sigma \boldsymbol{\theta}' \Phi^j \mathbf{e}, \quad j = 0, 1, 2, \dots \quad (9)$$

From equations (4) and (9) we easily find that $EY_n = 0$ and

$$\gamma_Y(h) := \text{Cov}(Y_{t+h}, Y_t) = \sigma^2 \boldsymbol{\theta}' \Phi^{|h|} \Xi \boldsymbol{\theta} \quad (10)$$

where $\Xi = E[\mathbf{X}_t \mathbf{X}_t'] = \sum_{j=0}^{\infty} \Phi^j \mathbf{e} \mathbf{e}' \Phi'^j$. The ARMA process, as we have defined it, has mean 0. We can, of course, easily extend the definition by saying that $\{U_n\}$ is an ARMA process with mean μ if $\{U_n - \mu\}$ is a zero-mean ARMA process.

Remark 1 In the case when $p > q$ and the reciprocals, $\xi_1, \dots, \xi_p$, of the zeros of the polynomial $\phi(z)$ are distinct, it follows from equations (4) and (9) and the spectral representation of the matrix Φ , that $\{Y_n\}$ has a *canonical representation* as a linear combination of (possibly complex-valued) ARMA(1,0) processes, $\{Y_{r,n}\}$. Thus

$$Y_n = \sum_{r=1}^p \beta_r Y_{r,n} \quad (11)$$

where $Y_{r,n} = \sum_{j=-\infty}^n \xi_r^{n-j} Z_j$ and $\beta_r = -\sigma \xi_r \theta(\xi_r^{-1}) / \phi'(\xi_r^{-1})$. Notice that the driving white noise sequence is the same for each of the component processes $\{Y_{r,n}\}$, so they are not independent. Corresponding to the canonical decomposition (11), there is an analogous representation of the autocovariance function, namely,

$$\gamma_Y(h) = \sum_{j=1}^p \gamma_j \xi_j^{|h|}, \quad (12)$$

where $\gamma_j = -\sigma^2 \xi_j \theta(\xi_j) \theta(\xi_j^{-1}) / [\phi(\xi_j) \phi'(\xi_j^{-1})]$.

Remark 2 Parallel to the *time domain* representation (4) of $\{Y_n\}$, there is a *spectral* or *frequency domain* representation,

$$Y_n = \int_{-\pi}^{\pi} \sigma \frac{\theta(e^{-i\omega})}{\phi(e^{-i\omega})} e^{i\omega n} dZ(\omega) \quad (13)$$

where $\{Z(\omega), -\pi \leq \omega \leq \pi\}$ is an orthogonal increment process with mean 0 and $E|dZ(\omega)|^2 = d\omega / (2\pi)$, and the autocovariance function has the corresponding spectral representation $\gamma_Y(h) = \int_{-\pi}^{\pi} f(\omega) e^{ih\omega} d\omega$, where the spectral density function f is given by

$$f(\omega) = \frac{\sigma^2}{2\pi} \left| \frac{\theta(e^{-i\omega})}{\phi(e^{-i\omega})} \right|^2, \quad \omega \in [-\pi, \pi] \quad (14)$$

Remark 3 So far we have considered only second-order properties of $\{Y_n\}$. If we make the stronger assumption that the driving white noise sequence is iid with log characteristic function, $\log E \exp(i\omega Z_n) = \xi(\omega)$, and if $t_1, \dots, t_n$ are integers such that $t_1 < t_2 < \dots < t_n$ and $\omega_1, \dots, \omega_n \in \mathbf{R}$, then we can use the representation (4) to derive the log of the joint characteristic function,

$$\begin{aligned} \log E \exp(i \sum_{r=1}^n \omega_r Y_{t_r}) \\ = \sum_{k=0}^{n-1} \sum_{j=t_{n-k-1}+1}^{t_{n-k}} \xi \left(\sum_{r=0}^k \omega_{n-r} g_{t_{n-k}-j} \right) \end{aligned} \quad (15)$$

where $t_0 := -\infty$.

CARMA Processes

A natural analog, in continuous time, of the stochastic difference equation (1) is the stochastic differential equation,

$$a(D)Y(t) = \sigma b(D)DL(t) \quad (16)$$

where σ is a strictly positive scale parameter, D denotes differentiation with respect to t , $a(z) := z^p + a_1 z^{p-1} + \dots + a_p$, $b(z) := b_0 + b_1 z + \dots + b_{p-1} z^{p-1}$, and the coefficients b_j satisfy $b_q = 1$ and $b_j = 0$ for $q < j < p$. To avoid trivial and easily eliminated complications, we shall assume that $a(z)$ and $b(z)$ have no common factors.

The continuous-time analogs of the driving noise terms Z_n in equation (1) are the increments of the process L . We shall assume that L is a Lévy process on $(-\infty, \infty)$, i.e., a process with homogeneous independent increments, continuous in probability, with cadlag sample-paths and $L(0) = 0$. We shall also restrict attention to second-order Lévy processes, i.e., those satisfying the condition $E(L(1)^2) < \infty$, and suppose, without further loss of generality, that $\text{Var}L(1) = 1$ and $EL(t) = \mu t$ for some $\mu \in \mathbb{R}$. The increments of L on disjoint intervals of equal length are then iid random variables with finite variance and some infinitely divisible distribution which could, for example, be Gaussian, gamma, compound Poisson, inverse Gaussian, or one of many other possibilities.

Since the derivatives on the right of equation (16) do not exist in the usual sense, we write the equation in state-space form (cf. equations (6) and (7)),

$$Y(t) = \sigma \mathbf{b}' \mathbf{X}(t) \quad (17)$$

and

$$d\mathbf{X}(t) - \mathbf{A}\mathbf{X}(t) dt = \mathbf{e} dL(t) \quad (18)$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -a_p & -a_{p-1} & -a_{p-2} & \dots & -a_1 \end{bmatrix},$$

$$\mathbf{e} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-2} \\ b_{p-1} \end{bmatrix}$$

with solution satisfying

$$\mathbf{X}(t) = e^{A(t-s)} \mathbf{X}(s) + \int_s^t e^{A(t-u)} \mathbf{e} dL(u), \quad \text{for all } t > s \quad (19)$$

In the Gaussian case L is Brownian motion, equation (18) is interpreted as an Ito equation and the integral in equation (19) is an Ito integral. In the general case, the integral in equation (19) is defined as in [12].

If we restrict attention to causal solutions, i.e., if we make the assumption that $\mathbf{X}(s)$ is independent of $\{L(t) - L(s), t > s\}$ for every s , then necessary and sufficient conditions for the existence of a strictly stationary process $\mathbf{X}$ satisfying equation (19) (see [6]) are

$$\text{Re}(\lambda_r) < 0, \quad r = 1, \dots, p \quad (20)$$

and, under these conditions, the stationary solution must satisfy

$$\mathbf{X}(t) \text{ is distributed as } \int_0^\infty e^{A u} \mathbf{e} dL(u) \text{ for all } t \quad (21)$$

where $\lambda_1, \dots, \lambda_p$ are the eigenvalues of A (which are also the zeros of the autoregressive polynomial $a(z)$). Condition (21) specifies the stationary marginal distribution of $\mathbf{X}(t)$ and condition (20) is the continuous-time analog of the causality condition (3).

If we assume that the conditions (20) and (21) hold, and let $s \rightarrow -\infty$ in equation (19) we find that

$$\mathbf{X}(t) = \int_{-\infty}^t e^{A(t-u)} \mathbf{e} dL(u) \quad (22)$$

Conversely, if $\mathbf{X}(t)$ is defined by equation (22) then $\mathbf{X}$ is a strictly stationary causal process satisfying equation (19). Consequently, $\{\mathbf{X}(t)\}$ as defined in equation (22) is the unique, strictly stationary causal function of L satisfying equation (19).

We now define the strictly stationary causal CARMA process by equation (17) with $\mathbf{X}$ given by equation (22). Thus,

$$Y(t) = \int_{-\infty}^t \sigma \mathbf{b}' e^{A(t-u)} \mathbf{e} dL(u) \quad (23)$$

This is the continuous-time analog of equation (4) with the discrete-time kernel, $\{g_j = \sigma \theta' \Phi^j \mathbf{e}, j = 0, 1, 2, \dots\}$, replaced by the continuous-time kernel,

$$g(t) = \sigma \mathbf{b}' e^{At} \mathbf{e}, \quad t > 0 \quad (24)$$

From equation (23), we find that $EY(t) = \sigma \mu b_0/a_p$ (where $\mu = EL(1)$) and

$$\gamma_Y(h) := \text{Cov}[Y(t+h), Y(t)] = \sigma^2 \mathbf{b}' e^{A|h|} \Sigma \mathbf{b} \quad (25)$$

where $\Sigma = E[\mathbf{X}(t)\mathbf{X}(t)'] = \int_0^\infty e^{Ay} \mathbf{e} \mathbf{e}' e^{A'y} dy$.

Remark 4 If the zeros, $\lambda_1, \dots, \lambda_p$, of the polynomial $a(z)$ are distinct, it follows from equation (23) and the spectral representation of the matrix A that $\{Y_n\}$ has a *canonical representation* as a linear combination of (possibly complex-valued) CARMA(1,0) processes, $\{Y_r(t)\}$. Thus,

$$Y(t) = \sum_{r=1}^p \beta_r Y_r(t) \quad (26)$$

where $Y_r(t) = \int_{-\infty}^t e^{\lambda_r(t-u)} dL(u)$ and $\beta_r = \sigma b(\lambda_r)/a'(\lambda_r)$. Notice that the driving process L is the same for each of the component processes $\{Y_r(t)\}$, so they are not independent. Corresponding to the canonical decomposition (26), there is an analogous representation of the autocovariance function, namely,

$$\gamma(h) = \sum_{r=1}^p \gamma_r e^{\lambda_r |h|} \quad (27)$$

where $\gamma_r = \sigma^2 b(\lambda_r) b(-\lambda_r) / [a(-\lambda_r) a'(\lambda_r)]$.

Remark 5 The mean-corrected process $\{Y^*(t) = Y(t) - \sigma \mu b_0/a_p\}$ has a spectral representation

$$Y^*(t) = \int_{-\infty}^{\infty} \sigma \frac{b(i\omega)}{a(i\omega)} e^{i\omega t} dZ(\omega) \quad (28)$$

where $\{Z(\omega), -\infty < \omega < \infty\}$ is an orthogonal increment process with mean 0 and $E|dZ(\omega)|^2 = d\omega/(2\pi)$, and the autocovariance function of Y (and of Y^*) has the corresponding spectral representation $\gamma_Y(h) = \int_{-\infty}^{\infty} f(\omega) e^{i\omega h} d\omega$, where the spectral density function f is given by

$$f(\omega) = \frac{\sigma^2}{2\pi} \left| \frac{b(i\omega)}{a(i\omega)} \right|^2, \quad \omega \in (-\infty, \infty) \quad (29)$$

Gaussian processes with such rational spectral densities were discussed extensively in [9].

Remark 6 If $\log E \exp(i\omega L(1)) = \xi(\omega)$, and if $t_1, \dots, t_n$ satisfy $t_1 < t_2 < \dots < t_n$ and $\omega_1, \dots, \omega_n \in \mathbf{R}$, then from the representation (23),

$$\begin{aligned} \log E \exp(i \sum_{r=1}^n \omega_r Y(t_r)) \\ = \sum_{k=0}^{n-1} \int_{t_{n-k-1}}^{t_{n-k}} \xi \left(\sum_{r=0}^k \omega_{n-r} g(t_{n-r} - u) \right) du \end{aligned} \quad (30)$$

where $t_0 := -\infty$.

Prediction and Inference

From observations at discrete times, $t_1 < t_2 < \dots < t_{k-1}$, of either an ARMA or CARMA process $\{Y(t)\}$ with specified parameter values, various techniques are available for computing the minimum mean-squared error linear predictor,

$$\hat{Y}(t_k) = \alpha_{k,0} + \alpha_{k,1} Y(t_{k-1}) + \dots + \alpha_{k,k-1} Y(t_1) \quad (31)$$

of $Y(t_k)$, $t_k > t_{k-1}$, and its mean-squared error, $v_k = E(Y(t_k) - \hat{Y}(t_k))^2$. For details see, for example [5] and [11]. If we assume that the series is Gaussian, we can write the joint density of $(Y(t_1), \dots, Y(t_n))$ at $(y(t_1), \dots, y(t_n))$ as

$$g(y(t_1), \dots, y(t_n)) = \prod_{k=1}^n \frac{1}{\sqrt{v_k}} f \left(\frac{y(t_k) - m(t_k)}{\sqrt{v_k}} \right) \quad (32)$$

where f is the standard normal density function, $m(t_k) = \alpha_{k,0} + \alpha_{k,1} y(t_{k-1}) + \dots + \alpha_{k,k-1} y(t_1)$ for $k > 1$, $m_1 = EY(t)$ and $v_1 = \text{Var}(Y(t))$. For given values of p and q , maximum Gaussian likelihood (MGL) estimators of the parameters are obtained by maximizing g with respect to the process parameters. This is a nonlinear maximization problem, and requires the use of efficient numerical techniques. Many other estimation procedures are available, with varying asymptotic efficiencies, but it is known that under rather mild conditions and even when the process $\{Y(t)\}$ is not in fact Gaussian, MGL estimators

have good asymptotic properties. Model selection, i.e., choice of p and q , is usually made by minimization of an information criterion, well-known examples being the AIC, AICC, and BIC. See [5] and [10] for details.

Some Extensions and Applications

One of the features frequently observed in financial and geophysical time series is the very slow rate of decay with increasing lag of the sample autocorrelation function. In order to model this phenomenon, *fractionally integrated* ARMA and CARMA processes were introduced. These have the property that the autocorrelation function decreases asymptotically at a hyperbolic rate, rather than at the exponential rate of decay for both ARMA and CARMA processes. For information on fractionally integrated processes see [2, 6, 8].

Nonlinear generalizations of ARMA and CARMA processes, in particular threshold models (see [14]), have also found a wide variety of applications.

In mathematical finance, the CARMA(1,0) (or stationary Ornstein–Uhlenbeck) process, driven by a nondecreasing Lévy process, has been used to represent stochastic volatility in the stochastic volatility model of Barndorff-Nielsen and Shephard [1]. Higher order CARMA models for stochastic volatility have also been used to allow for more complex autocorrelation functions (see [4] and [13]). In the COGARCH models of (see Further Reading), the volatility process is a nonlinear self-exciting CARMA process and for COGARCH models in which the volatility process has finite variance, the volatility has the autocorrelation function of a CARMA process, a property analogous to the corresponding result for the volatility of a GARCH process, which, if the volatility has finite variance, has the autocorrelation function of an ARMA process.

Acknowledgments

The author gratefully acknowledges support of this work by the National Science Foundation under Grant, DMS-0744058.

References

- [1] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck based models and some of their uses in financial economics (with discussion), *Journal of the Royal Statistical Society, Series B* **63**, 167–241.
- [2] Beran, J. (1994). *Statistics for Long-Memory Processes*, Chapman & Hall, New York.
- [3] Brockwell, P.J. (2001). Lévy-driven CARMA processes, *Annals of the Institute of Statistical Mathematics* **53**, 113–124.
- [4] Brockwell, P.J. (2004). Representations of continuous-time ARMA processes, *Journal of Applied Probability* **41A**, 375–382.
- [5] Brockwell, P.J. & Davis, R.A. (1991). *Time Series: Theory and Methods*, 2nd Edition, Springer, New York.
- [6] Brockwell, P.J. & Marquardt, T. (2005). Fractionally integrated continuous-time ARMA processes, *Statistica Sinica* **15**, 477–494.
- [7] Cline, D.B.H. & Brockwell, P.J. (1985). Linear prediction of ARMA processes with infinite variance, *Stochastic Processes and their Applications* **19**, 281–296.
- [8] Comte, F. & Renault, E. (1996). Long memory continuous time models, *Journal of Econometrics* **73**, 101–149.
- [9] Doob, J.L. (1944). The elementary Gaussian processes, *Annals of the Mathematics and Statistics* **25**, 229–282.
- [10] Hannan, E.J. (1973). The asymptotic theory of linear time series models, *Journal of Applied Probability* **10**, 130–145.
- [11] Jones, R.H. (1985). Time series analysis with unequally spaced data, in *Time Series in the Time Domain, Handbook of Statistics*, E.J. Hannan, P.R. Krishnaiah & M.M. Rao, eds, North Holland, Amsterdam, Vol. 5, pp. 157–178.
- [12] Protter, P.E. (2004). *Stochastic Integration and Differential Equations*, 2nd Edition, Springer, New York.
- [13] Todorov, V. & Tauchen, G. (2006). Simulation methods for Lévy-driven CARMA stochastic volatility models, *Journal of Business and Economic Statistics* **24**, 455–469.
- [14] Tong, H. (1990). *Non-linear Time Series : A Dynamical System Approach*, Clarendon Press, Oxford.

Further Reading

- Brockwell, P.J., Chandraa, E. & Lindner, A. (2006). Continuous-time GARCH processes, *Annals of Applied Probability* **16**, 790–826.
- Klüppelberg, C., Lindner, A. & Maller, R. (2004). A continuous time GARCH process driven by a Lévy process: stationarity and second order behaviour, *Journal of Applied Probability* **41**, 601–622.

PETER J. BROCKWELL

Copulas in Econometrics

Copulas and Multivariate Models

The knowledge of the multivariate (conditional) distribution (especially fat tails, asymmetry, positive or negative dependence) is essential in many important financial applications, including portfolio selection, option pricing, asset pricing models, Value-at-Risk (market risk, credit risk, liquidity risk) calculations, and forecasting. Copulas provide a general, flexible tool in modeling the complex dependence structure among multiple random variables and in building multivariate models from univariate models; see [6, 20, 26] for discussions and references on the applications of copulas in economics, finance, insurance, and risk management. Patton [23] provides an excellent review of copula models for financial time series.

A copula is a multivariate distribution function with uniform marginal distributions on $[0, 1]$ that links together univariate distribution functions to form a multivariate distribution function. To simplify the exposition, the focus is on the bivariate case in this article. Let X, Y denote two scalar random variables with the joint distribution function H and marginal distribution functions F, G . The characterization theorem of Sklar [25] implies that there exists a copula $C(u, v)$: $(u, v) \in [0, 1]^2$ such that

$$H(x, y) = C(F(x), G(y)) \text{ for all } (x, y) \in \mathcal{R}^2 \quad (1)$$

Conversely, for any marginal distributions F, G and any copula function C , the function $C(F(\cdot), G(\cdot))$ is a bivariate distribution function with given marginal distributions F, G . This theorem provides the theoretical foundation for the widespread use of copulas in generating multivariate distributions (models) from univariate distributions (models).^a The fact that the copula function C , the marginal distribution functions F, G in equation (1) are not necessarily of the same type allows the researcher a great deal of flexibility in specifying a multivariate distribution (model). For example, Chen and Fan [2] propose a general class of multivariate dynamic time series models by first modeling each individual time series by a univariate generalized autoregressive conditional heteroskedasticity (GARCH)-type model and then modeling the joint distribution of the innovations of the univariate models by a distribution with

a parametric copula and unknown marginal distributions. We refer the reader to [2] for details and related references.

Suppose that X, Y are continuous random variables. Then the copula function C in equation (1) is unique.^b It extracts the dependence information on X, Y from their joint distribution H and is sometimes referred to as the *dependence function of X, Y* . The copula function of X, Y is invariant to monotone increasing transformations of X, Y and most rank dependence measures can be expressed in terms of the copula alone, including Kendall's τ , Spearman's ρ , and tail dependence coefficients; for details, see [17] and [22]. It is to be noted, however, that Pearson's linear correlation coefficient cannot be expressed in terms of the copula alone; it also depends on the marginal distributions.

Suppose that the density function of (X, Y) exists and is denoted by h . Then equation (1) implies

$$h(x, y) = f(x)g(y)c(F(x), G(y)) \quad \forall (x, y) \in \mathcal{R}^2 \quad (2)$$

where f, g are the marginal density functions and c is the copula density function. The decomposition of the joint density function h in equation (2) is the basis for the commonly used methods for estimating copula models. For a fully parametric copula model in which both the copula function and the marginal distribution functions are parameterized, both maximum likelihood (ML) method and two-step method can be used. Joe [17] and Nelsen [22] present numerous parametric copula functions including the Gaussian copula, Student's t copula, and Archimedean copulas. For a semiparametric copula model in which the copula function is parameterized, but the marginal distribution functions are not, both two-step method and sieve ML method have been developed; see [15] for details on the two-step method and [5] for details on the sieve ML method. To introduce the two-step approach and ML or sieve ML approach, suppose that we have a random sample $\{X_i, Y_i\}_{i=1}^n$ from $h(x, y)$. Then equation (2) implies that the log-likelihood function is given as

$$\begin{aligned} \sum_{i=1}^n \ln f(X_i) + \sum_{i=1}^n \ln g(Y_i) \\ + \sum_{i=1}^n \ln c(F(X_i), G(Y_i)) \end{aligned} \quad (3)$$

In the two-step approach, the marginal distribution functions F and G are estimated in the first step and the copula function is estimated in the second step. For example, for a semiparametric copula model, the marginal distribution functions are estimated in the first step by the corresponding (rescaled) empirical distribution functions denoted as $\hat{F}(x)$, $\hat{G}(y)$ and the copula function is estimated in the second step by maximizing $\sum_{i=1}^n \ln c(\hat{F}(X_i), \hat{G}(Y_i))$. We refer the reader to [15] for details on the two-step approach for random samples and [2, 3] for extensions to copula-based multivariate and univariate time series models, respectively. In the ML or sieve ML approach, the marginal distribution functions and the copula function are estimated simultaneously. For a fully parametric copula model, this is accomplished by maximum likelihood estimation (MLE) obtained from maximizing the log-likelihood function (3) with respect to parameters in both the marginal distribution functions and the copula function simultaneously, whereas for a semiparametric copula model, the maximization of the log-likelihood function (3) is carried out over respective sieve spaces approximating the spaces of marginal distribution functions F and G and the parameter space for the copula parameter; see [5] for details on sieve spaces and the sieve MLE. In general, the two-step estimator is asymptotically less efficient than the ML estimator in a parametric copula model or the sieve ML estimator in a semiparametric copula model.

In a parametric or semiparametric copula model in which the copula is parameterized, it is important to select an appropriate parametric copula and to evaluate how well the selected copula model fits the data at hand. Commonly used parametric copula families include the Gaussian copula family, Student's t copula family, Gumbel copula family, Clayton copula family, and Frank copula family. The latter three copula families are members of the general Archimedean copula family. These copula families have different dependence properties and are thus appropriate for modeling different dependence properties of economic and financial variables. For example, the Gaussian and Frank copula families have no tail dependence and are thus not appropriate for modeling financial variables that tend to take extreme values together. Instead, Student's t copula family has both lower and upper tail dependence and thus is appropriate for modeling economic or financial variables taking extremely large or small values together. On the

other hand, the Gumbel copula family has upper tail dependence and no lower tail dependence, whereas the Clayton copula family has lower tail dependence and no upper tail dependence. Chen and Fan [2, 4] developed model selection tests for copula families and [13], among others, developed goodness-of-fit tests for copulas. For applications of estimation and model selection procedures for copula-based models, the reader is referred to [4, 7]. Additional references can be found in [23].

The General Fréchet Problem

In addition to its role in building multivariate models from univariate models, copulas have also proven to be useful in finding solutions to the general Fréchet problem. Broadly speaking, the Fréchet problem concerns the problem of finding sharp bounds on the distributions of functions of random variables X, Y such as $X + Y$ and $X - Y$ when the marginal distribution functions F, G are fixed. The reader is referred to [24, 27] for discussions and references on the general Fréchet problem and its solutions, to [20] for its applications in finance and risk management, and to [11, 12] for its applications in program evaluation.

For $(u, v) \in [0, 1]^2$, let $C^L(u, v) = \max(u + v - 1, 0)$ and $C^U(u, v) = \min(u, v)$ denote the Fréchet–Hoeffding lower and upper bounds for a copula, that is, $C^L(u, v) \leq C(u, v) \leq C^U(u, v)$. Then for any $(x, y) \in \mathcal{R}^2$, the following inequality holds:

$$C^L(F(x), G(y)) \leq H(x, y) \leq C^U(F(x), G(y)) \quad (4)$$

The bivariate distribution functions $C^L(F(\cdot), G(\cdot))$ and $C^U(F(\cdot), G(\cdot))$ are referred to as the *Fréchet–Hoeffding lower and upper bounds* for bivariate distribution functions with fixed marginal distributions F and G . They are distributions of perfectly negatively dependent and perfectly positively dependent random variables respectively; see [22] for more discussions. The inequality (4) directly implies sharp bounds on the correlation coefficient between X and Y : $\rho_L \leq \rho \leq \rho_U$, where

$$\rho_L = \frac{\int \int [C^L(F(x), G(y)) - F(x)G(y)] dx dy}{\sigma_X \sigma_Y}$$

$$\rho_U = \frac{\int \int [C^U(F(x), G(y)) - F(x)G(y)] dx dy}{\sigma_X \sigma_Y} \quad (5)$$

in which σ_X^2 and σ_Y^2 are variances of X and Y , respectively. It is known that ρ_L and ρ_U are sharp; see [9, 20] for detailed discussions on this result and other properties and pitfalls of the correlation coefficient as a dependence measure.

Although inequality (4) does not directly imply bounds on distribution functions of $X + Y$ and $X - Y$, copulas provide a useful tool in finding such bounds; see [14]. In what follows, sharp bounds on the distribution function of $X - Y$ and sharp bounds on the quantile function of $X + Y$ are provided and their applications pointed out.

Let $\Delta = X - Y$. Δ is referred to as the *individual treatment effect* in the literature on program evaluation. In this context, X is the potential outcome from treatment and Y is the potential outcome without treatment; see [11, 12] for details. Denote the distribution function of Δ by $F_\Delta(\cdot)$. Given the marginals F and G , sharp bounds on the distribution of Δ can be found in [27]: $F_\Delta^L(\delta) \leq F_\Delta(\delta) \leq F_\Delta^U(\delta)$, where

$$\begin{aligned} F_\Delta^L(\delta) &= \sup_y \max(F(y) - G(y - \delta), 0), \\ F_\Delta^U(\delta) &= 1 + \inf_y \min(F(y) - G(y - \delta), 0) \end{aligned} \quad (6)$$

Let $\Lambda = X + Y$ and $Q_\Lambda(p)$ ($= F_\Lambda^{-1}(p)$) denote its generalized quantile function at quantile level p : $0 < p < 1$. Given the generalized quantile functions F^{-1} and G^{-1} , sharp bounds on $Q_\Lambda(p)$ can be found in [20, 27]: $Q_\Lambda^L(p) \leq Q_\Lambda(p) \leq Q_\Lambda^U(p)$, where

$$\begin{aligned} Q_\Lambda^L(p) &= \inf_{u \in [p, 1]} [F^{-1}(u) + G^{-1}(u - p + 1)], \\ Q_\Lambda^U(p) &= \sup_{u \in [0, p]} [F^{-1}(u) + G^{-1}(u - p)] \end{aligned} \quad (7)$$

McNeil *et al.* [20] discuss applications of equation (7) in risk management and provide an extensive list of references.

The expressions in equations (5)–(7) share one common feature, that is, they depend only on the marginal distributions. As a result, they can be consistently estimated without requiring any dependence

information between X and Y , provided that univariate samples from F, G are available. For example, consider a linear portfolio Λ of market risk X and credit risk Y . Estimation of the Value-at-Risk of Λ requires a bivariate sample from the joint distribution of X, Y , which may not always be available. In contrast, equation (7) implies that the smallest and largest Value-at-Risk can be estimated with univariate samples from F, G . Inference on ρ , $F_\Delta(\delta)$, and $Q_\Lambda(p)$ belongs to the recent, but fast-growing area in econometrics, namely, inference for partially identified parameters. The reader is referred to [10] for discussion and references.

Conclusion

This article concludes with a mention of some recent work on the construction of higher dimensional copulas and the Fréchet problem in higher dimensions. Copulas have found to be most successful in bivariate modeling, as there are numerous parametric families of bivariate copulas that the researcher can choose from; see [17, 22]. Compared with bivariate parametric copulas, the number of higher dimensional parametric copulas is rather limited. So far, most applications in higher dimensions have focused on Gaussian copulas and Student's t copulas. It is well known that the construction of higher dimensional copulas is very difficult. Several methods have been proposed recently, including methods for constructing higher dimensional Archimedean copulas and the general pair-copula construction approach known as *vines*; see [1, 19] and the references therein.

The Fréchet problem in higher dimensions has drawn attention recently. The reader is referred to [8] for a brief account and references. Finally, the reader is referred to [21] and the associated discussions and rejoinder for a lively discussion on the value of copulas in multivariate modeling.

End Notes

^aJoe [17] and Nelsen [22] provide a detailed introduction to copulas and their mathematical and statistical foundations. Kallsen and Tankov [18] present a version of Sklar's theorem that allows to construct a general multivariate Lévy process from arbitrary univariate Lévy processes and an arbitrary Lévy copula.

^bGenest and Nešlehová [16] present an excellent discussion on copulas for discrete data.

References

- [1] Aas, K., Czado, C. Frigessi, A. & Bakken, H. (2009). Pair-copula constructions of multiple dependence, *Insurance, Mathematics, and Economics* **44**(2), 182–198.
- [2] Chen, X. & Fan, Y. (2006). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification, *Journal of Econometrics* **135**, 125–154.
- [3] Chen, X. & Fan, Y. (2006). Semiparametric estimation of copula-based time series models, *Journal of Econometrics* **130**, 307–335.
- [4] Chen, X. & Fan, Y. (2007). Model selection test for bivariate failure-time data, *Econometric Theory* **23**, 414–439.
- [5] Chen, X., Fan, Y. & Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models, *Journal of the American Statistical Association* **101**, 1228–1240.
- [6] Cherubini, U., Luciano, E. & Vecchiato, W. (2004). *Copula Methods in Finance*, John Wiley & Sons, England.
- [7] Dias, A. & Embrechts, P. (2003). *Dynamic Copula Models for Multivariate High-Frequency Data in Finance*, Mimeo.
- [8] Embrechts, P. (2008). Copulas: a personal view, Forthcoming in *Journal of Risk and Insurance*.
- [9] Embrechts, P., McNeil, A. & Straumann, D. (2002). Correlation and dependence properties in risk management: properties and pitfalls, in *Risk Management: Value at Risk and Beyond*, M. Dempster, ed., Cambridge University Press, pp. 176–223.
- [10] Fan, Y. & Park, S. (2007). *Confidence Sets for Some Partially Identified Parameters*, Manuscript, Vanderbilt University.
- [11] Fan, Y. & Park, S. (2009). Sharp bounds on the distribution of treatment effects and their statistical inference, Forthcoming in *Econometric Theory*.
- [12] Fan, Y. & Wu, J. (2007). *Partial Identification of the Distribution of Treatment Effects in Switching Regimes Models and its Confidence Sets*, Manuscript, Vanderbilt University.
- [13] Fermanian, J.-D. (2005). Goodness of fit tests for copulas, *Journal of Multivariate Analysis* **95**, 119–152.
- [14] Frank, M.J., Nelsen, R.B. & Schweizer, B. (1987). Best-possible bounds on the distribution of a sum – a problem of Kolmogorov, *Probability Theory and Related Fields* **74**, 199–211.
- [15] Genest, C., Ghoudi, K. & Rivest, L. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions, *Biometrika* **82**(3), 543–552.
- [16] Genest, C. & Nešlehova, J. (2007). A primer on copulas for count data, Forthcoming in *The ASTIN Bulletin* **37**, 475–515.
- [17] Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall/CRC.
- [18] Kallsen, J. & Tankov, P. (2006). Characterization of dependence of multidimensional Lévy processes using Lévy Copulas, *Journal of Multivariate Analysis* **97**, 1551–1572.
- [19] Kurowicka, D. & Cooke, R. (2006). *Uncertain Analysis with High Dimensional Dependence Modelling*, Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd.
- [20] McNeil, A.J., Frey, R. & Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton University Press, New Jersey.
- [21] Mikosch, T. (2006). Copulas: tales and facts, with discussion and rejoinder, *Extremes* **9**, 3–62.
- [22] Nelsen, R.B. (1999). *An Introduction to Copulas*, Springer.
- [23] Patton, A.J. (2007). Copula-based models for financial time series, in *Handbook of Financial Time Series*, T.G. Anderson, R.A. Davis, J.-P. Kreiss & T. Mikosch, eds, Springer-Verlag, pp. 767–786.
- [24] Schweizer, B. & Sklar, A. (1983). *Probabilistic Metric Spaces*, North-Holland, New York.
- [25] Sklar, A. (1959). Fonctions de répartition à n dimensionset leurs marges, *Publications de l'Institut de Statistique de l'Université de Paris* **8**, 229–231.
- [26] Trivedi, P. & Zimmer, D.M. (2007). Copula modeling: an introduction for practitioners, *Foundations and Trends in Econometrics* **1**(1), 1–110.
- [27] Williamson, R.C. & Downs, T. (1990). Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds, *International Journal of Approximate Reasoning* **4**, 89–158.

Related Articles

Copulas in Insurance; Copulas: Estimation; Correlation Risk; Lévy Copulas; Multivariate Distributions.

YANQIN FAN

Measurements Errors

In physical sciences, errors in measurement are present everywhere owing to the unavoidable inaccuracy of the measuring devices, and most often they are thought of as being independent and centered variables, usually with a density, and which add up to the “true” values of the quantity of interest. In finance, the quantities of interest are usually prices, which can be exactly observed by virtually any person. Some factual errors, like wrong records, may occur, but those are relatively infrequent and easy to detect (giving rise to the “cleaning” of data), and may be modeled with independent centered variables with no density and a rather large mass at 0 (= no error).

However, mathematical models in finance are often continuous-time models describing the dynamics of a process $(X_t : t \geq 0)$. This process is the basic underlying “price”, or log-price, of an asset, or several assets if it is multivariate, and it satisfies the rules of mathematical finance, and, in particular, it is a semimartingale by the first fundamental theorem of asset pricing. This process cannot be exactly observed because (i) any observed price is an integer (a number of elementary units, like a cent or a hundredth of a cent); and (ii) one may think that extraneous reasons, like some erratic behavior of price makers or erratic orders, “locally” change the price in a way that does not affect the underlying and should not influence the model for it.

In practice, one then observes prices Z_t that are possibly different from the underlying, or “efficient”, prices X_t , and at finitely many times $t_0, t_1, \dots, t_n$. This may look like an academic difference and it may appear that, in practice, we ought to consider Z_t and not X_t at all. However, currently we have high-frequency or very high-frequency data, like tick data, and it is empirically clear that we need to differentiate between X_t and Z_t . For example, when $Z_t = X_t$ (no error), the so-called realized volatility, which is the sum

$$V_n = (Z_{t_1} - Z_{t_0})^2 + (Z_{t_2} - Z_{t_1})^2 + \dots + (Z_{t_n} - Z_{t_{n-1}})^2 \quad (1)$$

is a quantity that converges, when the observations become more and more frequent over a fixed time interval $[0, T]$, to a given quantity V (the integrated

volatility over $[0, T]$ when X_t is continuous, and, more generally, to the quadratic variation of the process X_t). Moreover, at least when observation times are equidistant or almost equidistant, the “rate of convergence” is $\sqrt{n}$, which means that $\sqrt{n} (V_n - V)$ does neither explode nor go to 0 (and in fact it usually converges to some limit). Now, all empirical studies show that V_n increases indefinitely when the frequency of observations increases: this goes back to [15] and [8], or more recently to, for example, [2] or [12]. Hence, the observations Z_t differ from the underlying X_t in a nonnegligible way, and the difference $\epsilon_t = Z_t - X_t$ is referred to as the *microstructure noise*, or *market frictions*, and should be taken into account.

We then have to give a model for this microstructure noise, on top of the semimartingale model for X_t , like Black–Scholes, possibly with stochastic volatility, or jump diffusion, or others. The microstructure noise models that have been proposed so far in the literature are mostly the following:

1. The errors ϵ_t are independent, centered, identically distributed variables, and also independent of the underlying X_t . This is the simplest and most well-understood situation, in which the uncorrected realized volatility V_n increases to ∞ at the rate n . When the efficient price X_t is continuous and one is interested in estimating the integrated volatility V over $[0, T]$, that is, the quadratic variation of X_t at time T , a variety of variants of V_n have been proposed, using various smoothing kernels: the best methods give a rate of convergence which is $n^{1/4}$, known as the *optimal rate* when X_t is a Brownian motion, and asymptotic variances which are nearly optimal. Note that no distributional assumption is made on the errors, and, in particular, the variance of the errors is not supposed to be known.

Such results can be extended to (i) errors that are not independent, but present some sort of “weak” dependence, and are still globally independent of X_t ; (ii) errors that are not identically distributed, provided their second moments stay constant; (iii) errors that are multiplicative instead of additive, meaning that the variables Z_t/X_t are independent instead of the $Z_t - X_t$, and “centered” is replaced by “with mean 1”. The literature about this question is huge, and among many articles we can quote [1, 3–6, 13, 14, 16, 17].

2. The previous models more or less take care of reason (ii) for microstructure noise. As for reason (i), and if it were the only reason for the noise, one should simply take the observation Z_t to be $Z_t = \alpha[X_t/\alpha]$, where $[x]$ denotes the integer part of any number x , and α is the precision with which prices are written (1 cent or 1/100 cent for example). Then again some results for the estimation of the integrated volatility V are available, asymptotically when n increases whereas the rounding level α decreases to 0. Very few papers have been devoted so far to this subject; see, for example, [7].
3. Now, with the frequency attained by observations of prices, one cannot satisfactorily consider that the rounding level α is “small”, and we have a serious problem here: if α is fixed and the frequency increases, then the realized volatility increases to ∞ with the rate $\sqrt{n}$, and the integrated volatility is no longer identifiable. The best one can know about the process X_t is the times at which it crosses the level $i\alpha$ for any integer i , or equivalently the family of so-called “local times” $L(i\alpha)$ at levels $i\alpha$ for all integers i (and time T , the time horizon). This is very far from V , although V is indeed the limit of $\alpha \sum_{i \geq 0} L(i\alpha)$ as $\alpha \rightarrow 0$ (see [9]).
4. In practice, models like (1) above are not fully satisfactory because they do not account for the—obvious and ubiquitous—fact that rounding-off occurs. Models like (2) with a fixed level α are not satisfactory either, on a mathematical level (they do not permit to retrieve the quantities of interest like V) and on an empirical level they give rise to sequences of prices that stay constant for relatively many successive times and then oscillate wildly between two adjacent multiples of α , then stay constant again: this is not quite observed in practice; usually prices stay constant for a few successive times, with occasionally a jump back and forth of one unit. Probably a more reasonable model consists in a mixture of (1) and (2): we have additive errors ϵ_t , and then rounding, that is, we observe $Z_t = \alpha[(X_t + \epsilon_t)/\alpha]$. It turns out—perhaps surprisingly—that if the errors ϵ_t have a density that is positive on an arbitrarily located interval of length at least α , then it is possible to retrieve the integrated volatility V in pretty much the same way as when rounding

is absent, and at the same rate. Some recent attempts in this direction may be found in [11] and [10].

5. One might also look for more microeconomically oriented models. That is, in addition to the rounding effect, which is probably unavoidable, one tries to model the behavior of the various agents so that the observed prices Z_t behave according to empirical data, whereas the underlying X_t , which is a sort of “local limit” of the Z_t ’s, satisfies a model that fits the usual requirements of mathematical finance (no-arbitrage, completeness, and so on). This looks like the two-scale diffusions, but so far this has not really been put in use, because of the mathematical difficulties, and even more because the underlying microeconomics is still at a rather rudimentary level.

References

- [1] Ait-Sahalia, Y., Mykland, P.A. & Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise, *Review of Financial Studies* **18**, 351–416.
- [2] Andersen, T.G., Bollerslev, T., Diebold, F.X. & Labys, P. (2000). Great realizations, *Risk* **13**, 105–108.
- [3] Andersen, T.G., Bollerslev, T., Diebold, F.X. & Labys, P. (2001). The distribution of realized exchange rate volatility, *Journal of the American Statistical Association* **96**, 42–55.
- [4] Bandi, F.M. & Russell, J.R. (2006b). Separating microstructure noise from volatility, *Journal of Financial Economics* **79**, 655–692.
- [5] Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. & Shephard, N. (2008). Designing realised kernels to measure ex-post variation of equity prices in the presence of noise, *Econometrica* **76**(6), 1481–1536.
- [6] Barndorff-Nielsen, O.E. & Shephard, N. (2002). Econometric analysis of realised volatility and its use in estimating stochastic volatility models, *Journal of the Royal Statistical Society, B* **64**, 253–280.
- [7] Delattre, S. & Jacod, J. (1997). A central limit theorem for normalized functions of the increments of a diffusion process, in the presence of round-off errors, *Bernoulli* **3**, 1–28.
- [8] Hasbrouck, J. (1993). Assessing the quality of a security market: a new approach to transaction-cost measurement, *Review of Financial Studies* **6**, 191–212.
- [9] Jacod, J. (1996). La variation quadratique du Brownien en présence d’erreurs d’arrondi, *Astérisque* **236**, 155–162.
- [10] Jacod, J., Li, Y., Mykland, P.A., Podolskij, M. & Vetter, M. (2009). Microstructure noise in the continuous

-
- case: the pre-averaging approach, *Stochastic Processes and their Applications* **119**(7), 2249–2276.
- [11] Li, Y. & Mykland, P.A. (2007). Are volatility estimators robust with respect to modeling assumptions? *Bernoulli* **13**, 601–622.
 - [12] Mykland, P.A. & Zhang, L. (2005). Discussion of paper “a selective overview of nonparametric methods in financial econometrics” by J. Fan, *Statistical Science* **20**, 347–350.
 - [13] Oomen, R.A.A. (2005). Properties of bias corrected realized variance in calendar time and business time, *Journal of Financial Econometrics* **3**, 258–272.
 - [14] Podolskij, M. & Vetter, M. (2006). Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps, Technical Report, Ruhr-Universität, Bochum. To appear in *Bernoulli*.
 - [15] Roll, R. (1984). A simple model of the implicit bid-ask spread in an efficient market, *Journal of Finance* **39**, 1127–1139.
 - [16] Zhang, L. (2006). Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach, *Bernoulli* **12**, 1019–1043.
 - [17] Zhang, L., Mykland, P.A. & Aït-Sahalia, Y. (2005). A tale of two time scales: determining integrated volatility with noisy high-frequency data, *Journal of the American Statistical Association* **100**, 1394–1411.

JEAN JACOD

Time Change

The mathematical concept of time-changing continuous-time stochastic processes has first been studied in [11, 17, 18, 23, 24, 29, 39] and has later been introduced to the finance literature in [14]. Today, it can be regarded as one of the standard tools for building financial models (see also [22, 32]).

Let $X = (X_t)_{t \geq 0}$ denote a stochastic process, sometimes referred to as the *base process*, and let $T = (T_s)_{s \geq 0}$ denote a nonnegative, nondecreasing stochastic process not necessarily independent of X . The *time-changed process* is then defined as $Y = (Y_s)_{s \geq 0}$, where

$$Y_s = X_{T_s} \quad (1)$$

The process X is said to evolve in *operational time*. The process T is referred to as a *time change*, *stochastic clock*, *chronometer*, or *business time*. It reflects the varying *speed* of Y .

This article is structured as follows. First we link the use of time-changed stochastic processes to the construction of univariate stochastic volatility models in finance. Then we focus on the choice of an appropriate base process and, next, on popular examples for the time change. Finally, we briefly discuss the potential and the limitations of the methodology of time-changing stochastic processes to construct multivariate models in finance.

Stochastic Volatility Models

The use of time-changed stochastic processes in finance is closely linked to the concept of stochastic volatility models for asset prices. Numerous empirical studies have revealed the fact that asset price volatility tends to be time varying and tends to show clustering effects. The concept of stochastic volatility (see **Volatility**) in continuous-time asset price models on a filtered probability space $(\Omega, \mathcal{A}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ can basically be introduced by two methods.

One of the methods is to use time-changed stochastic processes as in equation (1), where natural assumptions are that $X = (X_t)_{t \geq 0}$ is an $\mathcal{F}_t$ -semimartingale and $(T_s)_{s \geq 0}$ is an increasing family of $\mathcal{F}_t$ -stopping times. The base process X is often assumed to possess some homogeneity properties,

whereas a nonlinear time change can induce deviations from homogeneity.

The other method is to use stochastic integrals (see **Stochastic Integrals**) of the form

$$Y_t = \int_0^t \sigma_{r-} dX_r \quad (2)$$

where $\sigma = (\sigma_t)_{t \geq 0}$ is a nonnegative $\mathcal{F}_t$ -predictable stochastic volatility process and $X = (X_t, t \geq 0)$ is an $\mathcal{F}_t$ -semimartingale. Often, X is assumed to possess some homogeneity properties so that nonconstant σ can induce deviations from homogeneity.

Under certain conditions, the models (1) and (2) lead to equivalent models. However, in general, this is not true, and we highlight in the following text some of the main differences between these two modeling approaches.

Time-changed Lévy Processes

In the finance literature, the main focus is on time-changed Brownian motion or, more generally, on time-changed Lévy processes (see **Lévy Processes**; **Time-changed Lévy Process**), since these processes possess natural homogeneity properties of stationary and independent increments (returns).

Time-changed Brownian motion has first been used as a model for (logarithmic) asset prices by Clark [14]. He investigated the case where $X = B = (B_t)_{t \geq 0}$ is a standard Brownian motion and where T is an independent continuous-time change. Clearly, in such a setting, the time-changed process $Y_s = B_{T_s}$ has a mixed normal distribution, that is, $Y_s | T_s \sim N(0, T_s)$, and is a continuous local martingale. The power of such models and those with the assumptions of independence and/or continuity relaxed is expressed in the following key results.

Theorem 1 (*Dubins–Schwarz*). [18, 34, 35] *Every continuous local martingale $M = (M_s)_{s \geq 0}$ can be written as a time-changed Brownian motion $(B_{[M]_s})_{s \geq 0}$, where $[M] = ([M]_s)_{s \geq 0}$ is the (continuous) quadratic variation of M .*

Also $M_t = \int_0^t \sigma_{r-} dW_r$ for $X = W = (W_t)_{t \geq 0}$ Brownian motion and $\sigma = (\sigma_t)_{t \geq 0}$ independent nonnegative with càdlàg sample paths is a continuous local martingale with quadratic variation $[M]_t =$

$[\int_0^t \sigma_r dW_r]_t = \int_0^t \sigma_r^2 dr$, which is often referred to as *integrated variance*.

Corollary 1 *In the context of the Dubins–Schwarz theorem, for the local martingale $M_s = \int_0^s \sigma_r dW_r = B_{[M]_s}$, independence of W and σ is equivalent to independence of B and $T = [M]$.*

Therefore, the models (1) and (2) are equivalent if X is a Brownian motion and $T_s = \int_0^s \sigma_r^2 dr$ is an absolutely continuous time change. Regarding Clark’s independence assumption, we note that independence of σ and X is also equivalent in models (1) and (2); however, this excludes the leverage effect, that is, the usually negative correlation between asset returns and volatility.

Fundamentally, these results are due to the scaling property of Brownian motion X , which translates spatial scaling σX_t or model (2) into temporal scaling $X_{\sigma^2 t}$ or model (1). Also, if Brownian motion is replaced by an α -stable Lévy process and $T_t = \int_0^t \sigma_s^\alpha ds$, the two models (1) and (2) are basically equivalent [25, 26]. No other Lévy process has such a scaling property, and indeed, the models (1) and (2) will be different. If we relate higher volatility to higher market activity, then model (1) suggests that markets move at a higher speed, and model (2) suggests that the volumes traded are higher.

Theorem 2 (Monroe [33]). *Every (càdlàg) semimartingale $Z = (Z_s)_{s \geq 0}$ can be written as a time-changed Brownian motion $(B_{T_s})_{s \geq 0}$ for a (càdlàg) family of stopping times $(T_s)_{s \geq 0}$ on a suitably extended probability space.*

In the light of the fundamental theorem of asset pricing (see **Fundamental Theorem of Asset Pricing**), this means that every arbitrage-free model can be viewed as time-changed Brownian motion. However, this result is of limited use for the construction of simple and natural parametric models.

In the finance literature, we find many asset price models where the base process is chosen to be a Lévy process other than Brownian motion (see, e.g., [12, 13]). Here, we refer to **Time-changed Lévy Process** for a detailed treatment of this class of models.

We also mention Lamperti’s representations [27, 28] of continuous-state branching processes and of self-similar Markov processes as time-changed Lévy processes very much in the spirit of the

Dubins–Schwarz Theorem. Salminen and Yor [36] use Lamperti’s time change to study Dufresne’s functional [20] that arises in the computation of discounted values in certain models of continuously payable perpetuities in actuarial science.

Choice of Time Change

There are different methods for choosing a time change that is suitable for financial models. Two classes of such processes are particularly popular: subordinators and absolutely continuous time changes [1].

In the finance literature, the terms *time change* and *subordinator* are sometimes used synonymously. However, in probability theory, the term subordinator describes a particular class of stochastic processes (as defined below) and does not include all time changes.

Subordinators

Subordinators are nondecreasing Lévy processes [10, 15, 37] and hence possess stationary and independent increments. They are pure jump processes of possibly infinite activity plus a deterministic linear drift. Clearly, they have no Brownian component and are of finite variation. Important examples include simple Poisson processes, increasing compound Poisson processes, gamma processes, and (tempered) stable subordinators. Note that many popular models in finance are based on time-changed Brownian motion where the time change is chosen to be a subordinator. For example, the variance gamma process [30, 31] can be represented as Brownian motion time-changed by a gamma process, the normal tempered stable process (including the normal inverse gaussian process [2, 3]) can be written as a Brownian motion time-changed by a tempered stable subordinator. Brownian motion, time-changed by an independent subordinator, yields a Lévy process always.

Absolutely Continuous Time Changes

Another important class of time changes is given by the class of absolutely continuous time changes of the form $T_s = \int_0^s \tau_u du$, for a positive and integrable process $\tau = (\tau_s)_{s \geq 0}$. Note that in such a setting T is always continuous, but τ can exhibit jumps. The process τ is often called *instantaneous (business)*

activity rate. Such models have been studied in the context when X is a Lévy process by Carr *et al.* [12] and [13] among others. The advantage of this model class is that it leads to affine models that are highly analytically tractable (see, e.g., [19, 25]), whereas stochastic integrals with respect to Lévy processes are in general not affine. Popular examples for the instantaneous activity rate are given by the Cox–Ingersoll–Ross process and the non-Gaussian Ornstein–Uhlenbeck process [7, 8, 16].

Multivariate Setting

Time-changed stochastic processes are also used in the finance literature to construct multivariate models. Usually, the base process X is then assumed to be multivariate (e.g., a multivariate Brownian motion) and the time-change process T is still assumed to be univariate ([15, 21] and the references therein). The advantage of such a modeling framework is the fact that such processes are highly analytically tractable and easy to simulate from. However, the range of dependence between the univariate components in such a multivariate model is rather limited (and, in particular, does not even include complete independence). Furthermore, such a model does not allow for an arbitrary choice of univariate models for the components [15]. The mathematical properties of multivariate processes X time-changed by a multivariate vector of time changes T have recently been studied in [5], but this has so far not been considered as a standard tool for constructing multivariate models in finance. So far, it has been much more common to use multivariate extensions of a stochastic integral (2) to construct multivariate stochastic volatility models, than to use time-changed processes. Finally, note that the concept of subordinators can also be generalized to matrix subordinators [4, 6] and their applicability in financial models is subject to ongoing research (see, e.g., [9, 38] for some first references).

Acknowledgments

We would like to thank Marc Yor and Ole Barndorff-Nielsen for suggesting relevant references. The first author acknowledges financial support by the Center for Research in Econometric Analysis of Time Series, CREATES, funded by the Danish National Research Foundation.

References

- [1] Ané, T. & Geman, H. (2000). Order flow, transaction clock, and normality of asset returns, *The Journal of Finance* **55**(5), 2259–2284.
- [2] Barndorff-Nielsen, O.E. (1997). Normal inverse gaussian distributions and stochastic volatility modelling, *Scandinavian Journal of Statistics* **24**, 1–13.
- [3] Barndorff-Nielsen, O.E. (1998). Processes of normal inverse Gaussian type, *Finance and Stochastics* **2**, 41–68.
- [4] Barndorff-Nielsen, O.E., Maejima, M. & Sato, K.-I. (2006). Some classes of multivariate infinitely divisible distributions admitting stochastic integral representation, *Bernoulli* **12**, 1–33.
- [5] Barndorff-Nielsen, O.E., Pedersen, J. & Sato, K.-I. (2001). Multivariate subordination, selfdecomposability and stability, *Advances in Applied Probability* **33**, 160–187.
- [6] Barndorff-Nielsen, O.E. & Pérez-Abreu, V. (2008). Matrix subordinators and related upilon transformations, *Theory of Probability and its Applications* **52**(1), 1–23.
- [7] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein–Uhlenbeck–based models and some of their uses in financial economics (with discussion), *Journal of the Royal Statistical Society, Series B-Statistical Methodology* **63**, 167–241.
- [8] Barndorff-Nielsen, O.E. & Shephard, N. (2002). Econometric analysis of realised volatility and its use in estimating stochastic volatility models, *Journal of the Royal Statistical Society. Series B* **64**, 253–280.
- [9] Barndorff-Nielsen, O.E. & Stelzer, R. (2007). Positive-definite matrix processes of finite variation, *Probability and Mathematical Statistics* **27**(1), 3–43.
- [10] Bertoin, J. (1996). *Lévy Processes*, Cambridge University Press, Cambridge.
- [11] Bochner, S. (1949). Diffusion equation and stochastic processes, *Proceedings of the National Academy of Sciences of the United States of America* **85**, 369–370.
- [12] Carr, P., Geman, H., Madan, D.B. & Yor, M. (2003). Stochastic volatility for Lévy processes, *Mathematical Finance* **13**(3), 345–382.
- [13] Carr, P. & Wu, L. (2004). Time-changed Lévy processes and option pricing, *Journal of Financial Economics* **71**, 113–141.
- [14] Clark, P.K. (1973). A subordinated stochastic process model with fixed variance for speculative prices, *Econometrica* **41**, 135–156.
- [15] Cont, R. & Tankov, P. (2004). *Financial Modelling With Jump Processes*, *Financial Mathematics Series*, Chapman & Hall, Boca Raton, Florida.
- [16] Cox, J.C., Ingersoll, J.E. & Ross, S.A. (1985). An intertemporal general equilibrium model of asset prices, *Econometrica* **53**, 363–384.
- [17] Doëblin, W. (2000). Sur l'équation de Kolmogoroff, *Les Comptes Rendus de l'Académie des Sciences* **I**(331),

- 1031–1187. Pli cacheté déposé le 26 février 1940, ouvert le 18 mai 2000.
- [18] Dubins, L. & Schwarz, G. (1965). On continuous martingales, *Proceedings of the National Academy of Sciences of the United States of America* **53**(5), 913–916.
- [19] Duffie, D., Filipovic, D. & Schachermayer, W. (2003). Affine processes and applications in finance, *Annals of Applied Probability* **13**, 984–1053.
- [20] Dufresne, D. (1990). The distribution of a perpetuity, with applications to risk theory and pension funding, *Scandinavian Actuarial Journal* **1990**, 39–79.
- [21] Eberlein, E. (2001). Application of generalized hyperbolic Lévy motion to finance, in *Lévy Processes: Theory and Applications*, O.E. Barndorff-Nielsen, T. Mikosch & S. Resnick, eds, Birkhäuser, Basel, pp. 319–337.
- [22] Geman, H. (2005). From measure changes to time changes in asset pricing, *Journal of Banking and Finance* **29**, 2701–2722.
- [23] Hunt, G. (1957). Markov processes and potentials I, II and III, *Illinois Journal of Mathematics* **1**, 44–93; 316–396; (1958). **2**, 151–213.
- [24] Itô, K. & McKean, H.P. (1965). *Diffusion Processes and Their Sample Paths*, Springer–Verlag, Berlin.
- [25] Kallsen, J. (2006). A didactic note on affine stochastic volatility models, in *From Stochastic Calculus to Mathematical Finance*, Y. Kabanov, R. Liptser & J. Stoyanov, eds, Springer–Verlag, Berlin, pp. 343–368.
- [26] Kallsen, J. & Shiryaev, A. (2002). Time change representation of stochastic integrals, *Theory of Probability and its Applications* **46**, 522–528.
- [27] Lamperti, J. (1967). Continuous–state branching processes, *Bulletin of The American Mathematical Society* **73**, 382–386.
- [28] Lamperti, J. (1972). *Semi–Stable Markov Processes. I, Probability Theory and Related Fields*, Springer–Verlag, Vol. 22, pp. 205–225.
- [29] Lévy, P. (1948). *Processus Stochastiques Et Mouvement Brownien*, Gauthier–Villars, Paris.
- [30] Madan, D., Carr, P. & Chang, E. (1998). The variance gamma process and option pricing, *European Finance Review* **2**, 79–105.
- [31] Madan, D.B. & Seneta, E. (1990). The VG model for share market returns, *Journal of Business* **63**, 511–524.
- [32] McKean, H. (2002). Scale and clock, in *Mathematical Finance—Bachelier Congress 2000*, Springer Finance, H. Geman, D. Madan, S. Pliska & T. Vorst, eds, Springer–Verlag, Berlin.
- [33] Monroe, I. (1978). Processes that can be embedded in Brownian motion, *The Annals of Probability* **6**(1), 42–56.
- [34] Protter, P.E. (2004). *Stochastic Integration and Differential Equations*, 2nd Edition, Springer, London.
- [35] Revuz, D. & Yor, M. (2001). *Continuous Martingales and Brownian Motion*, 3rd Edition, Springer, Berlin.
- [36] Salminen, P. & Yor, M. (2005). Perpetual integral functionals as hitting and occupation times, *Electronic Journal of Probability* **10**, 371–419.
- [37] Sato, K. (1999). *Lévy processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge.
- [38] Stelzer, R.J. (2007). *Multivariate Continuous time stochastic volatility models driven by a Lévy process*, Dissertation, Technische Universität München, München. Urn: <http://d-nb.info/986220337>.
- [39] Volkonski, V. (1958). Random time changes in strong Markov processes, *Theory of Probability and its Applications* **3**, 310–326.

Further Reading

- Fu, M., Jarrow, R., Yen, J.-Y. & Elliott, R. (eds) (2007). *Advances in Mathematical Finance, Applied and Numerical Harmonic Analysis*, Birkhäuser, Boston.
- Jeanblanc, M., Yor, M. & Chesney, M. (2009). *Mathematical Methods for Financial Markets*, Springer Finance, Springer–Verlag, Berlin.

Related Articles

Jump Processes; Lévy Processes; Martingales; Normal Inverse Gaussian Model; Realized Volatility and Multipower Variation; Stochastic Integrals; Time-changed Lévy Process; Variance-gamma Model; Volatility.

ALMUT E.D. VERAART & MATTHIAS WINKEL

Stylized Properties of Asset Returns

Since prices of different stocks are not necessarily influenced by the same events or information sets, price series obtained from different assets and—*a fortiori*—from different markets may exhibit different (statistical) properties. After all, why should properties of corn futures be similar to those of IBM shares or the dollar/yen exchange rate? Nevertheless, the result of more than half a century of empirical studies on financial time series indicates that this is the case if one examines their properties from a *statistical* point of view. The seemingly random variations of asset prices *do* share some nontrivial statistical properties. Such properties, common across a wide range of instruments, markets, and time periods, are called *stylized empirical facts* [10].

This assertion is backed by a vast research literature dealing with the statistical properties of financial time series which has developed during the last 50 years. Some early references include [25, 26, 29]. Surveys of stylized facts of financial time series are given in [17] for foreign exchange rates and in [5, 8, 10, 16, 30] for stocks and indices. Scaling and time-aggregation properties for high-frequency data are studied in [2] for stocks and in [17, 18, 27, 28] for foreign exchange rates. Numerous empirical studies on the scaling properties of returns can also be found in the physics literature [7].

Stylized facts are thus obtained by taking a common denominator among the properties observed in studies of different markets and instruments. Obviously, by doing so one gains in generality but tends to lose in the precision of the statements one can make about asset returns. Indeed, stylized facts are usually formulated in terms of *qualitative properties* of asset returns and may not be precise enough to distinguish among different parametric models. Nevertheless, albeit qualitative, these stylized facts are so constraining that it is not even easy to exhibit an (*ad hoc*) stochastic process which possesses the same set of properties, and stochastic modeling has gone to great lengths to reproduce these stylized facts.

Stylized facts are usually formulated in terms of asset (log-)returns defined as

$$r_t(\Delta) = \ln \frac{S_{t+\Delta}}{S_t} \quad (1)$$

where S_t is the price of an asset at date t and Δ is a time horizon. The properties of returns series $r_t(\Delta)$ greatly depend on the sampling frequency, that is, on Δ .

We focus here on empirical properties for *daily* (log-)returns of stocks and stock market indices in liquid markets and refer to **High-frequency Data** and **Market Microstructure Effects** for properties of intraday data.

1. **Heavy tails:** The (unconditional) distributions of returns display—at time horizons ranging from an hour to several weeks—heavy tails and positive sample excess kurtosis, properties that both reject the Gaussian assumption (Figure 1). A more precise formulation requires measuring the heaviness of the tails. For heavy tails with “regular variation” (*see Extreme Value Theory*), that is, of the type

$$P(r_t \geq x) = \frac{l_+(x)}{|x|^{\alpha_+}}, \quad P(r_t \leq x) = \frac{l_-(x)}{|x|^{\alpha_-}} \quad (2)$$

where l_+ (respectively l_-) is slowly varying at $+\infty$ (respectively $-\infty$) and $\alpha_+, \alpha_- > 0$ are the (right/left) tail exponents, the heaviness of the tails can be measured by the value of α_+, α_- . $\xi_+ = 1/\alpha_+$ is sometimes called the (right) *tail index*. Light tailed distributions with exponential-type tail have a zero tail index, while a strictly positive tail index indicates heavy tails of regularly varying type. Tail exponents can be estimated using a Hill estimator or using methods based on the extreme value theory (*see Extreme Value Theory*). Figure 2 shows the estimation of right and left tail indices for SP100 daily returns: in this example we find $\alpha_+ \simeq 4$ and $\alpha_- \simeq 3$. For stock and index returns, tail exponents are often found to be finite, higher than two and less than five for most data sets studied [20, 23, 24]. In particular, this excludes stable laws with infinite variance and the normal distribution. Power-law or Pareto tails reproduce such behavior. But, as argued, for example, in Heyde and Kou [19], with small samples exponential tails (“semi-heavy tails”) may be difficult to distinguish from Pareto tails.

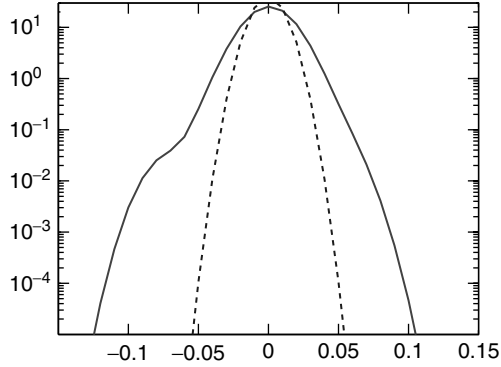


Figure 1 Kernel estimator of empirical density of SP100 daily returns (solid line) vs Gaussian density (dotted line) with same mean and variance, on a logarithmic scale. Note the dominance of the tails with respect to the Gaussian and the asymmetry of the empirical density

2. **Absence of autocorrelations:** (Linear) autocorrelations of asset returns are often insignificant, except at short lags ($\simeq 10$ min) for which microstructure effects come into play (*see Market Microstructure Effects*). Figure 3 illustrates this fact for the yen–US dollar exchange rate. This absence of significant autocorrelation can be attributed to the existence of arbitrageurs who “whiten the spectrum of price changes”, making their second-order properties indistinguishable from a white noise sequence [26]. A consequence of this absence of autocorrelation is the additivity of variance or “square-root” scaling rule: the variance of returns at horizon $n\Delta$ is n times the variance at horizon Δ in the absence of autocorrelations; thus the standard deviation should be scaled by $\sqrt{n}$. This scaling is the basis of the variance ratio test

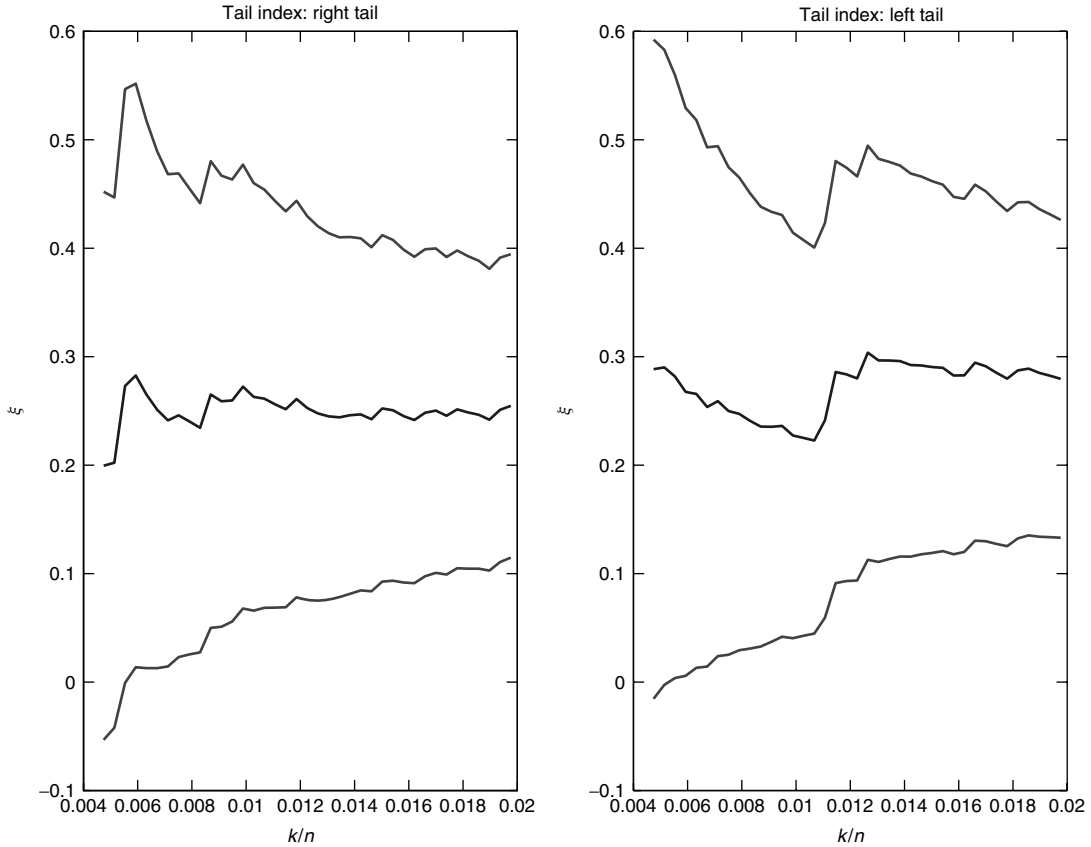


Figure 2 Hill estimators for the tail index of SP100 daily returns. We obtain $\alpha_+ \simeq 4$ and $\alpha_- \simeq 3$. Note the large standard errors

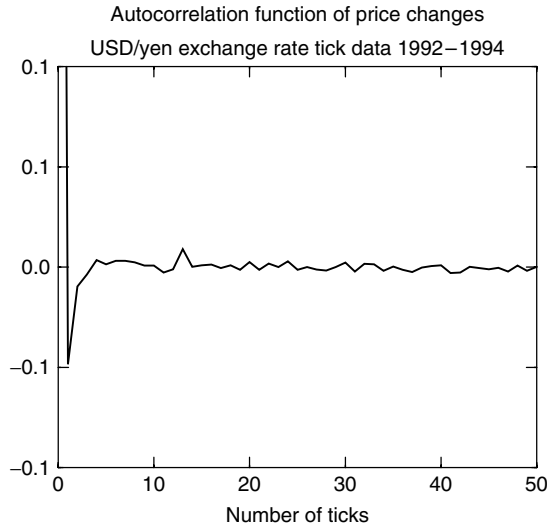


Figure 3 Autocorrelation function of log-returns, USD/yen exchange rate, $\Delta = 5$ min

for autocorrelation [8]. Note that this scaling rule breaks down for illiquid assets for which one typically observes high autocorrelation of returns.

3. **Gain/loss asymmetry:** Unconditional distributions of stock returns are often asymmetric. Although, this asymmetry may go undetected, when using moment-based measures such as skewness, one often observes a left tail which is heavier than the right tail, as attested by the tail index (*see Heavy Tails*) or extremal index (*see Extreme Value Theory*) estimates [20, 24]. For exchange rates between major currencies, there is a higher symmetry in up/down moves and the skewness can change sign from one period to another.

Daily changes in credit spreads (e.g., credit default swap spreads, *see Credit Default Swaps*) show an opposite asymmetry: they exhibit a heavier right (upper) tail and there are more frequent large upward jumps in credit spreads.

4. **Aggregational normality:** As one increases the timescale Δ over which returns are calculated, their distribution looks more and more like a normal distribution. In particular, the shape of the distribution is not the same at different timescales: the heavy-tailed feature

becomes less pronounced as the time horizon is increased.

5. **Volatility clustering:** “Large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes” [25]. A quantitative manifestation of this fact is that, while returns themselves are uncorrelated, absolute returns $|r_t(\Delta)|$, their squares, as well as other implied or historical volatility indicators display a positive, significant, and slowly decaying autocorrelation function. In particular, this observation rules out the assumption of independence of returns and points to nonlinear dependence in returns across time.

This property is sometimes called the “ARCH effect” in the econometrics literature, because of the fact that ARCH models (*see GARCH Models*) have this property. This is a misnomer: volatility clustering is a model-free property of data that has nothing to do with ARCH models.

6. **Conditional heavy tails:** Even after adjusting returns for volatility clustering (i.e. normalizing returns by a dynamic estimator of volatility, such as in GARCH-type models), the residuals (normalized returns) still exhibit heavy tails [5, 14]. However, the tails are less heavy than in the unconditional distribution of returns.

This remark has motivated the development of conditionally heteroskedastic models driven by non-Gaussian noise in discrete time and Lévy-driven stochastic volatility models in continuous time [4, 5, 13, 14] (*see GARCH Models; Jump Processes; Time-changed Lévy Process; Regime-switching Models*).

7. **Slow decay of autocorrelation in absolute returns:** The autocorrelation function of absolute (or squared) returns decays slowly as a function of the time lag, roughly as a power-law with an exponent $\beta \leq 1$. An example is shown in Figure 4. This is sometimes interpreted as a sign of long-range dependence in volatility (*see Long Range Dependence; Multifractals*).

It has been argued [4, 22] that the decay of the autocorrelation of $|r_t|$ can also be reproduced by a superposition of several exponentials, indicating that the dependence is characterized by multiple timescales. In fact, an operational definition of long-range dependence is that the timescale of dependence in a sample of length

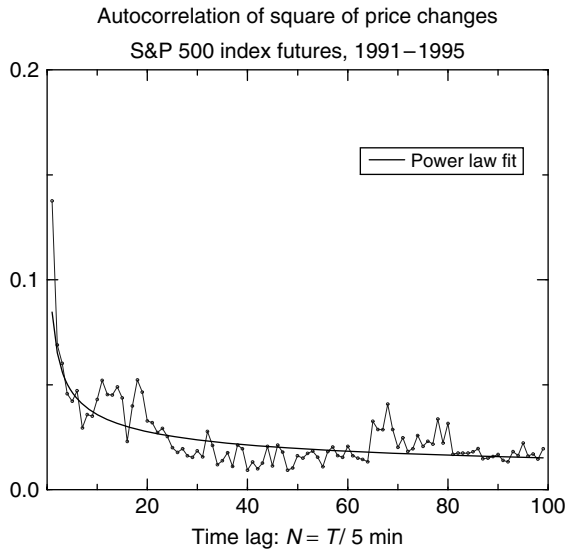


Figure 4 Autocorrelation function of squared log-returns: S&P500 futures, $\Delta = 30$ min

T is found to be of the order of T : dependence extends over the whole sample. Interestingly, the largest timescale in [22] is found to be of the order of the sample size, a prediction which would be compatible with long-range dependence [11].

8. **“Leverage” effect:** Volatility increases when the stock price falls. More precisely, changes in implied or historical indicators of volatility of an asset are negatively correlated with the returns of that asset. In fact, there is evidence of a lead–lag relationship: a negative correlation between *past* returns and (current) changes in volatility [6].

The name “leverage effect” comes from a supposed explanation based on a firm’s capital structure: a decrease in the market valuation of a firm’s equity increases the degree of leverage in its capital structure, leading to an increase in stock volatility. However, empirical studies do not seem to confirm this “explanation” [3] and the “leverage effect” should be seen more in terms of an asymmetric impact of news on returns and volatility [15].

9. **Volume–volatility correlation:** Trading volume in a given period (e.g., daily trading volume) is positively correlated with measures of volatility (e.g., realized volatility on the same day).

The same holds for other measures of market activity such as the number of trades.

These observations have motivated the construction of models in which the dynamics of prices is modeled as a function of cumulative volume [9] or number of transaction, and has inspired models based on time-changed processes (*see Time-changed Lévy Process; Mixture of Distribution Hypothesis; Time Change*).

10. **Asymmetry in time scales:** Volatility indicators computed at lower frequency predict future volatility at higher frequency, whereas past values of high-frequency volatility indicators do not necessarily have predictive power for future low-frequency volatility [18]. This observation has led to the development of “random cascade” models, in which market shocks affect price behavior at low frequencies (fine scales) and their impact trickles down to lower scales (*see Multifractals; [27]*).

It may be noted that we have not included “self-similarity” or scale invariance among the above properties. In fact, there is ample empirical evidence showing that asset returns are *not* scale invariant or self-similar [12]. For example, the shape of the distribution of returns $r_t(\Delta)$ at a horizon Δ greatly varies with Δ and “scales” in a nontrivial way, which is not captured by simple self-similar models such as α -stable Lévy processes (*see Lévy Processes*) or (fractional) Brownian motion (*see Fractional Brownian Motion*).

Stylized Facts as Model Benchmarks

Stylized empirical facts have motivated research in two different directions. One is the development of stochastic models—in discrete or continuous time—whose aim is to match some or all of the observed properties of financial time series. For instance, the (G)ARCH family of models (*see GARCH Models*) was mainly developed to model volatility clustering and heavy tails in asset returns, while stochastic volatility models were developed to model volatility clustering and the leverage effect. Similarly, models based on Lévy processes (*see Exponential Lévy Models*) were initially introduced to capture the time-aggregation properties (heavy tails

at high-frequency, scaling with respect to the horizon Δ) of the distribution of returns.

Another strand of research has aimed at proposing economic explanations for the origins of stylized facts. Representative agent models attempt to explain stylized facts in terms of the behavior of a hypothetical representative investor whose behavior generates market price changes. Ait-Sahalia [1] shows that this approach can go a long way in terms of reproducing, if not explaining, stylized empirical facts. By contrast, agent-based market models [11, 21] tend to view stylized facts as properties *emerging* from the collective behavior of a large number of market participants. In all the cases, stylized empirical properties of asset returns serve as benchmarks for evaluating the plausibility of financial models before any estimation or testing is done.

References

- [1] Ait-Sahalia, Y. (1998). Dynamic equilibrium and volatility in financial asset markets, *Journal of Econometrics* **84**, 93–128.
- [2] Andersen, T. & Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets, *Journal of Empirical Finance* **4**, 115–158.
- [3] Aydemir, C., Gallmeyer, M. & Hollifield, B. (2007). *Financial Leverage and the Leverage Effect: A Market and Firm Analysis*. Working Paper 2007-031, Carnegie Mellon University.
- [4] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck based models and some of their uses in financial econometrics, *Journal of the Royal Statistical Society Series B* **63**, 167–241.
- [5] Bollerslev, T., Chou, R. & Kroner, K. (1992). ARCH modeling in finance, *Journal of Econometrics* **52**, 5–59.
- [6] Bouchaud, J.-P., Matacz, A. & Potters, M. (2001). The leverage effect in financial markets: retarded volatility and market panic, *Physical Review Letters* **87**, 2287–2301.
- [7] Bouchaud, J. & Potters, M. (1997). *Théorie des Risques Financiers*, Aléa, Saclay.
- [8] Campbell, J., Lo, A. & McKinlay, C. (1996). *The Econometrics of Financial Markets*, Princeton University Press, Princeton.
- [9] Clark, P.K. (1973). A subordinated stochastic process model with finite variance for speculative prices, *Econometrica* **41**, 135–155.
- [10] Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues, *Quantitative Finance* **1**, 1–14.
- [11] Cont, R. (2005). Volatility clustering in financial markets: empirical facts and agent-based models, in *Long Memory in Economics*, A. Kirman & G. Teyssiere, eds, Springer, Berlin.
- [12] Cont, R., Bouchaud, J.-P. & Potters, M. (1997). Scaling in financial data: stable laws and beyond, in *Scale Invariance and Beyond*, B. Dubrulle, F. Graner & D. Sornette, eds, Springer, Berlin.
- [13] Cont, R. & Tankov, P. (2004). *Financial Modelling with Jump Processes*, Chapman & Hall/CRC Press.
- [14] Engle, R. (ed.) (1995). ARCH: selected readings, in *Advanced Texts in Econometrics*, Oxford University Press.
- [15] Engle, R. & Ng, V. (1993). Measuring and testing the impact of news on volatility, *Journal of Finance* **48**, 1749–1778.
- [16] Fama, E. (1965). The behavior of stock market prices, *Journal of Business* **38**, 34–105.
- [17] Guillaume, D., Dacorogna, M., Davé, R., Müller, U., Olsen, R. & Pictet, O. (1997). From the birds eye view to the microscope: a survey of new stylized facts of the intraday foreign exchange markets, *Finance and Stochastics* **1**, 95–131.
- [18] Guillaume, D., Dacorogna, M., Davé, R., Müller, U., Olsen, R. & Pictet, O. (1997). Volatilities at different time resolutions: analyzing the dynamics of market components, *Journal of Empirical Finance* **4**, 213–239.
- [19] Heyde, C. & Kou, S. (2004). On the controversy over tailweight of distributions, *Operations Research Letters* **32**, 399–408.
- [20] Hols, M. & De Vries, C. (1991). The limiting distribution of extremal exchange rate returns, *Journal of Applied Econometrics* **6**, 287–302.
- [21] LeBaron, B. (2000). Agent-based computational finance: suggested readings and early research, *Journal of Economic Dynamics and Control* **24**, 679–702.
- [22] LeBaron, B. (2001). Stochastic volatility as a simple generator of apparent financial power laws and long memory, *Quantitative Finance* **1**, 621–631.
- [23] Longin, F. (1996). The asymptotic distribution of extreme stock market returns, *Journal of Business* **69**, 383–408.
- [24] Longin, F. (2005). The choice of the distribution of asset prices : how extreme value theory can help, *Journal of Banking and Finance* **29**, 1017–1035.
- [25] Mandelbrot, B.B. (1963). The variation of certain speculative prices, *Journal of Business* **XXXVI**, 392–417.
- [26] Mandelbrot, B.B. (1971). When can price be arbitrated efficiently? A limit to the validity of the random walk and martingale models, *The Review of Economics and Statistics* **53**, 225–236.
- [27] Mandelbrot, B.B., Calvet, L. & Fisher, A. (1997). *The Multifractality of the Deutschmark/US Dollar Exchange Rate*, Discussion paper 1166, Cowles Foundation for Economics, Yale University.
- [28] Müller, U., Dacorogna, M. & Pictet, O. (1998). Heavy tails in high-frequency financial data, in *A Practical Guide to Heavy Tails: Statistical Techniques for*

- Analysing Heavy Tailed Distributions*, R. Adler, R. Feldman & M. Taqqu, eds, Birkhäuser, Boston, pp. 55–77.
- [29] Officer, R. (1972). The distribution of stock returns, *Journal of the American Statistical Association* **67**, 807–812.
- [30] Pagan, A. (1996). The econometrics of financial markets, *Journal of Empirical Finance* **3**, 15–102.

Long Range Dependence; Mixture of Distribution Hypothesis; Random Matrix Theory; Realized Volatility and Multipower Variation; Stochastic Volatility Models; Extremal Behavior; Stochastic Volatility Models; Time Change; Time-changed Lévy Process; Volatility.

RAMA CONT

Related Articles

Exponential Lévy Models; Extreme Value Theory; GARCH Models; Heavy Tails; Jump Processes;

Stochastic Volatility Models: Extremal Behavior

Extreme value theory for financial models mostly concerns the martingale (see **Martingales**) component of the logarithm of a price process, since random volatility determines the extreme risk in price fluctuations. The increments $(Y_n)_{n \in \mathbb{Z}}$ and $(Y_t)_{t \in \mathbb{R}}$ of length 1 of this martingale part often have the structure

$$Y_n = \sigma_n \varepsilon_n, \quad n \in \mathbb{Z}, \quad \text{or} \quad Y_t = \int_{(t-1, t)} \sigma_{s-} dL_s, \quad t \in \mathbb{R} \quad (1)$$

for a discrete-time or continuous-time model, respectively. Here, the volatility is modeled by σ , and $(\varepsilon_n)_{n \in \mathbb{Z}}$ or $(L_t)_{t \in \mathbb{R}}$ are typically independent and identically distributed (i.i.d.) sequences or a Lévy process (see **Lévy Processes**), respectively. The usual prerequisite of extreme value theory for a stochastic process is its strict stationarity. Note that, in most cases, strict stationarity of the log price increment process Y is inherited from stationarity of the volatility process σ . Consequently, we present conditions for strict stationarity of the models below, followed by the extremal analysis of a stationary version.

The importance of extreme value theory for such pricing models is twofold. First, the *tail behavior* for large absolute arguments describes the fluctuations of the prices. We distinguish between light- and heavy-tailed models: light tailed means normal or exponential models, whereas heavy-tailed models are defined *via* regular variation. We say a random variable X has regularly varying tail, if there are $\alpha > 0$ and slowly varying $\ell : (0, \infty) \rightarrow (0, \infty)$, that is, $\lim_{x \rightarrow \infty} \ell(tx)/\ell(x) = 1$ for all $t > 0$, such that

$$P(X > x) = \ell(x)x^{-\alpha}, \quad x > 0 \quad (2)$$

We write $X \in \mathcal{R}(-\alpha)$. If the volatility σ has regularly varying tail and the noise ε or L is light tailed and independent of the volatility, then the log price increments are again regularly varying by Breiman's classical result. The same holds for a light-tailed volatility and independent regularly varying noise. If both, volatility and noise, are light

tailed, the distribution of the product depends even asymptotically on both factors and may be more difficult to estimate. Symmetry or tail balance of the right and left tails of the noise often simplifies the calculations.

Second, *volatility clusters on high levels* induce extreme price clusters, which can cause a particularly risky situation. For discrete-time models, such clusters are described by a limit of point processes of exceedances over high thresholds. For certain models, this limit is simply a Poisson process, indicating that exceedances happen as single points and at completely random times. However, more realistic models capture the fact that the limit process is not a Poisson process but a compound Poisson process, which describes also the cluster size distribution. A crude, but simple, measure of the cluster size of extremes is the *extremal index* $\theta \in (0, 1]$, where $1/\theta$ can be interpreted as the mean size of a cluster. It also appears in the distributional limit of the running maxima

$$M_n^Z = \max\{Z_1, \dots, Z_n\}, \quad n \in \mathbb{N} \quad (3)$$

of a stationary sequence $(Z_n)_{n \in \mathbb{Z}}$. More precisely, under weak conditions on Z_1 , there exist $a_n > 0$ and $b_n \in \mathbb{R}$ such that the *Poisson condition*

$$nP(Z_1 > a_n x + b_n) = -\log G(x), \quad x \in \mathbb{R} \quad (4)$$

holds, and if the process satisfies further mixing conditions, then the extremal index of $(Z_n)_{n \in \mathbb{Z}}$ is the unique number $\theta \in (0, 1]$ such that

$$\lim_{n \rightarrow \infty} P(a_n^{-1}(M_n^Z - b_n) \leq x) = G^\theta(x), \quad x \in \mathbb{R} \quad (5)$$

Here, G is an extreme value distribution, that is, it is of the same type as a Fréchet distribution $\Phi_\alpha(x) = \mathbf{1}_{(0, \infty)}(x) \exp(-x^{-\alpha})$ for some $\alpha > 0$ (heavy-tailed case), a Gumbel distribution $\Lambda(x) = \exp(-e^{-x})$ (light-tailed case), or a Weibull distribution Ψ_α for some $\alpha > 0$. We refer to the books [11, 24] for this and further information about extreme value theory.

For $\theta = 1$, we interpret the stochastic process as a process without clusters in the extremes, whereas if $\theta < 1$ we speak of a process with cluster possibilities in the extremes. Using a point process approach, the whole cluster size distribution can be derived, giving

much more insight into the extremal behavior of time series models. This, however, would go beyond the scope of this article and we refer to the references given throughout for more details.

As for discrete-time models, the extremal behavior of continuous-time models can be described by the limit behavior of point processes. A simple measure is again given by an extension of the extremal index: if $(Z_t)_{t \in \mathbb{R}}$ is a continuous-time process, we define for fixed $h > 0$ the discrete-time process

$$M_k^Z(h) := \sup_{(k-1)h \leq t \leq kh} Z_t, \quad k \in \mathbb{Z} \quad (6)$$

Denoting $\theta(h)$ as the extremal index of the sequence $(M_k^Z(h))_{k \in \mathbb{Z}}$, we follow [13] and call $\theta(h)$ for $h \in (0, \infty)$ the *extremal index function* of Z . The function θ is increasing, and we shall say that the continuous-time process Z has extremal clusters if $\lim_{h \rightarrow 0} \theta(h) < 1$, that is, if there is some $h > 0$ such that the discrete skeleton $(M_k^Z(h))_{k \in \mathbb{Z}}$ has extremal index less than 1, that is, it clusters.

There exist various publications on extreme value theory for time-dependent data; we mention, for example, [7, 8, 11, 17–20, 24] and references therein.

Discrete-time Stochastic Volatility Models

The simple stochastic volatility model is given by

$$Y_n = \sigma_n \varepsilon_n, \quad \log \sigma_n^2 = \alpha_0 + \sum_{j=0}^{\infty} c_j \eta_{n-j}, \quad n \in \mathbb{Z} \quad (7)$$

where $(\eta_n)_{n \in \mathbb{Z}}$ is i.i.d. $N(0, s^2)$ with $\sum_{j=0}^{\infty} c_j^2 < \infty$ and $(\varepsilon_n)_{n \in \mathbb{Z}}$ is i.i.d., independent of $(\eta_n)_{n \in \mathbb{Z}}$. This covers the case when $(\log \sigma_n^2)_{n \in \mathbb{Z}}$ is a causal ARMA process with i.i.d. Gaussian noise $(\eta_n)_{n \in \mathbb{Z}}$, the most prominent case being the volatility model of Taylor [28]:

$$Y_n = \sigma_n \varepsilon_n, \quad \log \sigma_n^2 = \alpha_0 + \psi \log \sigma_{n-1}^2 + \eta_n, \quad n \in \mathbb{Z}, \quad |\psi| < 1 \quad (8)$$

when the log volatility is a causal Gaussian AR(1) process.

Extreme value analysis for equation (7) is based on the transformation

$$X_n = \log Y_n^2 = \alpha_0 + \sum_{j=0}^{\infty} c_j \eta_{n-j} + \log \varepsilon_n^2, \quad n \in \mathbb{Z} \quad (9)$$

which is a Gaussian linear process plus an i.i.d. noise. From [5, 7] we have the following theorem:

Theorem 1 (Tail Behavior and Extremes of the Stochastic Volatility Model). *Assume the stochastic volatility model (7) as above with X_n defined by equation (9).*

(a) *If ε_1 is $N(0, 1)$ denote $\tilde{s}^2 = s^2 \sum_{j=0}^{\infty} c_j^2$ and $k = \log(2/\tilde{s}^2)$.*

Then the tail of the stationary process $(X_n)_{n \in \mathbb{Z}}$ satisfies as $x \rightarrow \infty$

$$P(X_1 > x + \alpha_0) = \frac{\tilde{s}^2}{\sqrt{\pi}} \exp \left\{ -\frac{x^2}{2\tilde{s}^2} + \frac{x \log x}{\tilde{s}^2} + \frac{(k-1)x}{\tilde{s}^2} + \frac{(k+\tilde{s}^2) \log x}{\tilde{s}^2} - \frac{(\log x)^2}{2\tilde{s}^2} - \frac{k^2}{2\tilde{s}^2} + O\left(\frac{(\log x)^2}{x}\right) \right\} \quad (10)$$

The tail of the stationary log price increments is given by

$$P(Y_1 > \sqrt{y}) = \frac{1}{2} P(|Y_1| > \sqrt{y}) = \frac{1}{2} P(X_1 > \log y), \quad n \in \mathbb{Z} \quad (11)$$

If, furthermore, the autocorrelation function ρ of $(\log \sigma_n^2)_{n \in \mathbb{Z}}$ satisfies

$$\rho(h) = \text{corr}(\log \sigma_n^2, \log \sigma_{n+h}^2) = o((\log h)^{-1}), \quad h \rightarrow \infty \quad (12)$$

then each of the sequences $(X_n)_{n \in \mathbb{Z}}$ and $(Y_n)_{n \in \mathbb{Z}}$ has extremal index 1, and the extreme value distribution G appearing in equations (4) and (5) is the Gumbel distribution Λ for both X and Y .

(b) Assume that $\varepsilon_1 \in \mathcal{R}(-\alpha)$ for some $\alpha > 0$. Then as $x \rightarrow \infty$ we have

$$P(Y_1 > x) \sim E(\sigma_1^\alpha) P(\varepsilon_1 > x) \quad (13)$$

If, moreover, for $p \in (0, 1]$ the tail balance condition for $x \rightarrow \infty$ $P(\varepsilon_1 > x) \sim p P(|\varepsilon_1| > x)$ holds, then $(Y_n)_{n \in \mathbb{Z}}$ has extremal index 1 and the extreme value distribution G appearing in (4) and (5) is the Fréchet distribution Φ_α .

Here, $f(x) \sim g(x)$ as $x \rightarrow \infty$ for two strictly positive functions f and g means that $\lim_{x \rightarrow \infty} f(x)/g(x) = 1$. Berman's condition (12) is very weak and, for example, satisfied, whenever $\log \sigma_n^2$ follows a causal ARMA equation (in particular, for the volatility model (8) of Taylor). This means that for most stochastic volatility models (7) with Gaussian η and either light- or heavy-tailed noise ε , the extremal index is 1, so that the point processes of exceedances over high thresholds converge to a Poisson process. The model cannot model clusters of extremes.

Extensions of the η_n to non-Gaussian random variables in equation (7) have been considered. In most cases, the qualitative behavior remains the same provided that the processes σ and ε are independent and η_1 is light-tailed [9, 10, 22].

EGARCH model

A model related to stochastic volatility models is the EGARCH model of Nelson [26] given by

$$Y_n = \sigma_n \varepsilon_n, \quad \log \sigma_n^2 = \alpha_0 + \sum_{j=1}^{\infty} c_j g(\varepsilon_{n-j}), \quad n \in \mathbb{Z} \quad (14)$$

where $(\varepsilon_n)_{n \in \mathbb{Z}}$ is i.i.d. normal (or more generally follows a generalized error distribution $\text{GED}(\nu)$), the real coefficients α_0 and $(c_j)_{j \in \mathbb{N}}$ decay sufficiently fast, and g is a deterministic function. The infinite moving average representation for $\log \sigma^2$ typically arises from EGARCH(p, q) equations of the form

$$\log \sigma_n^2 = \alpha_0 + \sum_{j=1}^p \alpha_j g(\varepsilon_{n-j}) + \sum_{j=1}^q \beta_j \log(\sigma_{n-j}^2) \quad (15)$$

with real coefficients α_j and β_j . The standard choice for g is $g(x) = \varphi x + \gamma(|x| - E|\varepsilon_0|)$, where φ and γ are real constants, so that g is an affine linear function and $g(\varepsilon_n)$ allows the volatility to respond asymmetrically to negative and positive innovations. Depending on the size of φ and γ , different cases arise, but the most important one is for $\gamma - \varphi > \gamma + \varphi > 0$, in which case a negative innovation increases the volatility more than a positive innovation of the same modulus (i.e., it models the leverage effect).

As for stochastic volatility models, extreme value analysis for equation (14) is based on the transformed process (which is stationary by equation 14)

$$\begin{aligned} X_n &= \log Y_n^2 \\ &= \alpha_0 + \sum_{j=1}^{\infty} c_j g(\varepsilon_{n-j}) + \log \varepsilon_n^2, \quad n \in \mathbb{Z} \end{aligned} \quad (16)$$

The following theorem follows from [9, 10, 22]:

Theorem 2 (Tail Behavior and Extremes of EGARCH). Assume the EGARCH model as above with $(\varepsilon_n)_{n \in \mathbb{N}}$ i.i.d. $N(0, 1)$ and $\gamma - \varphi > \gamma + \varphi > 0$. Suppose further that the coefficients $(c_j)_{j \in \mathbb{N}}$ are nonnegative and that $c_j = O(j^{-\delta})$ as $j \rightarrow \infty$ for some $\delta > 1$. Denote

$$A := (\gamma - \varphi)^2 \sum_{j=1}^{\infty} c_j^2 \quad \text{and} \quad B := -\gamma E|\varepsilon_0| \sum_{j=1}^{\infty} c_j \quad (17)$$

Then the stationary processes $(X_n)_{n \in \mathbb{Z}}$ and $(Y_n)_{n \in \mathbb{Z}}$ satisfy, as $x \rightarrow \infty$,

$$\begin{aligned} P(X_1 > \alpha_0 + x) &= \exp(-[x^2 - x \log x \\ &\quad - (B + \log(2/A) - 1)x]/A + o(x)) \end{aligned} \quad (18)$$

and (11), respectively. Both processes $(X_n)_{n \in \mathbb{Z}}$ and $(Y_n)_{n \in \mathbb{Z}}$ have extremal index 1, and the extreme value distribution G appearing in equation (4) and (5) is the Gumbel distribution.

Observe that for an EGARCH(p, q) process as in equation (15) with $\alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_q \geq 0$ and such that $\sum_{j=1}^q \beta_j < 1$, the coefficients $(c_j)_{j \in \mathbb{N}}$

in equation (14) are automatically nonnegative and decay exponentially, so that Theorem 2 applies. In particular, the EGARCH process with normal innovations cannot cluster. Extensions to other light-tailed innovations ε such as GED(ν) distributions with $\nu > 1$ are possible; cf [10].

GARCH(1,1) Model

In the GARCH(1,1) model (see **GARCH Models**) of Engle [12] and Bollerslev [4] the log price increments $(Y_n)_{n \in \mathbb{Z}}$ and the volatilities $(\sigma_n)_{n \in \mathbb{Z}}$ are given by

$$Y_n = \sigma_n \varepsilon_n, \quad \sigma_n^2 = \gamma + \alpha Y_{n-1}^2 + \beta \sigma_{n-1}^2, \quad n \in \mathbb{Z} \quad (19)$$

where $(\varepsilon_n)_{n \in \mathbb{Z}}$ are i.i.d. and $\alpha, \beta, \gamma > 0$. Rewriting

$$\sigma_{n+1}^2 = \gamma + (\alpha \varepsilon_n^2 + \beta) \sigma_n^2, \quad n \in \mathbb{Z} \quad (20)$$

it becomes clear that a stationary and causal solution exists if and only if $E \log(\alpha \varepsilon_1^2 + \beta) < 0$. The following result on the tail behavior is included in Kesten's seminal work; see [8] for further results and references. The calculation of the extremal index can be found in [19, 25].

Theorem 3 (Tail Behavior and Extremes of GARCH(1,1)). *Assume the GARCH(1,1) model such that ε_1 is symmetric and has a positive density on $\mathbb{R}$ such that $E(|\varepsilon_1|^h) < \infty$ for $h < h_0$ and $E(|\varepsilon_1|^{h_0}) = \infty$ for $h \geq h_0$ for some $h_0 \in (0, \infty]$. Then there exist unique $\kappa > 0$ and $c > 0$ such that*

$$E(\alpha \varepsilon_1^2 + \beta)^{\kappa/2} = 1 \quad (21)$$

and the stationary distributions have tails for $x \rightarrow \infty$

$$\begin{aligned} P(\sigma_1 > x) &\sim c x^{-\kappa} \quad \text{and} \\ P(Y_1 > x) &\sim \frac{1}{c} E(|\varepsilon_1|^\kappa) x^{-\kappa} \end{aligned} \quad (22)$$

The extremal indices of $(\sigma_n)_{n \in \mathbb{Z}}$ and $(|Y_n|)_{n \in \mathbb{Z}}$ are given by

$$\theta_\sigma = \int_1^\infty P \left(\sup_{n \geq 1} \prod_{j=1}^n (\alpha \varepsilon_j^2 + \beta) \leq y^{-1} \right)$$

$$\begin{aligned} &\times \frac{\kappa}{2} y^{-(\kappa/2)-1} dy \in (0, 1) \quad \text{and} \\ \theta_{|Y|} &= \frac{E \left(|\varepsilon_1|^\kappa - \sup_{m \geq 1} |\varepsilon_{m+1}|^\kappa \prod_{j=1}^m (\alpha \varepsilon_j^2 + \beta)^{\kappa/2} \right)^+}{E |\varepsilon_1|^\kappa} \\ &\in (0, 1) \end{aligned} \quad (23)$$

respectively. Also the extremal index $\theta_Y \in (0, 1)$, and for all three sequences σ , $|Y|$ and Y , the extreme value distribution G appearing in equations (4) and (5) is the Fréchet distribution Φ_κ .

We conclude that the GARCH(1,1) process is able to model clusters in the extremes. Extensions to higher order GARCH processes are given in [3, 8].

Continuous-time Models

While for discrete-time volatility models many results on the extremal behavior are formulated for both the volatility process and the log price increments process, with few exceptions, most of the literature on extremes for continuous-time models concentrates on the volatility process. Hence, in the following, we often state results concerning the volatility process only.

The Wiggins Model

In the volatility model of Wiggins [30] (see also [27]) the log volatility is modeled as a Gaussian Ornstein–Uhlenbeck (OU) process. More precisely, the log price increments Y_t and the volatility σ_t are given by

$$\begin{aligned} Y_t &= \int_{(t-1, t)} \sigma_s_- dB_s, \\ d \log \sigma_t^2 &= (b_1 - b_2 \log \sigma_t^2) dt + \delta dW_t, \\ t &\in \mathbb{R} \end{aligned} \quad (24)$$

with two independent standard Brownian motions B and W , and real constants b_1 , b_2 and $\delta \neq 0$. The volatility has a stationary solution if and only if $b_2 > 0$, in which case it is given by

$$\log \sigma_t^2 = \int_{-\infty}^t e^{-b_2(t-s)} (b_1 ds + \delta dW_s), \quad t \in \mathbb{R}$$

(25)

Sampling the log volatility at integer points results in a causal Gaussian AR(1) process, so that the Euler type approximation $\bar{Y}_n := \sigma_{n-1}(B_n - B_{n-1})$ to equation (24) is the discrete-time volatility model (8) of Taylor.

From the above representation, it is clear that $\log \sigma_t^2$ is $N(b_1/b_2, \delta^2/(2b_2))$ distributed. The extremal index function of $\log \sigma^2$ and hence σ^2 follows from results in [24] as shown in [13].

Theorem 4 (Extremal Index Function of the Volatility). *Under the assumptions above, the extremal index function $\theta_\sigma(h)$ for $h \in (0, \infty)$ of the stationary volatility process σ^2 in (24) is identical to 1.*

We conclude that the volatility in the model (24) does not allow for extremal clusters. This continues to hold if the Gaussian OU process for the log volatility in equation (24) is replaced by any Gaussian process with continuous sample paths satisfying equation (12). While it is easy to show that the stationary log price increment Y_1 in equation (24) is distributed as $\left(\int_0^1 \sigma_t^2 dt\right)^{1/2} \varepsilon_1$ with ε_1 standard normally distributed and independent of $(\sigma_t)_{t \in \mathbb{R}}$, we are not aware of any explicit expressions for the tail behavior and the extremal index function of Y_1 or $\log Y_1^2$ as in Theorem 1(a).

Barndorff-Nielsen and Shephard (BNS) Model

In [1, 2], Barndorff-Nielsen and Shephard (BNS) model (see **Barndorff-Nielsen and Shephard (BNS) Models**) the volatility process as a Lévy-driven OU process, which results in the model

$$\begin{aligned} Y_t &= \int_{(t-1, t)} \sigma_{s-} dB_s, \\ \sigma_t^2 &= \int_{-\infty}^t e^{-\lambda(t-s)} dL_{\lambda s}, \quad t \in \mathbb{R} \end{aligned} \quad (26)$$

where B is Brownian motion, $\lambda > 0$ and L is a Lévy process with increasing sample paths (i.e., a subordinator), independent of B . The volatility process is stationary and satisfies the stochastic differential equation (SDE) $d\sigma_t^2 = -\lambda\sigma_t^2 dt + dL_{\lambda t}$.

The extremal behavior of this model depends on the driving Lévy process and has been analyzed in

[13–15]. For regularly varying noise, as shown in [14], one obtains the following.

Theorem 5 (Tail and Extremes for Noise in $\mathcal{R}(-\alpha)$). *Consider the stationary BNS-model (26) and assume that $L_1 \in \mathcal{R}(-\alpha)$ with $\alpha > 0$. Then $\sigma_1 \in \mathcal{R}(-2\alpha)$, $Y_1 \in \mathcal{R}(-2\alpha)$ and we have for $x \rightarrow \infty$*

$$\begin{aligned} P(\sigma_1^2 > x) &\sim \alpha^{-1} P(L_1 > x) \\ P(Y_1^2 > x) &\sim E(|\varepsilon_1|^{2\alpha}) \left(\frac{(1 - e^{-\lambda})^\alpha}{\alpha \lambda^\alpha} \right. \\ &\quad \left. + \frac{1}{\lambda^\alpha} \int_0^1 (1 - e^{s-\lambda})^\alpha ds \right) P(L_1 > x) \\ P(Y_1 > x) &= \frac{1}{2} P(Y_1^2 > x^2) \end{aligned} \quad (27)$$

respectively, where ε_1 is a standard normal random variable. The extremal index function θ_σ of the volatility process is furthermore given by

$$\theta_\sigma(h) = (h\alpha\lambda)(h\alpha\lambda + 1)^{-1}, \quad h > 0 \quad (28)$$

Observe that for a regularly varying noise process, the tail of σ^2 is of the same order as that of the driving noise process. Also, since $\theta_\sigma(h) < 1$, the process exhibits cluster possibilities. This is in contrast to the case when L has exponential tail as in the next theorem; see [13, 18]:

Theorem 6 (Tail and Extremes for Exponential-type Noise). *Consider the stationary BNS-model (26) as above and assume that L_1 has an exponential-type distribution tail:*

$$\begin{aligned} P(L_1 > x) &= c(x) \exp \left\{ - \int_0^x (a(y))^{-1} dy \right\}, \quad x > 0 \end{aligned} \quad (29)$$

where $\lim_{x \rightarrow \infty} c(x) = c > 0$ and $a > 0$ is absolutely continuous with $\lim_{x \rightarrow \infty} a(x) = \gamma^{-1}$ and $\lim_{x \rightarrow \infty} a'(x) = 0$, where $\gamma \in [0, \infty)$. Assume also that $L_1 \in \mathcal{S}(\gamma)$, that is, $P(L_2 > x) \sim E(e^{\gamma L_1}) P(L_1 > x)$ as $x \rightarrow \infty$ with $E(e^{\gamma L_1}) < \infty$. Then

$$P(\sigma_1^2 > x) \sim \frac{a(x)}{x} \frac{E e^{\gamma \sigma_1^2}}{E e^{\gamma L_1}} P(L_1 > x), \quad x \rightarrow \infty \quad (30)$$

In particular, $P(\sigma_1^2 > x) = o(P(L_1 > x))$ for $x \rightarrow \infty$. The extremal index function θ_σ is equal to 1, that is, $\theta_\sigma(h) = 1$ for all $h > 0$.

Brockwell [6] suggests to model the volatility in equation (26) by Lévy-driven continuous-time ARMA (CARMA) processes, with the CAR(1) process being the OU process. As shown in [13–15], for CARMA processes driven by regularly varying noise processes, clusters occur as in Theorem 5, while for driving Lévy processes as described in Theorem 6, CARMA processes may or may not model clusters, depending on the corresponding kernel function. Todorov and Tauchen [29] suggest to model the volatility by a CARMA(2,1) process with a mixture of gamma distributions as driving noise process. For this model the results presented here do not apply.

Continuous-time GARCH(1,1) Models

As a diffusion limit of GARCH(1,1) processes, Nelson [26] obtained

$$\begin{aligned} Y_t &= \int_{(t-1,t)} \sigma_{s-} dB_s, \\ d\sigma_t^2 &= (\beta - \varphi\sigma_t^2) dt + \lambda\sigma_t^2 dW_t, \quad t \geq 1 \end{aligned} \quad (31)$$

where B and W are independent standard Brownian motions and $\beta \geq 0, \lambda > 0$ and $\varphi \in \mathbb{R}$ are parameters. It has a strictly stationary solution if and only if $2\varphi/\lambda^2 > -1$ and $\beta > 0$, in which case the marginal distribution is inverse gamma. The two independent driving processes in equation (31) are in contrast to the situation for discrete-time GARCH processes, where price and volatility are both driven by the same noise sequence $(\varepsilon_n)_{n \in \mathbb{Z}}$. Inspired by this, Klüppelberg *et al.* [23] constructed another continuous-time GARCH model, termed COGARCH(1,1), which meets the features of discrete-time GARCH better and for which the volatility jumps, unlike for the diffusion limit (31). Let $(L_t)_{t \geq 0}$ be a Lévy process with nonzero Lévy measure and $\eta, \varphi, \beta > 0$ be parameters. Defining the auxiliary Lévy process

$$R_t = \eta t - \sum_{0 < s \leq t} \log(1 + \varphi(\Delta L_s)^2), \quad t \geq 0 \quad (32)$$

the log price increments Y and the volatility σ are given by

$$\begin{aligned} Y_t &= \int_{(t-1,t)} \sigma_{s-} dL_s, \\ \sigma_t^2 &= \left(\beta \int_0^t e^{R_{s-}} ds + \sigma_0^2 \right) e^{-R_t}, \quad t \geq 1 \end{aligned} \quad (33)$$

where σ_0^2 is independent of L . A sufficient condition for strict stationarity of (33) is the existence of some $\kappa > 0$ such that

$$|L_1|^\kappa \log^+ |L_1| < \infty \quad \text{and} \quad E(e^{-R_1 \kappa/2}) = 1 \quad (34)$$

Observe that the volatility in Nelson's diffusion limit (31) has also a solution (33), with R_t defined by

$$R_t := (\varphi + \lambda^2/2)t - \lambda W_t, \quad t \geq 0 \quad (35)$$

For the stationary choice, we have

$$E(e^{-R_1 \kappa/2}) = 1 \quad \text{with} \quad \kappa := 2 + 4\varphi/\lambda^2 > 0 \quad (36)$$

The following result is from [16, 18]:

Theorem 7 (Tail and Extremes of Continuous-time GARCH). *Consider the stationary diffusion limit (31) or COGARCH(1,1) process as above with $\kappa > 0$ as given by equations (36) or (34), respectively. Then there exists a constant $c > 0$ such that*

$$P(\sigma_1 > x) \sim cx^{-\kappa}, \quad x \rightarrow \infty \quad (37)$$

In the case of the COGARCH(1,1) process, assume further that there is $d > \max\{1, \kappa\}$ such that $E|L_1|^{2d} < \infty$ with $\kappa > 0$ as defined in equation (34) and that L is not the negative of a subordinator. Denote $M_t := B_t$ for the diffusion limit (31) and $M_t := L_t$ for the COGARCH(1,1) process. Then

$$\begin{aligned} P(Y_1 > x) &\sim E \left[\left(\int_0^1 e^{-R_{t-}/2} dM_t \right)^+ \right]^\kappa \\ P(\sigma_1 > x), \quad x &\rightarrow \infty \end{aligned} \quad (38)$$

and σ has extremal index function

$$\theta_\sigma(h) = \frac{E \left(\sup_{0 \leq t \leq h} e^{-R_t \kappa/2} - \sup_{t \geq h} e^{-R_t \kappa/2} \right)^+}{E \left(\sup_{0 \leq t \leq h} e^{-R_t \kappa/2} \right)}$$

$$< 1, \quad h > 0 \quad (39)$$

The extremal index of the discrete-time process $(Y_n)_{n \in \mathbb{N}}$ of the log price increments at integer times is given by

$$\theta = \frac{E \left[(I_1^k - \max_{k \geq 2} I_k^k)^+ \right]}{E[I_1^k]} < 1 \quad (40)$$

It follows that both the diffusion limit and the COGARCH(1,1) can model extremal clusters. Since the diffusion limit (31) has continuous sample paths, one can also consider its clustering behavior via epsilon-upcrossings. Choosing such an approach, the diffusion limit of Nelson does not cluster, as reported in [18]. In particular, both notions of extremal clustering for processes with continuous sample paths do not lead to the same interpretation.

Similar to the definition of COGARCH, a continuous-time analog to the EGARCH process has been proposed in [21]. So far, no analog to Theorem 2 for the continuous-time EGARCH process seems to be available.

References

- [1] Barndorff-Nielsen, O.E. & Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics (with discussion), *Journal of the Royal Statistical Society. Series B* **63**, 167–241.
- [2] Barndorff-Nielsen, O.E. & Shephard, N. (2002). Econometric analysis of realised volatility and its use in estimating stochastic volatility models, *Journal of the Royal Statistical Society. Series B* **64**, 253–280.
- [3] Basrak, B., Davis, R.A. & Mikosch, T. (2002). Regular variation of GARCH processes, *Stochastic Process and their Applications* **99**, 95–116.
- [4] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**, 307–327.
- [5] Breidt, F.J. & Davis, R.A. (1998). Extremes of stochastic volatility models, *Annals of Applied Probability* **8**, 664–675.
- [6] Brockwell, P.J. (2004). Representations of continuous time ARMA processes, *Journal of Applied Probability* **41A**, 375–382.
- [7] Davis, R.A. & Mikosch, T. (2008). Extremes of stochastic volatility models, in *Handbook of Financial Time Series*, T.G. Andersen, R.A. Davis, J.-P. Kreiss & T. Mikosch, eds, Springer, Heidelberg, 355–364.
- [8] Davis, R.A. & Mikosch, T. (2008). Extreme value theory for GARCH processes, in *Handbook of Financial Time Series*, T.G. Andersen, R.A. Davis, J.-P. Kreiss & T. Mikosch, eds, Springer, Heidelberg, 187–199.
- [9] Drude, N. (2006). *Extremwertverhalten von unendlichen Moving Average Prozessen mit leicht taillierten Innovationen und Anwendungen auf EGARCH Prozesse*. Diplomarbeit, Technische Universität München.
- [10] Drude, N. & Lindner, A. (2008). *Extremes of Sums of Infinite Moving Average Processes with Light Tails and Applications to EGARCH*, in preparation.
- [11] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.
- [12] Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica* **50**, 987–1008.
- [13] Fasen, V.M. (2004). *Extremes of Lévy Driven Moving Average Processes with Applications in Finance*. PhD thesis, Technische Universität München.
- [14] Fasen, V. (2005). Extremes of regularly varying mixed moving average processes, *Advances in Applied Probability* **37**, 993–1014.
- [15] Fasen, V. (2009). *Extremes of Lévy Driven Mixed MA Processes with Convolution Equivalent Distributions* **12**(3) 265–296.
- [16] Fasen, V. (2008). *Asymptotic Results for Sample Autocovariance Functions and Extremes of Integrated Generalized Ornstein-Uhlenbeck Processes*, to appear.
- [17] Fasen, V. (2008). Extremes of continuous-time processes, in *Handbook of Financial Time Series*, T.G. Andersen, R.A. Davis, J.-P. Kreiss & T. Mikosch, eds, Springer, Heidelberg, 653–667.
- [18] Fasen, V., Klüppelberg, C. & Lindner, A. (2006). Extremal behavior of stochastic volatility models, in *Stochastic Finance*, M.D.R. Grossinho, A.N. Shiryaev, M. Esquivel & P.E. Oliveira, eds, Springer, New York, pp. 107–155.
- [19] Fasen, V., Klüppelberg, C. & Schlather, M. (2007). *High-level Dependence in Time Series Models*, to appear.
- [20] Finkenstädt, B. & Rootzén, H. (eds) (2004). *Extreme Values in Finance, Telecommunications, and the Environment*, Chapman & Hall/CRC, Boca Raton.
- [21] Haug, S. & Czado, C. (2007). An exponential continuous time GARCH process, *Journal of Applied Probability* **44**, 960–976.
- [22] Klüppelberg, C. & Lindner, A. (2005). Extreme value theory for moving average processes with light-tailed innovations, *Bernoulli* **11**, 381–410.
- [23] Klüppelberg, C., Lindner, A. & Maller, R. (2004). A continuous time GARCH process driven by a Lévy process: stationarity and second order behaviour, *Journal of Applied Probability* **41**, 601–622.
- [24] Leadbetter, M.R., Lindgren, G. & Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer, Berlin.
- [25] Mikosch, T. & Stărică, C. (2000). Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process, *Annals of Statistics* **28**, 1427–1451.

- [26] Nelson, D.B. (1991). Conditional heteroskedasticity in asset returns: a new approach, *Econometrica* **59**, 347–370.
- [27] Scott, L.O. (1987). Option pricing when the variance changes randomly: theory, estimation and application, *Journal of Financial Quantitative Analysis* **22**, 419–439.
- [28] Taylor, S.J. (1982). Financial returns modelled by the product of two stochastic processes – a study of daily sugar prices 1961–79, in *Time Series Analysis: Theory and Practice I*, O.D. Anderson, ed., North-Holland, Amsterdam, pp. 203–226.
- [29] Todorov, V. & Tauchen, G. (2006). Simulation methods for Lévy-driven CARMA stochastic volatility models, *Journal of Business and Economic Statistics* **24**, 455–469.
- [30] Wiggins, J.B. (1987). Option values under stochastic volatility: theory and empirical estimates, *Journal of Financial Economics* **19**, 351–372.

Related Articles

Autoregressive Moving Average (ARMA) Processes; Barndorff-Nielsen and Shephard (BNS) Models; Bates Model; Extreme Value Theory; GARCH Models; Heavy Tails; Risk Measures; Statistical Estimation; Stochastic Volatility Models; Stylized Properties of Asset Returns.

CLAUDIA KLÜPPELBERG & A. LINDNER

Securitization

Securitization is the process by which tradable securities are created and introduced into capital markets. Such a process generally involves, typically, the isolation of a pool of assets and their repackaging into securities. Securitization enables the originator of such a transaction to manage and diversify its risk, to optimize the use of capital, while investors in these new securities are looking for diversification and higher return. Various types of risks have been securitized and different structures have been issued in the capital markets.

Asset-backed Securities

The best-known example of securitization is certainly that of mortgages, with the first issue of mortgage-backed securities (MBS) taking place in 1970 in the United States. This transaction was the first of a very long series and the beginning of an extremely successful market. When the collateral of a securitization is made with more general assets than simple mortgages, the resulting securities are called *asset-backed securities (ABS)*. The collateral may be, for instance, credit card receivables or student loans.

As of January 1, 2006 [1], the estimated outstanding in the United States is \$8 trillion, which represents the largest segment of the American-fixed income market—31% of the total outstanding bond market (23% for the MBS and 8% for the ABS)—while the T-bond market only counts for 16%. The US market is the largest market in terms of securitization of mortgages in the world (80% of the total market). In Europe and Asia, the situation is slightly different since the legal framework for a securitization process to take place has been missing for a long time: new rules were adopted in the United Kingdom in 1987, in France in 1988, . . . , and in Italy only in 1999. In Asia, the first Indian ABS took place in 1992 (the first MBS in 2000), the first MBS took place in 1998 in Japan (Japan is the largest Asian market), 2005 in China, and 2003 in Taiwan.

Securitization of mortgages in the United States often require the participation of one of the three government sponsored enterprises (GSEs), the Federal National Mortgage Association (FNMA or Fannie

Mae), the Government National Mortgage Association (GNMA or Ginnie Mae), and the Federal Home Loan Mortgage Corporation (FHLMC or Freddie Mac). The securities backed by one of the GSEs are called *agency ABS*. Because of their government sponsorship (in particular that of the US Treasury), each of the GSEs could perfectly act as credit enhancer. Since most of the MBS are issued by these agencies, they implicitly have the full backing of the US treasury and therefore their credit risk is almost inexistent.

There are also ABS issued without the sponsorship of one of the GSEs. They are called *nonagency ABS*. In this case, the collateral is not standard. It is typically made of jumbo loans (i.e., huge mortgages with a rather good credit quality), Alt-A loans (i.e., loans where a borrower could not provide complete documentation of his assets or income), subprime loans (i.e., loans to borrowers with heavy credit burden and poor credit quality), or commercial loans (i.e., loans on retail, office, or industrial properties). Until the recent crisis of 2007, the subprime market was the largest part of the nonagency ABS. Several factors can explain their importance; in particular, the growing number of homeowners with heavy debt burden, the aggressiveness of the marketing of subprime issuers, and the lending practices in this market. The characteristics of the ABS with this nonstandard collateral are very different and make them attractive to investors: not only do they offer a higher yield to compensate for the higher credit risk but also the prepayment risk is typically very limited. Therefore, they offer an alternative with a different risk profile.

There are two main types of structures of ABS. The first type—*pass-through structures*—is the simplest. It corresponds to a securitization based directly on the collateral. The securities give a particular ownership right to the pool of *assets*. In this case, the cash flows of interest and principal received by the SPV are directly “passed through” to the investors. The investor receives a fraction of the payments on the pro rata of the number of shares he or she owns.

The second type—*pay-through structures*—is more complex. The securities, such as collateralized mortgage obligations (CMOs), are not directly based on the assets ceded by the originator and the securitization process requires some structuring. The SPV buys the collateral from the originator and then issues a variety of instruments, which divide

the collateral cash flows into different classes or *tranches*. Such a structure enables to issue securities with different characteristics compared to the original assets; in particular, the structure and timing of payments, the associated risk can be different.

Motivations

Legal, regulatory, and accounting rules applicable to securitization transactions differ widely among countries and are subject to the ongoing modification and revision. It is necessary in structuring a securitization transaction to deal with legal, regulatory, and tax rules that may affect the sale and conveyance of the assets from the originator to the SPV. Another important issue relates to the legal framework governing the creation, maintenance, and operation of special purposes vehicles. Finally, depending on the originator's objectives, the balance sheet effects and accounting treatment and consequences of a particular securitization are essential and will frequently influence the ultimate structure of the transaction. In particular, the question of regulatory capital, or in other words, the amount of provisions to be put aside by financial institutions to be authorized by the regulators to have an activity, is essential and has a huge impact on the structure of transactions, as securitization can be used for capital relief to reduce regulatory capital allocations. In the case of insurance-linked securitization, things can be even more complex.

Advantages and Limitations

Securitization offers several advantages to the originator. Among them, one can mention the transfer of risk, the reduction in the asset-liability mismatch and lower capital requirements by removing risk from the balance sheet (improvement of various financial ratios, more efficient use of capital, and compliance with capital standards). It is also a way to lock-in profits by cashing some money today and removing the uncertainty associated with the future cash flows. It can appear as a lower cost source of financing. Obviously, such a process is not for everyone since it is very costly: securitizations are expensive and take time. The more complicated and unusual the transaction is, the more expensive and time consuming it will be. There is also a critical mass to achieve. Only large originators or pools of originators can think of

securitization as a cost-effective way of transferring risks.

As far as investors are concerned, ABS offer an opportunity to invest in a specific pool of assets, with a high credit quality and a relatively large rate of return. This can enhance the diversification of their existing portfolios. There are, however, some risks associated with these financial products: liquidity risk, credit risk, event risk, and prepayment risk appear to be the main sources of uncertainties to which the investors are exposed when buying these products.

Globally, many public, social, and economic benefits can result from securitization. For instance, the existence of a liquid and efficient secondary securitization market increases the availability and reduces the cost of financing for the primary market lenders. It can also reduce geographical and regional disparities in the availability and cost of borrowing by linking local credit activities to global capital markets. Also, securitization can be used as a way to improve capital allocation and reduce risks to individual institutions and systemic risks within the financial system.

General Structure and Key Agents Involved in a Securitization

A typical securitization begins with one (or several) originator having some assets in its book that it wants to divest. The securitization process itself requires several steps, as illustrated in Figure 1. Various agents are involved at different stages of any securitization process.

Issuer and Investors

Even if ABS have been issued on a basis of a large variety of different underlying risks, their overall structure is rather generic. Everything starts with an originator. It initiates the contracts that will be securitized at the end. In other words, the originator provides a product to a customer, who agrees to make a series of payment over a certain period of time. The present value of these payments (capital and interests) constitutes an asset of the originator. Instead of keeping it in its balance sheet, the originator is now able to move it off balance sheet through the securitization process.

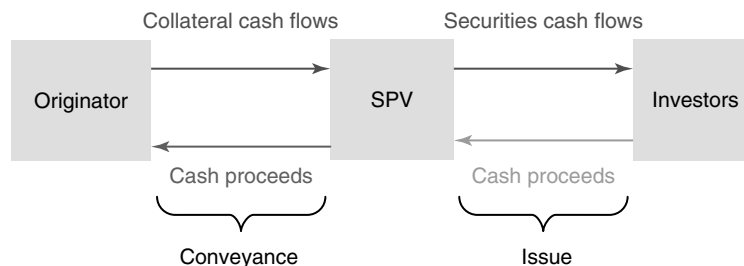


Figure 1 General structure of a securitization process

Special Purpose Vehicle (SPV)

More practically, to issue these new securities, a special purpose vehicle (SPV) is created. This is a passive financial entity, bankruptcy remote, that exists for the sole purpose of purchasing and housing the assets from the originator and issuing securities using the assets as *collateral*. The collateral consists of assets provided to secure an obligation. The name of the securities itself comes from the nature of the collateral: as already mentioned, when the assets are mortgage loans, the securities are collateralized or backed by mortgage loans and so are called *MBS*. When it is a general asset that backs the securities, they are called *ABS*.

Once the assets have been transferred from the originator to the SPV, a process called *conveyance*, the SPV is responsible for managing the collateral (asset) and administrating its different cash flows. It is also responsible for designing the securities to be issued. Since the management of the cash flows is passive and tax-free, the whole cash flows are transferred to the investors. Note that, in general, the conveyance of assets from the originator to the SPV has to be conducted in a manner that results in a *true sale* (i.e., legal isolation), in order to remove the assets from the bankruptcy or insolvency of the originator. In particular, if the originator was to become bankrupt or insolvent, a trustee would not be able to reach the transferred assets to satisfy claims of creditors of the originator.

In summary, after the transfer of the collateral asset to the SPV, the latter issues securities to investors, who, in exchange, pay some money to the SPV in order to buy the securities. Then, the SPV pays all the proceeds from the securities issue to the originator, while the latter will pay the different cash

flows coming from the asset during its life to the SPV so that they can be transferred to the investors.

Cash and Synthetic Structures

In the securitization process described here, the collateral is made of assets from the originator's balance sheet. Some more sophisticated securitization processes also exist without this particular feature, especially in credit risk securitization. The issued securities are usually called *synthetic* (versus cash) since the collateral is recreated synthetically.

A standard example is the (cash) collateralized debt obligations (CDOs), based upon a collateral that is a pool of assets, usually comprising loans and debt instruments (but generally nonmortgage assets). The collateral pool can be made of various assets: investment-grade bonds, high-yield bonds, sovereign debt, bank loans, or even tranches of other CDOs. More precisely, a CDO consists of the issue of several notes, also called *tranches*, having different characteristics (maturity, seniority, etc.). The credit risk of the collateral is borne by the investors of the CDO, but the exposure is different according to the tranche. However, CDOs can also be synthetic (funded or unfunded). In this case, the underlying risk is the credit risk associated with a portfolio of credit default swaps signed by the originator and the SPV. In the funded case, the collateral is constituted by the purchase by the SPV of high credit quality financial instruments such as T-bonds with the initial proceeds from the investors. In the unfunded case, there is no collateral as such and the CDO is therefore similar to a swap contract.

There is an important difference between the cash and synthetic structures coming from the motivation behind the securitization itself: in a standard (cash) securitization, the originator is really at the origins of

the transaction, with the idea of removing some assets and risks from its balance sheet and using the capital in an optimal way. In a synthetic case, the motivations may come more naturally from the investors' side, and the securitization is an answer to the demand for specific products.

Agents Involved before the Transaction

Before the transaction, in addition to the advisor, typically an investment bank designing the overall transaction, and the originator, many other types of agents take part in the structuring of the transaction.

Advisors and Modeling Firms. In a standard securitization process, the originator works with an advisor, typically an investment bank. The advisor is here to design the transaction in its every aspect, even the underwriting. It cannot, however, retain any risk. Its role is that of a pure intermediary, making the overall process feasible. This comes at a cost, the fees being typically 1% of the notional amount of the transaction.

In some cases, especially when the underlying risks are not standard (this is usually the case for insurance-linked risks such as natural catastrophe or life), a modeling firm is called upon to help with the pooling of the risk to be securitized and the valuation of those risks.

Rating Agencies. The risk of any transaction has to be assessed before it can go through. Rating agencies are firms specialized in the risk assessment. The largest are Moodys, Standard and Poors, and Fitch Ratings. Their role is to provide not only an independent, timely, and prospective opinion on the credit worthiness but also the structural integrity and other attributes of the transaction (and also, more generally, an entity or even a government).

As to achieve the desired credit risk profile for the securities, depending upon the nature of the transaction and the assets involved, the asset pool may need to be supported by some types of credit support, called *credit enhancement*. Credit enhancement mechanisms can take various forms, both from external parties and within the structure of the deal. Internal credit enhancement mechanisms are a form of self-insurance, whereas external credit enhancement mechanisms are third party-guarantees and typically involve the participation of a bond insurer or monoline.

Before the issue, rating agencies have to give an assessment of the risk taken by the investor when buying this particular product, that is, risk of not only losing their investment due to a credit event but also any triggering event underlying the transaction (natural catastrophe in the case of a cat-bond, for instance). Once issued, the rating agencies usually (but not always) monitor the performance of the transaction and adjust the rating accordingly. Ratings are assigned and reviewed by a committee process. The rating will be based upon the structure of the transaction, the quality of the various parties, and the risks underlying the new securities. Obviously, this process involves the knowledge of the models available for these particular risks. Rating agencies may therefore work closely with modeling firms when the underlying is not standard. Various credit enhancement strategies can be used to improve the credit quality of a securitization. The rating system generates many problems as revealed by the credit crisis of 2007–2008. In particular, the use of the same rating system for entities and transactions leads to a confusion among investors. While an AAA entity represents a default risk-free entity (this is the case for the G7 countries), an AAA transaction (or tranche) is not exempt from risk. The rating of a structure should also include some other forms of risks (liquidity or market value risks, for instance, which are very important especially for highly sophisticated structures).

Lawyers

The next step before receiving the approval from the regulatory bodies is to write the legal agreement related to the deal. Even if most advisors have an internal legal department, most of them use an independent law firm to write down the contract, as to avoid any responsibility from the legal point of view. The task of these law firms can be very tricky as it requires the knowledge of various legal frameworks depending on the country, or even the state in the United States.

Regulatory Authorities

Finally, any transaction, before going to the financial markets, has to be approved by the regulatory authority of the relevant country or even the relevant state in the United States. In the United Kingdom, it is the

role of the Financial Services Authority (FSA) and in the United States that of the Securities and Exchange Commission (SEC). Both are independent nondepartmental public bodies regulating the financial services.

Insurance-linked Securitization

The convergence of insurance with the capital markets has not only opened up an alternative channel for insurers to transfer risk, raise capital, and optimize their regulatory reserves but also offered institutions a source of potentially liquid and uncorrelated investment exposure. One of the financial instruments allowing for the cession of insurance-related risks to the capital markets is called *insurance-linked securities (ILS)* [2]. Growth in the ILS market has been steady over the past decade, with a number of different structures, risk layers, and trigger types being introduced. When discussing ILS, one needs to take into account the different structural features for high-frequency–low-severity *versus* low-frequency–high-severity transactions in the nonlife sector, or for

regulatory capital, longevity, and extreme mortality trades in the life sector. Motivations also differ according to the structure and the cedant/insured. Catastrophic risk is not only concerned by the securitization process. In theory, any insurance risk, on which policies can be written, can be subject to securitization, and in particular, life insurance risk.

References

- [1] Euromoney (2008). *Global Securitisation Review 2007/08*, Euromoney Yearbook.
- [2] Swiss Re (2006). *Sigma 07/2006, Securitization—New Opportunities for Insurers and Investors*, Swiss Reinsurance Company, Zürich.

Related Articles

ABS Indices; Collateralized Debt Obligations (CDO); Structured Finance Rating Methodologies.

PAULINE BARRIEU

ABS Indices

Asset back securities (ABS) are financial debt securities based on a pool of underlying assets or collateral assets. When the assets are backed by mortgages, they are called *mortgage backed securities* (MBS). ABSs are backed by nonmortgage assets. This includes auto loans, credit card receivables, home equity loans, student loans, and so on. Owing to government guarantees, MBS typically entail no credit risk. However, ABSs generally lack such guarantees, so they entail credit risk. It might be for this reason that ABS backed by subprime mortgages are still called *ABS*, or more explicitly, *subprime residential mortgage-backed security* (RMBS) instead of MBS.

The basic building blocks of credit derivatives are credit default swaps (CDS). It was the standardization of the settlement terms and conditions of the corporate CDS contracts by International Swaps and Derivatives Association (ISDA) in early 1990s that allowed the corporate CDS market to grow and evolve quickly. The credit risk in subprime RMBS differs in fundamental ways from credit risk in corporate bonds. ISDA published its first pay-as-you-go (PAUG) CDS documentation for RMBS in June 2005, and followed up with amended versions in April 2006 and November 2006. This ISDA documentation standardization for PAUG CDS on subprime RMBS helped create a synthetic CDS market on the subprime RMBS bonds. In late 2005, 15 security firms, representing major brokers and dealers, along with an index company CDS Index Co. and a data provider, Mark-It Partner came up with the ABS CDS benchmark index, ABX.HE, for subprime RMBS, and the trading of ABX started on January 19, 2006. The index is supposed to roll every six months with newly issued subprime RMBS bonds as collateral assets. Owing to the subprime crisis and a lack of newly issued subprime RMBS bonds, we have only four series of ABX indices up to now; they are simply called *ABX* 2006-1, *ABX* 2006-2, *ABX* 2007-1, and *ABX* 2007-2 based on the vintage year of underlying subprime RMBS bonds.

Tranched ABX or TABX is the benchmark index tranche product based on ABX. Owing to the relatively small number of collateral ABS bonds of each vintage ABX index, we use the underlying ABS bonds of two conjacent ABX indices as collateral assets for the TABX. Shortly after the introduction

of TABX, the subprime crisis emerged. In practice, TABX has never been liquidly traded. However, there are a large number of issuances of bespoke ABS collateralized debt obligations (CDOs) with non-ABX bonds as collateral assets.

We simply present the contract parts of the ABX and TABX and briefly discuss the modeling of TABX.

ABX

ABX.HE indices are CDS benchmark index for a standard basket of subprime home equity ABS bonds. For each vintage ABX.HE, we have five indices based upon the rating of reference ABS bonds: AAA, AA, A, BBB and BBB— by both Moody's and S&P. The selection of the included bonds is rule-based according to the deal size (at least 500 millions) of the subprime home equity ABS shelf program with a limit from the same loan originator (four deals) and from the same master servicer (to six deals), and the weighted average life (4–6 years from the issuance for other ratings except AAA, and more than 5 years for AAA). Each vintage ABX.HE has 20 underlying collateral bonds, and a new series of index is supposed to be issued every six months.

As stated early, ABX is a CDS spread index whose premium leg or fixed leg is based on two parts: up-front payment in terms of percentage and a running payment in terms of a spread in basis point. This running spread is called a *fixed rate*, which is determined at the beginning so that the initial quote for the index is 100 or we do not need to pay any up-front premium if you buy protection or short the index. The following shows the market quote for ABX on November 28, 2009. Owing to the subprime crisis, most of the current “price” is very low. This implies that we need to pay a very high up-front payment to buy protection. For example, if we want to buy default protection on 2006-01 BBB—, we need to pay $(100 - 5 \cdot 16/32)\% = 94.5\%$ up-front premium based on the notional amount and 267 bps (basis points) running spread per annum, but paid monthly, on the basis of the outstanding notional amount. The outstanding notional amount declines over time on the basis of the reference obligations amortization.

The loss leg includes interest shortfall, write-down, principal shortfall adjusted by interest rate shortfall reimbursement, principal shortfall reimbursement, and write-down reimbursement. The protection buyer receives any interest shortfall, but the

total amount is capped at a fixed rate. A credit event is defined as principal shortfall and write-down according to the 2005 ISDA PAUG template definition.

For example, if any one of the 20 referenced ABS bonds incurs a write-down of 1% in the amount of its current principal balance, and the referenced bond current factor is 70%, then the protection buyer of 100 millions receives the following payment:

$$\begin{aligned} & \text{Notional amount} \times \text{current factor} \times \text{weighting} \\ & \times \text{loss rate} = 100 \text{ millions} \times 0.7 \times 0.05 \\ & \times 0.1 = 35\,000 \end{aligned} \quad (1)$$

After this credit event, the notional amount based on which that fixed rate is paid would be reduced by \$35 000 until the earlier of the next credit event or scheduled termination. ABX has been actively traded since its inception on a daily basis. Table 1 provides a market quote for ABX on Nov 28, 2008.

TABX

TABX was introduced in February 2007 just before the subprime crisis emerged. TABX tranches reference a portfolio of 40 names from underlying bonds from 06-2 and 07-1 ABX.HE series and 07-1 and 07-2 ABX.HE series of a similar rating. Only two tranche baskets were traded initially: BBB and BBB−. The attachment and detachment points for the 06-2 and 07-1 portfolio and the fixed running spreads for all tranches are given in Table 2.

The tranche protection buyer receives protection against the write-downs and principal shortfalls on the reference ABS bonds. No interest shortfalls are

covered by the tranche protection. On the fixed side, all tranches have a specified or fixed running spreads in bps as given in Table 1, and are traded on price up front. Fixed running spreads are capped on roll date at 500 bps. The subordination level is the initial subordination notional amount less write-downs, principal shortfalls, and amortization. Therefore, the detachment and attachment points of the tranche can change over each period depending on the loss evolution and amortization. We could think of TABX, or, broadly speaking, of ABS CDOs as CDOs with changing subordination levels.

Modeling of ABX and TABX

The TABX market was a short-lived one. It was introduced in February 2007. Shortly after its introduction, we observed the first crash of the ABX indices. But there were billions of dollars of ABS CDOs structured and issued by Wall Street firms without ABX index bonds as collateral assets. There is no market consensus with respect to the ABS CDO valuation. Some practitioners tried to adapt the Gaussian copula model for ABS CDOs, which were hard to interpret since the results depended critically on the prepayment assumptions for the underlying subprime loans.

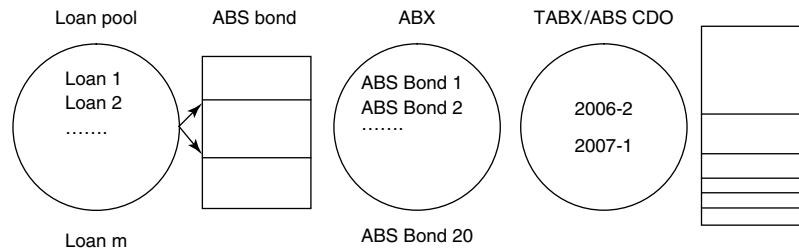
Figure 1 shows the whole securities involved in the TABX or ABS CDOs. The first layer of collateral assets for ABS bonds consists of individual mortgage loans underlying the ABS bonds. Each ABS bond takes about 5000–10 000 individual mortgage loans as collateral assets. Then ABX takes 20 ABS bonds as its underlying assets. The TABX takes 40 ABS bonds as collateral assets and goes through another tranche. The TABX is thus more like a CDO squared transaction. In the market, there is no consensus with

Table 1 Market quote for ABX, November 28, 2008

Series	Fixing rate	Rating	Quote	Series	Fixed rate	Rating	Quote
07-2	76	AAA	33-24/36-24	06-2	11	AAA	46-00/49-00
07-2	192	AA	6.00/8.00	06-2	17	AA	11-00/14-00
07-2	369	A	4.00/7.00	06-2	44	A	5-00/8-00
07-2	500	BBB	3.00/5.00	06-2	133	BBB	2-16/4-16
07-2	500	BBB−	3.00/5.00	06-2	242	BBB−	2-16/4-16
07-1	9	AAA	33-24/36-24	06-1	18	AAA	75-00/78-00
07-1	15	AA	4-16/6-16	06-1	32	AA	30-00/33-00
07-1	64	A	3.00/6-4-16	06-1	54	A	11-00/14-00
07-1	224	BBB	2-16/4-00	06-1	154	BBB	4-24/6-24
07-1	389	BBB−	2-16/4-00	06-1	267	BBB−	4-16/6-16

Table 2 Attachment and detachment point for TABX tranches (06-2 and 07-1 portfolio) and running spreads

Index pool	ABX rating	Attachment (%)	Detachment (%)	Fixed rate
06-2 and 07-1	BBB	0	3	500
06-2 and 07-1	BBB	3	7	500
06-2 and 07-1	BBB	7	12	500
06-2 and 07-1	BBB	12	20	467
06-2 and 07-1	BBB	20	35	200
06-2 and 07-1	BBB	35	100	51
06-2 and 07-1	BBB-	0	5	500
06-2 and 07-1	BBB-	5	10	500
06-2 and 07-1	BBB-	10	15	500
06-2 and 07-1	BBB-	15	25	500
06-2 and 07-1	BBB-	25	40	267
06-2 and 07-1	BBB-	40	100	72

**Figure 1** An overview of subprime RMBS products

respect to the CDO squared transaction even though attempts have been made to tackle this problem. One example is the approach in [1] in which the authors use a mixed Gaussian copula function along with the mapping of loss distributions. But this is only for corporate credits that have one risk or “decrement”. For subprime loans, each loan is subject to two decrements: default and prepayment.

We believe the only approach to capture all features and complexities of the TABX or ABS CDOs is a bottom-up approach. In this approach, we model the fundamental driving factors for the mortgage loan pool: prepayment and default. We then generate the cash flows for the ABS bonds using the waterfall structure for each collateral ABS bond (like a cash-flow CDO tranche). This can be accomplished by running Intex. These cash flows include the write-downs, principal, interest rate, and coupon cap shortfalls. We can then aggregate the cash flows for the whole underlying pool of the TABX. The TABX tranche can then be assessed by the aggregate cash flows for the collateral pool. The

dynamics and correlation have been introduced since there are common driving factors for prepayment and default, such as the home price appreciation (HPA) index. All other features are incorporated since we use the waterfall structure for each ABS bond. Calibration of the model to the observed ABX prices can be achieved if we use a few parameters for the prepayment and default curves.

Reference

- [1] Li, D.X. & Liang, M.H. (2005). *CDO Squared Pricing Using Gaussian Mixture Model with Transformation of Loss Distribution* at <http://ssrn.com/abstract=890766>.

Related Articles

CDO Square; Credit Default Swap (CDS) Indices; Securitization.

DAVID X. LI

Modigliani–Miller Theorem

The Modigliani–Miller theorem states that the value of a firm is invariant with respect to its leverage policy in an arbitrage-free market when there is no corporate income tax and no bankruptcy cost. Modigliani and Miller (M–M) [10] proved this “invariance” result using, for the first time, a “no-arbitrage” condition.

To present the M–M theorem, we consider a simple levered firm in which there are two basic claimants on the firm’s income:

- *debt holders*, whose security allows them to claim the coupon C at each time t as long as default is not declared;
- *equity holders* (the owners of the firm), whose security allows them, once debt holders have been paid, to claim the *residual cash flow* (if positive) as dividends.

Debt holders are paid *before* equity holders, but their claim does not allow them to receive more than the coupon, no matter what the net result is. On the contrary, dividends received by equity holders may be very high when the net result is very high but can also be very low (and even null) when the net result is very low. Equity holders are frequently called the *residual claimant* since they own the *residual income* (i.e., the net result) of the firm, that is, what remains when employees, debt holders, and government have been paid.

Let E^l and B be the market value of equity and debt, respectively, of a given firm at some point in time. To become the owner of this (levered) firm, an investor must buy not only the equity E^l but also the debt of the firm, so that the total payment is equal to $V^l = E^l + B$, where V^l is called the *value of the firm*. It is thus tempting to think that if we increase B , the value of the firm V^l will also increase. Finding the conditions under which debt affects the value of the firm is the central question of M–M analysis: given a *fixed* investment policy, what are the condition(s) under which debt does not affect the value of the firm?

Basic Statement of the Theorem

Consider a firm that has invested in some productive asset, which allows it to generate an operating income (i.e., the earning before interests and taxes (EBIT)) X_t at each time t . Roughly speaking, X_t is equal to the value of the sales minus production costs. When positive, X_t is used to pay debt holders (coupon), government (tax), and equity holders (dividends).

Unlevered Firm

In this case, the value of the firm at time t is denoted by v_t , and is simply equal to the net result, that is,

$$v_t = X_t(1 - T_a) \quad (1)$$

where T_a is the corporate tax income.

Let C be the coupon of the corporate debt and T be the maturity of the debt. In what follows, we not only consider the case of a *perpetual debt* so that T is infinite but we also assume that each firm distributes all the residual income (or cash flow) as dividends to equity holders. This thus implies that the firm does not have any cash reserve that it can use when the EBIT is not enough to pay the coupon C . Let v_t^l be the value of the levered firm at a given time t . We now compute v_t^l under various scenarios.

Levered Firm—No Default Risk

We say that a perpetual corporate debt is *risk-free* if the coupon C is paid at each time t to debt holders with probability one. Since the debt is risk-free, the net result $(X_t - C)(1 - T_a)$ is positive with probability one.^a The value of the levered firm at time t is thus equal to

$$v_t^l = C + (X_t - C)(1 - T_a) = X_t(1 - T_a) + T_a C \quad (2)$$

$$= v_t + T_a C \quad (3)$$

The value of the levered firm is equal to the value of the unlevered firm v_t , plus the value of the tax-shield $T_a C$, which comes from the fact that interest expenses are tax-deductible. In the particular case in which $T_a = 0$, we thus have the following equality:

$$v_t^l = v_t = X_t \quad (4)$$

When there is no corporate income tax, the value of the levered firm (at time t) is equal to the value of the unlevered one (at time t), that is, the value of the firm is invariant with respect to C as long as the coupon implies no default risk. This is the M–M theorem without default risk.

Levered Firm—Default Risk

We say that the corporate bond is risky if there is a positive probability that the coupon will not be paid to debt holders at some time t . Since all the residual income is distributed to equity holders as dividends, there are no available retained earnings so that default is declared at the first time for which $X_t < C$. Note importantly that in a default^b situation

- debt holders have an unconditional right to be paid before equity holders;
- the firm is not subject to corporate taxation;
- due to *limited liability* of equity holders, debt holders cannot force them to pay the difference $C - X_t$; and
- the bankruptcy cost (i.e., lawyer cost, administrative costs, etc.) is a fraction γ of X_t .

It follows thus that

$$v_t^l = \min\{X_t, C\}(1 - \gamma \mathbf{1}_{X_t < C}) + \max\{(X_t - C)(1 - T_a \mathbf{1}_{X_t \geq C}); 0\} \quad (5)$$

where $\mathbf{1}_{X_t \geq C}$ and $\mathbf{1}_{X_t < C}$ are indicator functions.^c When there is no default, the value of the firm is given by equation (3). However, when there is default, the value of the firm is equal to $X_t(1 - \gamma)$. In the particular case in which $\gamma = 0$ and $T_a = 0$, it follows that

$$v_t^l = \min\{X_t, C\} + \max\{(X_t - C); 0\} = v_t = X_t \quad (6)$$

When there is no bankruptcy cost and no corporate income tax, the value of the levered firm is equal to the value of the unlevered one, that is, the value of the firm is invariant with respect to the coupon. This is the M–M theorem with default risk.

Classical Presentation

In their original work, Modigliani and Miller [10] make the following assumptions.^d

1. Capital markets are perfectly competitive.
2. Individuals and firms can borrow and lend at the risk-free rate r .
3. All firms are assumed to be in the same *class risk*.

Assumption 1 implies that market investors have full (and symmetric) information concerning the return of the firm, that there are no transaction costs, and that there are no restriction on asset trade, that is, long and short positions are possible. Assumption 2 means that when firms or households borrow, they are not subject to default risk so that they can borrow at the risk-free rate. Assumption 3 means that the stream of EBIT is the same for all firms in the same class risk; if two firms, one levered and one unlevered, belong to the same class risk, then, they differ only with respect to leverage.

In their original work, Modigliani and Miller [10] consider a discrete-time model in which the family of random variable $\{X_t\}_{t \geq 1}$ is identically and independently distributed. Let $\mathbb{E}(X_t) = \bar{X}$ be the expectation of X_t under the probability measure $\mathbb{P}$. In the present framework, the corporate debt is riskless if $\mathbb{P}(X_t > C) = 1$.

Cost of Capital and Valuation Principle

In M–M [10], the valuation principle is fairly simple: the present value of a stream of *risky* cash flow is equal to the *expected* sum over time of the *discounted* value of the cash flow. The expectation is taken under the probability measure $\mathbb{P}$, and the discount factor, also called *risk adjusted rate of return* or, more simply, the *cost of capital* incorporates a risk premium. Since, in general, investors are risk averse, the riskier the cash flow, the higher is the cost of capital. Let r be the cost of capital of a riskless cash flow and $k_u > r$ be the cost of capital of a risky cash flow.

Using the above principle, the value of the unlevered firm is equal to the present value of the flow of v_t , that is,

$$V = E = \mathbb{E} \sum_{t=1}^{\infty} \frac{v_t}{(1 + k_u)^t} = \frac{\bar{X}(1 - T_a)}{k_u} \quad (7)$$

where V denotes the value of the unlevered firm, which is also equal to the value of equity E . To determine the value of the levered firm now, we have to compute the expected discounted value of the flow of v_t^l . Using equation (3), we just have to compute the value of the stream of tax shield. By assumption, the corporate debt is a perpetual risk-free one, so that the relevant cost of capital for the tax shield is r . It thus follows that

$$\begin{aligned} V^l &= \mathbb{E} \sum_{t=1}^{\infty} \frac{X_t(1 - T_a)}{(1 + k_u)^t} + \sum_{t=1}^{\infty} \frac{T_a C}{(1 + r)^t} \\ &= V + T_a \frac{C}{r} = V + T_a B \end{aligned} \quad (8)$$

where $B = \frac{C}{r}$ is the value of the perpetual corporate risk-free debt, and $T_a B$ the value of the perpetual tax shield. The value of the levered firm is thus equal to the value of the unlevered firm plus the value of the tax shield. We can now state the famous invariance result when $T_a = 0$. Let $\frac{B}{E^l}$ be the *debt–equity ratio* where $E^l = V^l - B$ is the value of equity of the levered firm.

Modigliani–Miller Theorem with Risk-free Debt

In addition to the above assumptions, suppose that there is no arbitrage opportunity and no corporate income tax (i.e., $T_a = 0$); then, $V = V^l$, that is, the value of the firm is invariant with respect to the debt–equity ratio.

This result implies that when there are no taxes and no default risk, the value of the firm is *invariant* with respect to the “financing policy”, which means that issuing only equity or issuing a mix of debt and equity does not affect the value of the firm. Fundamentally, the value of the firm is determined by the stream of EBIT, that is, by the investment policy; debt affects only the *split* of the value of the firm between debt holders and equity holders. As a consequence, different debt–equity ratios lead to different splits of the total value but have no impact on this total value.

From a mathematical point of view, the proof is trivial, since when $T_a = 0$, we immediately get that $V = V^l$. Actually, Modigliani and Miller became very famous not only for the result *per se*, but also because they offered, for the first time, a *no-arbitrage*

principle^c for security valuation: they showed that if $V^l > V$ when $T_a = 0$, then, there is an arbitrage opportunity.

While the proof is rather simple, it highlights their assumptions. Consider two firms that are in the same class risk. The first is unlevered while the second is levered. Since $T_a = 0$, the value of equity of the levered firm is equal to $E^l = V - C/r$ while the value of equity of the unlevered firm is $V = E$. Consider now an equity holder who owns $\kappa\%$ of the shares of the levered firm. At each time t , his/her claim D_t (i.e., dividend) is equal to

$$D_t = \kappa(X_t - C) \quad (9)$$

Suppose now that he/she sells his/her part of the levered firm for an amount κE^l and borrows an amount equal to $\kappa(C/r)$. With that money, he/she buys a fraction $\frac{\kappa(E^l + (C/r))}{V}$ of the shares of the unlevered firm. Actually, the equity holder borrows an amount $\kappa C/r$ to *replicate* the leverage of the firm, and this has been called a *homemade leverage*. Note importantly that by assumption 2, the investor borrows at the risk-free rate. He/she thus holds a portfolio in which he/she will not only perceive the dividends (in proportion of his/her part) of the unlevered firm at each time t but will also have to pay the interest expense equal to κC . The return of his/her portfolio is thus equal to $Y_t = \frac{\kappa(E^l + C/r)}{V} X_t - \kappa C$. By definition, $V^l = E^l + B$ so that Y_t is equal at each time t to

$$Y_t = \kappa \left(\frac{V^l}{V} X_t - C \right) \quad t \geq 1 \quad (10)$$

Since $V^l > V$ by assumption, it thus follows that $Y_t > D_t$ with probability one, which is an arbitrage opportunity. If we rule out such an arbitrage opportunity, then, $V^l = V$.

As Duffie [2] remarks in his review on M–M theorem, the result follows because equity holders can undertake the same financial transaction as the firm in exactly the same condition. If an equity holder is considered as defaultable so that he/she must borrow at a rate higher than the risk-free rate, then, he/she cannot offset the financial decision of the firm at the same price.

The Cost of Equity and the Debt–equity Ratio

From equation (7), when $T_a = 0$, the cost of capital $k_u = \bar{X}/E$ is the (expected) rate of return on equity. When the firm is levered, with $T_a = 0$, the (expected) rate of return of the levered firm k_e that we call *cost of equity* is equal to

$$k_e = \frac{\bar{X} - C}{E^l} \quad (11)$$

Since the firms are in the same class risk, the mean EBIT $\bar{X}$ of the two firms are equal. As a consequence, by writing equation (11) as $\bar{X} = E^l k_e + rB$ (recall that $C = rB$), since $\bar{X} = k_u E$, it thus follows that

$$E^l k_e + rB = k_u E \quad (12)$$

The M–M theorem implies that $V = V^l \equiv E^l + B$. Since $V = E$, from equation (12), we easily obtain that

$$k_e = k_u + (k_u - r) \frac{B}{E^l} \quad (13)$$

Equation (13) means that *ceteris paribus*, the higher the debt–equity ratio, the higher is the cost of equity of the levered firm. More precisely, the cost of equity is a *linear* increasing function of the debt–equity ratio, and is sometimes thought of as a pricing formula where $(k_u - r) \frac{B}{E^l}$ is the premium for the *financial risk*. See [14] for a presentation of the various existing formulas of the cost of capital.

Contrast with Empirical Evidence

In contrast to the M–M theorem, the value of firms is actually influenced by their capital structure: the debt–equity ratio affects the value of the firm. To explain the observed debt–equity ratio, a large part of the recent theory of capital structure has been to explore the role of the conflicts, inexistent in the M–M analysis, that arise among managers, equity holders, and debt holders. Informational asymmetries play an important role. The article of Harris and Raviv [5] and the books of Milgrom [9] or Hillier *et al.* [7] contain a very readable presentation of many aspects of the recent theory of capital structure.

Modern Presentation

We now present the M–M theorem in a continuous-time setting assuming that markets are arbitrage free

and complete^f so that there exists a *unique pricing measure* (also called risk-neutral measure) denoted by $\mathbb{Q}$. More precisely, we assume that under the unique risk-neutral measure $\mathbb{Q}$, the stochastic process $X = (X_t)_{t \geq 0}$ is given by

$$X_t = x e^{(\mu - 0.5\sigma^2)t + \sigma W_t} \iff \frac{dX_t}{X_t} = \mu dt + \sigma dW_t \quad (14)$$

where μ is the risk-neutral drift, σ is the volatility, W_t is a standard Brownian motion under $\mathbb{Q}$, and x is the value of the EBIT at time $t = 0$. In what follows, the expectation is taken under the pricing measure $\mathbb{Q}$.

Valuation Principle

In the “modern” approach, one also values a stream of risky cash flow as the *expected* sum over time of the *discounted* value of the cash flow. The discount rate is now always equal to the risk-free rate r , whether the cash flow is risky or not, and, to prevent arbitrage,^g the pricing measure $\mathbb{Q}$ is such that the *overall expected rate of return* of the unlevered equity, that is, expected capital gain plus dividend yield, is equal to the risk-free rate r .

The value of the unlevered firm seen from $t = 0$, which is also the value of equity, is equal to the expected sum of discounted v_t (see equation 1) over time, that is

$$\begin{aligned} V(x) \equiv V &= \mathbb{E} \int_0^\infty e^{-rt} v_t dt \\ &= \mathbb{E} \int_0^\infty e^{-rt} X_t (1 - T_a) dt \end{aligned} \quad (15)$$

$$= \frac{x(1 - T_a)}{r - \mu} \quad \mu < r \quad (16)$$

Seen from a given time $t > 0$, the value of the unlevered firm (or equity) is equal to

$$V_t = \frac{X_t(1 - T_a)}{r - \mu} \quad (17)$$

We assume, as previously, that equity is a *traded asset*, whose price is V_t at time t . Since all the residual cash flow is distributed to equity holders as dividends, the dividend yield is equal to $\frac{X_t(1 - T_a)}{V_t} =$

$r - \mu$, which means that $\mu < r$. As a consequence, the overall expected rate of return from holding the equity of the unlevered firm is equal to r , that is, the expected capital gain μ plus the dividend yield $r - \mu$.

Default Risk

The existence of default risk complicates the analysis of the preceding section, since the flow of v_t^l may be stopped at a random time τ . Moreover, if default occurs at time τ , a fraction γV is lost due to bankruptcy costs. As a consequence, the value of the levered firm is equal to V plus the value of the tax-shield TS minus the value of the bankruptcy costs BC . The value of the levered firm V^l at time $t = 0$ is thus equal to

$$V^l = V + TS - BC \quad (18)$$

To compute the value (at time $t = 0$) of the tax-shield TS and of the bankruptcy costs BC , we now have to specify the default declaration mechanism. It is assumed that default is triggered at the first time $t > 0$ for which $V_t \leq V_B$, with $V_B < C/r$. Let τ denote the default time. Since (almost all) the sample paths of the stochastic process V_t are continuous, default occurs at the first time for which $V_t = V_B$. It follows that

$$\tau = \inf\{t \in \mathbb{R}^+ : V_t = V_B\} \quad (19)$$

It can be shown^h that, for a perpetual corporate debt, the risk-neutral default probability is equal to

$$\left(\frac{V_B}{V}\right)^{2(\mu - 0.5\sigma^2)/\sigma^2} \quad (20)$$

To determine TS and BC , we follow [3] and first compute the present value of a contingent claim that pays €1 if the firm defaults. Seen from $t = 0$, if the default time τ were to be known, its value would be equal to $e^{-r\tau}$. Since $\tau > 0$ is not known, the value of such a contingent claim is equal to its expected value. It can be shownⁱ that

$$\lim_{T \rightarrow \infty} \mathbb{E}(e^{-r\tau} \mathbf{1}_{\tau < T}) = \left(\frac{V_B}{V}\right)^\xi \quad (21)$$

where $\xi = \frac{1}{\sigma^2}[(\mu - 0.5\sigma^2) + \sqrt{(\mu - 0.5\sigma^2)^2 + 2r\sigma^2}] > 0$, and $\mathbf{1}_{\tau < T}$ is the indicator function of the event $\tau < T$ (recall that T is the maturity of the debt).

Consider now the value of the bankruptcy costs. Seen from $t = 0$, when default occurs at time $\tau > 0$, their value is $e^{-r\tau} \gamma V_B$ while if default never occurs, their value is zero. Using equation (21), it thus follows that the value of the bankruptcy costs is equal to

$$BC = \lim_{T \rightarrow \infty} \mathbb{E}(\gamma V_B e^{-r\tau} \mathbf{1}_{\tau < T}) = \gamma V_B \left(\frac{V_B}{V}\right)^\xi \quad (22)$$

In the same way, since the tax shield is positive as long as default is not declared, its value is equal to

$$TS = \left(1 - \left(\frac{V_B}{V}\right)^\xi\right) T_a(C/r) \quad (23)$$

The value of the levered firm, defined by equation (18), is thus equal to

$$V^l = V + \left(1 - \left(\frac{V_B}{V}\right)^\xi\right) T_a(C/r) - \left(\frac{V_B}{V}\right)^\xi \gamma V_B \quad (24)$$

Note that when V (or x) tends to infinity, equation (24) reduces to equation (8) since default risk vanishes. When $T_a = 0$ and $\gamma = 0$ in equation (24), it thus follows that $V^l = V$.

Modigliani–Miller theorem with risky debt

Assume that markets are arbitrage free and complete and that there are no corporate income taxes (i.e., $T_a = 0$) and no bankruptcy costs (i.e., $\gamma = 0$). Then $V = V^l$, that is, the value of the firm is invariant with respect to the debt–equity ratio.

Default risk does not invalidate the M–M theorem as long as there are no taxes and no bankruptcy costs. When there are both tax and bankruptcy costs, the debt–equity ratio that maximizes the value of the firm is an optimal trade-off between the value of the tax shield and the value of the bankruptcy costs.^j However, it is assumed not only that there is no cash reserve (due to the particular dividend policy) but also that equity holders can *costlessly* issue new shares to avoid default if they want to do so.

Assume now that some fraction of the net income (when positive) is not distributed to equity holders as

dividends, it goes on a cash reserve that capitalizes at a rate equal to $r - \lambda$. The parameter λ reflects inefficiencies of managerial expenses (agency costs) so that some fraction of the cash can be considered as lost. When there are both issuance and agency costs, it has been shown that the *optimal dividend policy*^k is an optimal trade-off between these two costs.

End Notes

^aThis might not be the case if, for example, equity holders issue new equity to pay $C - X_t$ to bondholders when $X_t < C$.

^bHere, we make no difference between default and liquidation.

^cRecall that the indicator function of an event A is such that $\mathbf{1}_A = 1$ if $\omega \in A$ and 0 otherwise.

^dSee [13] for a lucid discussion of their assumptions.

^eTheir no-arbitrage proof was indeed misunderstood. See [6] and the reply in [11].

^fIn a section entitled Arrow-Debreu Securities, [12] shows the M–M theorem in a (static) complete market setting. For a treatment of the incomplete market case, see [1] or [4].

^gSee, for example, Kerry Back (2005). *A Course in Derivatives Securities*, Springer-Verlag, p. 42.

^hSee, for example, Bielecki, T. & Rutkowski, M. (2001). *Credit Risk: Modeling, Valuation and Hedging*, Springer-Verlag, p. 68.

ⁱSee Bielecki T. and Rutkowski M. (2001), p 81 for a probabilistic proof or [3], p 491 for an alternative proof.

^jThis model has been developed by Leland [8].

^kSee Decamps, J.P., Mariotti, T., Rochet, J.C. & Villeneuve, S. (2008). *Free Cash Flow, Issuance Costs, and Stock Price Volatility*, IDEI Working Paper, n° 518, September 2008.

References

- [1] De Marzo, P. (1988). An extension of the Modigliani-Miller theorem to stochastic economies with incomplete markets, *Journal of Economic Theory* **45**(2), 261–286.
- [2] Duffie, D. (1992). Modigliani-Miller theorem, in *The New Palgrave Dictionary of Money and Finance*, M. Milgate, P. Newman, & J. Eatwell, eds, McMillan Press, London.
- [3] Goldstein, R., Ju, N. & Leland, H. (2001). An EBIT-based model of dynamic capital structure, *Journal of Business* December, 483–512.
- [4] Gottardi, P. (1995). An analysis of the conditions for the validity of the Modigliani-Miller theorem with incomplete markets, *Economic Theory* **5**, 191–207.
- [5] Harris, A. & Raviv, A. (1991). The theory of capital structure, *Journal of Finance* March, 297–355.
- [6] Heins, J. & Sprenkle, C. (1969). A comment on the Modigliani-Miller cost of capital thesis, *American Economic Review* September, 590–592.
- [7] Hillier, D., Grinblatt, M. & Titman, S. (2008). *Financial Markets and Corporate Strategy*, European edition, Mac Graw Hill.
- [8] Leland, H. (1994). Corporate debt value, bond covenants, and optimal capital structure, *Journal of Finance* **49**, 1213–1252.
- [9] Milgrom, P. & Roberts, J. (1992). *Economics, Organization, and Management*, Prentice-Hall.
- [10] Modigliani, F. & Miller, M. (1958). The cost of capital, corporation finance, and the theory of investment, *American Economic Review* **48**, 261–297.
- [11] Modigliani, F. & Miller, M. (1969). Reply to Heins and Sprenkle, *American Economic Review* September, 592–595.
- [12] Stiglitz, J. (1969). A re-examination of the Modigliani-Miller theorem, *American Economic Review* **59**, 784–793.
- [13] Stiglitz, J. (1988). Why financial structure matters, *Journal of Economic Perspectives* **2**(4), 121–126.
- [14] Taggart, R. (1991). Consistent valuation and the cost of capital expressions with corporate and personal taxes, *Financial Management* Autumn, 8–20.

Further Reading

- Hellwig, M. (1981). Bankruptcy, limited liability and the Modigliani-Miller theorem, *American Economic Review* **71**(1), 155–170.
- Modigliani, F. & Miller, M. (1961). Dividend policy, growth and the valuation of shares, *Journal of Business* **34**(4), 411–433.
- Stiglitz, J. (1974). On the irrelevance of corporate financial policy, *American Economic Review* December, 851–866.

Related Articles

Arbitrage Pricing Theory; Modigliani, Franco.

YANN BRAOUEZEC

Multifractals

Since the early work of Mandelbrot on the fluctuations of the cotton price in the early sixties, it is well known that market price variations are poorly described by the standard geometric Brownian motion [14]. Extreme events are more probable than in a Gaussian world and volatility fluctuations are well known to be of an intermittent and correlated nature. As we discuss in this article, multifractal analysis has provided new concepts and tools to analyze market fluctuations and has inspired a particularly elegant family of models that accounts for the main observed empirical properties in a parsimonious way. These models while capturing the “heteroskedastic” nature of return fluctuations, still preserve, in some sense, the nice stability property across timescales of Brownian motion. This sharply contrasts with classical econometric models that can hardly be controlled as the timescale is changed.

Multifractal Scaling

Let $p(t)$ be the price of some asset at time t , $X(t) = \ln[p(t)]$ and $\delta_\ell X(t)$ the log return between times t and $t + \ell$: $\delta_\ell X(t) = X(t + \ell) - X(t)$.^a $X(t)$ will be called a *multifractal process* if, for a range of q , the order q absolute moment of $\delta_\ell X(t)$ behaves as a power law when ℓ varies

$$M(q, \ell) = \mathbb{E} [|\delta_\ell X|^q] \sim C_q \ell^{\zeta(q)}, \quad l \leq T \quad (1)$$

where $\mathbb{E}[\cdot]$ stands for the mathematical expectation, the exponent $\zeta(q)$ is a nonlinear concave function of q and T is a large characteristic timescale referred to as the *integral scale*.

The multifractal nature of market data is illustrated on the S&P 500 index in Figure 1 using traded prices sampled every 10 min. In perfect agreement with equation (1), log–log plots of $M(q, l)$ versus ℓ show linear behavior up to a scale of the order of $T \simeq 1$ year and the so-obtained $\zeta(q)$ spectrum (using linear regression) turns out to be well fitted by a parabolic shape. The S&P 500 future index can thus be considered, at least at this description level, as a multifractal signal. During the last decade, a lot of works have studied various market data within the framework of multifractal scaling analysis: on

FX rates, commodity markets, stock markets, future markets, and emerging markets (see [2] for a review). These studies point to multifractality as a stylized feature of financial time series.

The scaling of return moments as a function of the timescale ℓ is, in fact, intimately related to some self-similarity property of the process $X(t)$. A process $X(t)$ is called *self-similar* of exponent H if it has stationary increments and if $\forall s > 0$,

$$\delta_{s\ell} X(st) \stackrel{\text{Law}}{=} s^H \delta_\ell X(t) \quad (2)$$

This means that the original process and a dilated version of it are statistically undistinguishable. In that case, $M(q, \ell) \sim \ell^{qH}$ and $\zeta(q)$ is a linear function of q . To account for multifractality (i.e., nonlinear $\zeta(q)$), one can replace the deterministic factor s^H by a random one and get the stochastic self-similarity property:

$$\delta_{s\ell} X(st) \stackrel{\text{Law}}{=} W_s \delta_\ell X(t) \quad (3)$$

where $W_s = e^{\omega_s}$ is a positive random variable independent of the process X and which law is log infinitely divisible and only depends on s . From this equation, it is easy to show that $\zeta(q) = \ln \mathbb{E} [e^{-q\omega_s}] / \ln s$. The multifractality *strength* of the process is generally characterized by the curvature of $\zeta(q)$, that is, by the so-called intermittency coefficient $\lambda^2 = -\zeta''(0)$. The simplest nonlinear case is the log-normal model that corresponds to a parabolic $\zeta(q)$ (i.e., ω_s is Gaussian).

Let us point out that the stochastic self-similarity implies (i) that the smaller the ℓ is, the *fatter* the tails of the probability distribution function (PDF) of $\delta_\ell X$ are and (ii) that the covariance function of $\ln |\delta_\ell X|$ is slowly decreasing (up to scale T) as the logarithm of the lag. These two features are indeed observed on most financial data as illustrated on S&P 500 future data in Figures 2 and 3.

Random Cascades Models

Let us first note that, following an idea of Mandelbrot [14], if $\theta(t)$ (referred to as the *trading time*) is a nondecreasing multifractal process satisfying equation (3) and $B(t)$ a Brownian motion that is independent of $\theta(t)$, then the process

$$X(t) = B[\theta(t)] \quad (4)$$

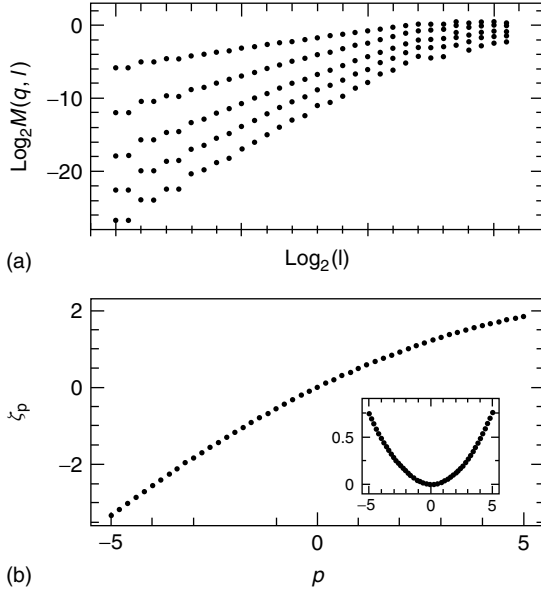


Figure 1 Multifractal Analysis of S&P 500 future index. Traded prices have been sampled every 10 min during the period 1988–1999. (a) Log–log plots of $M(q, l)$ versus l for $q = 1, 2, 3, 4, 5$. The timescales l range from 10 min to 6 years. They display linear behavior up to a scale $T \simeq 1$ year. (b) $\zeta(q)$ spectrum obtained by linear regression of plots on the left. The plot in the inset is the parabolic nonlinear part of $\zeta(q)$

is also multifractal with $\zeta_X(q) = \zeta_\theta(q/2)$. If we write $\theta(t) = \int_0^t d\theta(t)$, the *multifractal random measure* $d\theta(t)$ can be seen as the *instantaneous stochastic volatility*. Consequently, multifractal models of asset returns reduce to multifractal models of volatility measures.

Discrete multiplicative cascades were first introduced in the field of empirical finance by the pioneering work of Calvet *et al.* on exchange rates volatility [15]. Their construction is a direct translation of the self-similarity property (3). In the simplest case, the construction of $\theta(t)$ involves a dyadic tree in a timescale (t, s) half plane. The level n lies at scale $s_n = T2^{-n}$ and has 2^n nodes $\{p_{n,k}\}_{0 \leq k < 2^n}$, which are uniformly distributed on $[0, T]$ leading to a partition made of 2^n intervals $I_{n,k}$ of size s_n . Moreover, i.i.d. (positive) log infinitely divisible random factors $W_{n,k}$ are associated with the nodes $p_{n,k}$. The cascading process starts at the integral scale $s_0 = T$ where a measure is uniformly spread on $[0, T]$: $d\theta_0(t) = dt$. At step $n = 1$, the measure $d\theta_1(t)$ is obtained from

$d\theta_0(t)$ by multiplying its density on the interval $I_{1,0}$ (respectively, $I_{1,1}$) by the corresponding factors $W_{1,0}$ (respectively, $W_{1,1}$). At step n , $d\theta_n(t)$ is obtained by multiplying $d\theta_{n-1}(t)$ on each $I_{n-1,k}$ by $W_{n-1,k}$ leading to the representation

$$d\theta_n(u) = 2^{-n} e^{\sum_{i=1}^n \delta\omega_{i,u_k}} dt \quad (5)$$

where $W_{i,u_k} = e^{\delta\omega_{i,u_k}}$ and u_k corresponds to the integer part of $u2^n$. Finally, one gets $d\theta(t) = \lim_{n \rightarrow +\infty} d\theta_n(t)$.

Grid-bounded cascades, though simple, do not provide a satisfying solution to model volatility fluctuations: (i) they involve an arbitrary fixed scale ratio, (ii) they are not built in a causal way, and (iii) they are not stationary. In their “Poisson multifractal model”, Calvet and Fisher [7] get rid of (ii) and (iii), by basically replacing, for a fixed n , the points $\{p_{n,k}\}_k$ by random Poisson points with intensity $r(s_n) = s_n^{-1}$. Barral and Mandelbrot [5] go one step further (getting rid of (i)), by replacing the points $\{p_{n,k}\}_{n,k}$ and their associated weights $\{W_{n,k}\}_{n,k}$ by a nonhomogeneous spatial compound Poisson process, with the compound law W and the intensity $r(s, t) = s^{-1}$. This construction can be generalized by replacing the compound Poisson process by an arbitrary infinitely divisible random noise $d\omega(t, s)$ in the half plane (t, s) (with density $dt ds/s^2$). Equation (4) is replaced by

$$d\theta_s(t) = e^{\int_{(s', t') \in C_s(t)} d\omega(t', s')} dt \quad (6)$$

where $C_s(t)$ is a cone-like domain (pointing at $(0, t)$ and truncated below scale s). This construction [3] corresponds to the fully continuous generalization of equation (4). The limiting measure $d\theta(t) = \lim_{s \rightarrow 0} d\theta_s(t)$ is shown to be stochastic self-similar, with *exact* scaling ($M(q, l) = C_q l^{\zeta(q)}$, $\forall l \leq T$) and having none of the drawbacks (i)–(iii). Of course, the corresponding $X(t) = B(\theta(t))$ process shares exactly the same properties, it is referred to as a *multifractal random walk (MRW)*.

The “simplest” MRW model corresponds to the case W is log-normal. In that case $d\omega(t', s')$ is a Gaussian white noise, and it can be shown that $\theta(t)$ can be obtained directly by

$$\theta(t) = \lim_{\Delta \rightarrow 0} \int_0^t e^{2\omega_\Delta(u)} du \quad (7)$$

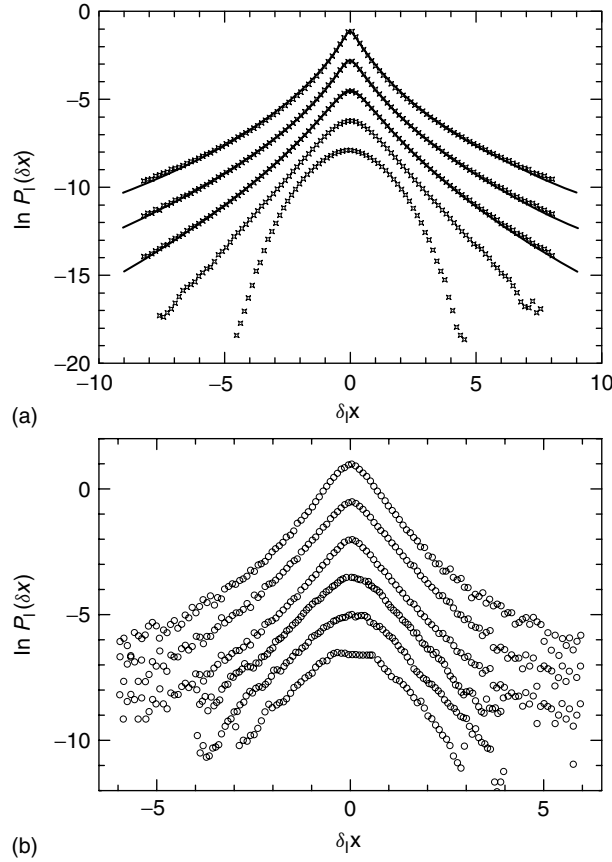


Figure 2 Continuous deformation of increment PDF's $\delta_\ell X$ across scales. Standardized PDF's (in logarithm scale) at large (b) to small (a) scales. Plots have been arbitrarily shifted along the vertical axis to illustrate the tails getting fatter as the scale is going smaller. (b) same S&P 500 data as in Figure 1. The scales range from 10 min to 1 month. (a) MRW Model with $\lambda^2 = 0.03$, $\Delta = 1/16$ and $T = 8192$. The scales range from 16 sampling size to two integral scales. Estimation ($\circ$) made from 500 MRW realizations of 2^{17} sampled points. The solid line corresponds to the prediction from the largest scale using equation (10)

where the *magnitude process* $\omega_\Delta(u)$ is a stationary Gaussian process fully defined by

$$\mathbb{E}[\omega_\Delta(t)] = -\lambda^2 \ln\left(\frac{T}{\Delta}\right) \quad (8)$$

$$R_\Delta(\tau) = \begin{cases} \lambda^2 \left(\ln\left(\frac{T}{\Delta}\right) + 1 - \frac{\tau}{\Delta} \right), & \tau \leq \Delta \\ \lambda^2 \ln\left(\frac{T}{\tau}\right), & \tau \in [\Delta, T] \\ 0, & \tau > T \end{cases} \quad (9)$$

where $R_\Delta(\tau) = \mathbb{Cov}(\omega_\Delta(t), \omega_\Delta(t + \tau))$.

Statistical issues such as parameter estimation remain to be studied (see [18]). Calvet and Fisher [7] introduced a Markov-switching version of their

cascade model that is amenable to maximum likelihood estimation. They were the first to propose a generalized method of moments (GMMs) to estimate the parameters of their model (see also [12, 13] for discrete cascades). The log-normal MRW model is fully defined by three parameters: σ^2 the variance of the Brownian motion, T the integral scale (decorrelation timescale of the volatility), and λ^2 the intermittency factor, which is involved in the nonlinear part of ζ_q (equation (8)). Since, as briefly discussed below, many of statistical moments associated with the MRW model can be analytically computed, GMM can be simply devised and all these parameters can be easily estimated [11]. We observed that for both

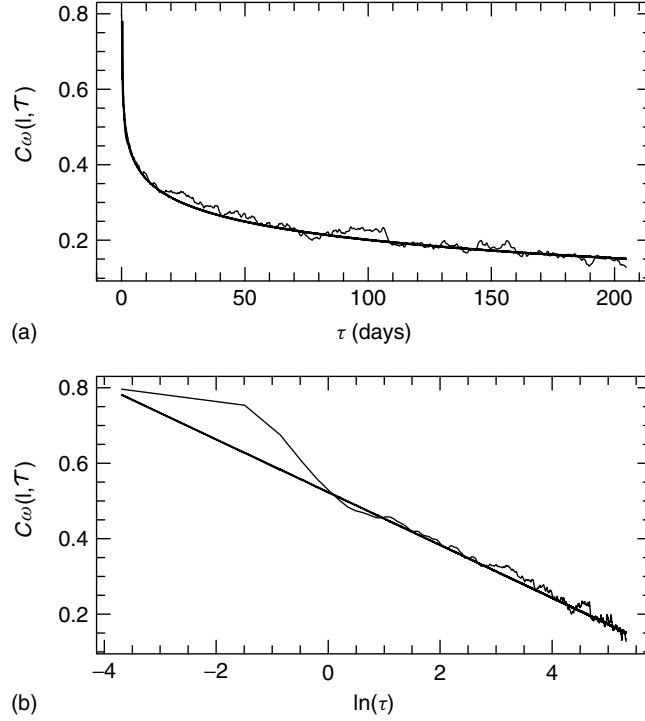


Figure 3 Log-volatility covariance $C_\omega(\ell, \tau)$ of the S&P 500 future. (a) $C_\omega(\ell, \tau)$ versus τ for $\tau = 10$ min. The solid line represents a fit according to the MRW model logarithmic expression. (b) $C_\omega(\ell, \tau)$ versus $\ln(\tau)$. The MRW model predicts a linear behavior that crosses the y-axis at $\ln(\tau) = \ln(T)$

the daily and intraday financial data, the correlation time T is very large (greater than one year) and the parameter λ^2 lies in $[0.02, 0.05]$.

Properties and Risk Forecasting

As far as multifractal properties are concerned, $X(t)$ verifies the exact stochastic self-similarity property as defined by equation (3) and therefore its moments behave exactly as a power law of the timescale ($\ell \leq T$) with a parabolic multifractal spectrum

$$\zeta(q) = q \left(\frac{1}{2} + \lambda^2 \right) - \frac{\lambda^2}{2} q^2 \quad (10)$$

Many other scaling laws can be computed and notably those of “ p -volatility” correlations $C_p(\ell, \tau) \equiv \mathbb{E} [|\delta_\ell X(\tau)|^p |\delta_\ell X(0)|^p]$. In [17], it is shown that

$$C_p(\ell, \tau) \sim K_p^2 \left(\frac{\ell}{T} \right)^{2\zeta_p} \left(\frac{\tau}{T} \right)^{-\lambda^2 p^2} \quad (11)$$

where the constant K_p can be computed analytically. For proofs and details, see [2, 3]. Although they did not explicitly refer to the framework of multifractal analysis, Ding and Granger [9] were the first to observe a nontrivial dependence of $C_p(\ell, \tau)$ as a function of p . Note that when $p \rightarrow 0$, one recovers the logarithmic slowly decreasing of the log-volatility covariance $C_\omega(\ell, \tau)$ as illustrated in Figure 3.

Equation (9) can be directly used to predict volatility. In [7], Calvet and Fisher have already shown that their cascade model provides better volatility forecasts as compared to GARCH(1,1) or Markov-switching GARCH (see also [13]). Bacry and Muzy [2, 4] have shown that linear volatility forecast provided by the MRW model outperforms GARCH(1,1) models.

Actually, the MRW model allows to estimate the full (conditional) return probability density function at all time horizons and scales. Very much like standard Brownian motion that is stable with respect to time aggregation, the self-similarity properties

of the MRW process $X(t)$ allows to control how the return probability law changes when varying the timescale. Indeed, it can be shown that, in a good approximation, when $\lambda^2 \ll 1$, the returns of the MRW process, can be written, in law, as

$$\delta_\ell X(t) \underset{\text{Law}}{\simeq} \epsilon_\ell(t) e^{\Omega_\ell(t)} \quad (12)$$

(see [4] for details and for the precise meaning of the symbol $\simeq$) for all $\ell \leq T$, where ϵ_ℓ is a Gaussian white noise of variance $\sigma^2 \ell$ and the process $\Omega_\ell(t)$ is the *renormalized magnitude*. It is a stationary Gaussian process whose mean and covariance are very similar to equation (7). In Figure 2, we see that, as observed for S&P 500 data, the MRW returns PDF strongly depends on the timescale and evolves from a “quasi-Gaussian” shape at large scales ($\ell \simeq T$) to PDF’s with high kurtosis at small scales. Even if the precise shape of the PDF tails returns is a matter of debate, the most commonly admitted form is a power law with a tail exponent μ within the interval [3, 5] (see, e.g., [6, 10]). The MRW model allowed to point out [1, 16] that this behavior is strongly related to the choice of the asymptotics in the estimation process. The “high-frequency asymptotics” (sampling frequency going to infinity) has to be distinguished from the “long duration asymptotics” (duration of the considered time period going to infinity).

All previous considerations can be extended to conditional laws and can be of practical interest for VaR forecasting. As shown in [2, 4], MRW predictions of VaR at various time horizons can be more reliable than predictions based on a GARCH(1,1) model.

End Notes

^aWe do not address here the modeling of intraday modulation of the data. This modulation can be removed following [8]. We consider that $\delta_\ell X(t)$ is stationary and has zero mean.

References

- [1] Bacry, E., Gloter, A., Hoffmann, M. & Muzy, J.F. (2009). Multifractal analysis in a mixed asymptotic framework, *Annals of Applied Probability* Preprint, arXiv.org/math/0805.0194. To appear.
- [2] Bacry, E., Kozhemyak, A. & Muzy, J.F. (2008). Continuous cascade models for asset returns, *Journal of Economic Dynamics and Control* **32**, 156–199.
- [3] Bacry, E. & Muzy, J.F. (2003). Log-infinitely divisible multifractal process, *Communications in Mathematical Physics* **236**, 449–475.
- [4] Bacry, E., Kozhemyak, A. & Muzy, J.F. (2009). Log-normal continuous cascades: aggregation properties and estimation, *Quantitative Finance* Preprint, arXiv.org/physics/0804.0185. To appear.
- [5] Barral, J. & Mandelbrot, B.B. (2002). Multifractal products of cylindrical pulses, *Probability Theory and Related Fields* **124**, 409–430.
- [6] Bouchaud, J.P. & Potters, M. (2003). *Theory of Financial Risk and Derivative Pricing*, Cambridge University Press, Cambridge.
- [7] Calvet, L. & Fisher, A. (2004). How to forecast long-run volatility, *Journal of Financial Econometrics* **2**, 49–83.
- [8] Dacorogna, M.M., Gençay, R., Müller, U.A., Olsen, R.B. & Pictet, O.V. (2001). *An introduction to High Frequency Finance*, Academic Press, San Diego.
- [9] Ding, Z. & Granger, C.W.J. (1996). Modeling volatility persistence of speculative returns: a new approach, *Journal of Econometrics* **73**, 185–215.
- [10] Gopikrishnam, P., Meyer, M., Amaral, L.A.N. & Stanley, H.E. (1998). Inverse cubic law for the distribution of stock price variations, *The European Physical Journal B* **3**, 139–140.
- [11] Lastwave gnu software, (2008). www.cmap.polytechnique.fr/~bacry/LastWave
- [12] Lux, T. (2001). Turbulence in financial markets, *Quantitative Finance* **1**, 632.
- [13] Lux, T. (2004). *The Markov-Switching Multi-Fractal Model of Asset Returns*, Economic, University of Kiel, working paper.
- [14] Mandelbrot, B.B. (1997). *Fractals and Scaling in Finance. Discontinuity, Concentration, Risk*, Springer, New York.
- [15] Mandelbrot, B.B., Calvet, L. & Fisher, A. (1997). *The Multifractal Model of Asset Returns*, Preprint, Cowles Foundation Discussion paper 1164.
- [16] Muzy, J.F., Bacry, E. & Kozhemyak, A. (2006). Extreme values and fat tails of multifractal fluctuations, *Physical Review E* **73**, 066114.
- [17] Muzy, J.F., Delour, J. & Bacry, E. (2000). Modelling fluctuations of financial time series, *European Journal of Physics B* **17**, 537–548.
- [18] Ossiander, M. & Waymire, E.C. (2000). Statistical estimation for multiplicative cascades, *The Annals of Statistics* **28**, 1533–1560.

Further Reading

- Kahane, J.P. & Peyrière, J. (1976). Sur certaines martingales de Benoît Mandelbrot, *Advances in Mathematics* **22**, 131–145.
- Molchan, G.M. (1996). Scaling Exponents and Multifractal Dimensions for independent random cascades, *Communications in Mathematical Physics* **179**, 681–702.

EMMANUEL BACRY & JEAN-FRAN MUZY

Random Matrix Theory

Random matrix theory is the study of matrices with random entries. It has been applied in quantitative finance for the estimation of large covariance matrices which are of interest in portfolio optimization (see **Modern Portfolio Theory; Risk–Return Analysis**). This estimation problem is linked to understanding the properties of *large-dimensional random matrices*.

If daily returns on n days for p different assets are represented as an $n \times p$ matrix X , the sample covariance matrix $\hat{\Sigma}$ is defined as

$$\hat{\Sigma} = \frac{1}{n-1}(X - \bar{X})(X - \bar{X})' \quad (1)$$

where X' is the transpose of X and $\bar{X}$ is a $(n \times p)$ matrix of means, that is, its k th column has constant value equal to the mean of the k th column of X . These matrices are important in practice because they are used to estimate from the data the (unknown) population covariance Σ , when the rows of X , $\{X_i\}_{i=1}^n$ are assumed to be independent identically distributed (i.i.d.) p -dimensional vectors with covariance Σ . When that is the case, $\hat{\Sigma}$ is an unbiased estimator of Σ , that is, $E(\hat{\Sigma}) = \Sigma$.

In the Markowitz portfolio optimization problem (see **Risk–Return Analysis**), optimal weights for different assets are allocated according to a formula involving the covariance Σ of the assets. Since Σ is unknown, it is necessary to estimate it, and the sample covariance matrix is often used as an estimator. Note that for large portfolios, p will be quite large, of order 100 for instance. Also, if looking at daily returns, the number of days, n , used to compute $\hat{\Sigma}$ is often chosen not to exceed a year or two of trading, and is also therefore of order a few 100s (roughly 250 to 500). Hence, in financial practice, it will often be the case that both p and n are “large”.

This is in contrast to the classical theory (see [1]), which is concerned with asymptotics where p is held fixed and n goes to infinity. In this setting of “small p , large n ”, the sample covariance matrix is known to be a good estimator of the population covariance. For instance, the sample eigenvalues can be shown to be close to the population eigenvalues, and similarly for corresponding eigenspaces. Practically, it means that using the sample covariance matrix in the solution of

a Markowitz portfolio optimization problem with a few assets will work “well”, if we observe the assets for a long enough time.

The “large p , large n ” situation is very different and is the one that is studied in random matrix theory. Roughly speaking, what happens in that situation is that, under “reasonable” assumptions detailed later, the eigenvalues of $\hat{\Sigma}$ become inconsistent estimators of the eigenvalues of Σ . Similarly, problems develop with eigenvectors.

The first result in this area was obtained by Marčenko–Pastur [21]. A subcase of the result shown there can, for instance, be interpreted as saying that if the entries of X are i.i.d $\mathcal{N}(0, 1)$, the histogram of eigenvalues of $\hat{\Sigma}$ is asymptotically *nonrandom* (though the matrix itself is random) and moreover its shape can be described. The limiting density of eigenvalues is, if $p/n \rightarrow r \in (0, 1)$,

$$f_r(x) = \frac{1}{2\pi r} \frac{\sqrt{(b-x)(x-a)}}{x} 1_{a \leq x \leq b} \quad (2)$$

In the above equation, $b = (1 + \sqrt{r})^2$ and $a = (1 - \sqrt{r})^2$. The probability distribution with corresponding density is often called the *Marčenko–Pastur law*. It is illustrated in Figure 1. (If $r > 1$, there is also a point mass of mass $(1 - 1/r)$ at 0.)

Several general facts about eigenvalues of random matrices can be observed in Figure 1. First, the sample eigenvalues are more dispersed than the population eigenvalues. Second, the extreme eigenvalues exhibit bias: instead of converging to the population (or “true”) extreme eigenvalues, which in the case above are concentrated at 1, the largest (respectively, smallest) sample eigenvalue is asymptotically at least as large as $b = (1 + \sqrt{r})^2$ (respectively, as small as $a = (1 - \sqrt{r})^2$). Under certain moment conditions on the entries of X , it can actually be shown that the extreme eigenvalues of $\hat{\Sigma}$ converge to b and a . We now turn to a more precise presentation of the results we have hinted at.

Eigenvalues of Sample Covariance Matrices

Two types of results are available for eigenvalues of sample covariance matrices: results dealing with the behavior of the “bulk” of eigenvalues and “edge” results, which deal with extreme (maximal or minimal) eigenvalues.

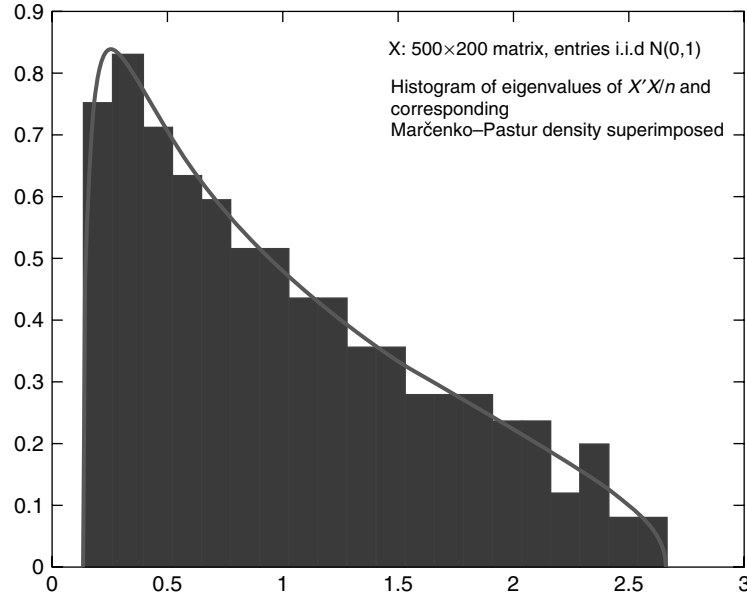


Figure 1 Density of Marčenko–Pastur law and histogram of eigenvalues. Here the histogram of eigenvalues of $X'X/n$ is plotted, for $p = 200$, $n = 500$. The entries of X are i.i.d $\mathcal{N}(0, 1)$. Since $\Sigma = \text{id}$, the “true” (or population) eigenvalues are all equal to 1, illustrating the fact that the sample covariance matrix is not a good estimator of the population covariance when $r \neq 0$. The curve is the theoretical prediction obtained from the Marčenko–Pastur law

“Bulk” results deal with the empirical distribution of the eigenvalues and study the convergence properties of this (random) probability measure. If $l_1 \geq l_2 \geq \dots \geq l_p$ denote the decreasingly ordered eigenvalues of the $p \times p$ matrix $\widehat{\Sigma}$, we call F_p the spectral distribution of $\widehat{\Sigma}$; by definition, we have

$$dF_p(x) = \frac{1}{p} \sum_{i=1}^p \delta_{l_i} dx \quad (3)$$

Here δ_x denotes a Dirac mass (of mass 1) at x . Note that F_p is a probability measure. Also, $F_p([x, \infty))$ is the proportion of eigenvalues of $\widehat{\Sigma}$ that are larger than x .

To study F_p in random matrix theory, it is common to use the Stieltjes transform of a distribution. By definition, the Stieltjes transform of a distribution G is a complex-valued function of one complex variable with strictly positive imaginary part, defined as

$$\begin{aligned} m_G(z) &= \int \frac{dG(x)}{x - z}, \text{ for } z \in \mathbb{C}^+ \\ &= \{z \in \mathbb{C} \text{ with } \text{Im}[z] > 0\} \end{aligned} \quad (4)$$

We note that m_G maps $\mathbb{C}^+$ into $\mathbb{C}^+$. The importance of Stieltjes transforms in this context comes from the fact that if F_p is a sequence of probability distributions and $m_{F_p}(z)$ converges, for all z in $\mathbb{C}^+$ to $m_F(z)$, where $m_F(z)$ is in $\mathbb{C}^+$, and $\lim_{y \rightarrow \infty} y m_F(iy) = -1$, then F_p converges weakly to F (and therefore F is a probability distribution). For a thorough introduction to the connection between convergence of Stieltjes transforms and convergence of probability distributions, we refer the reader to [15].

An important result concerning eigenvalues of sample covariance matrices is the following result shown in [26], which generalizes results in [21] and [29]. We note that the functional equation we are about to state was found first in [21] and is sometimes called the *Marčenko–Pastur* equation.

Theorem 1 (Silverstein [26]) *Let Y be an $n \times p$ matrix with i.i.d entries, denoted $Y_{i,j}$. Suppose that $Y_{i,j}$ has two moments, that is, $\mathbf{E}(Y_{i,j}^2) < \infty$, and variance 1. Let Σ_p be the population covariance matrix. In particular, Σ_p is positive semidefinite. Suppose the spectral distribution of Σ_p , H_p , has*

a limit, H , in the sense of weak convergence of distributions.

Let the data matrix be represented by $X = Y\Sigma_p^{1/2}$. Call F_p the spectral distribution of $X'X/n$ and m_p the Stieltjes transform of F_p . Call $v_p(z) = -(1 - p/n)/z + p/nm_p(z)$. (v_p is the Stieltjes transform of the spectral distribution of the matrix XX'/p .)

When $p/n \rightarrow r \neq 0$, F_p converges weakly almost surely (a.s.) to a deterministic (probability) distribution F . Moreover, $\forall z \in \mathbb{C}^+$, $v_p(z) \rightarrow v(z)$ a.s., and

$$-\frac{1}{v(z)} = z - r \int \frac{\lambda dH(\lambda)}{1 + \lambda v(z)} \quad (5)$$

It can be shown that there is a unique solution to the previous equation which is a Stieltjes transform.

Some comments are in order here. First, equation (5) relates the properties of the limit of the empirical spectral distribution (“encoded” by its Stieltjes transform in v) to those of the limiting population spectral distribution represented by H . Also, because finite rank perturbation of matrices does not affect limiting spectral distributions (see, for instance, [2]), the same result applies to the limiting spectral distribution of $(X - \bar{X})'(X - \bar{X})/n$, the sample covariance matrix used in practice. In other respects, the Marčenko–Pastur law can be obtained fairly simply from equation (5), as explained in [21] for instance. Finally, robustness properties (and lack thereof) of this functional equation have recently been investigated in [8], highlighting, in particular, geometric limitations for the data, implicitly implied by the model used in Theorem 1.

Properties of Extreme Eigenvalues

Limiting spectral distribution properties do not in general imply anything about properties of extreme eigenvalues, beside the fact that extreme eigenvalues need to be at the edge (or outside) of the support of the limiting spectral distribution. However, properties of these extreme eigenvalues are important in statistics, in particular for techniques such as principal component analysis (PCA; see [22]).

Results about extreme eigenvalues are available. An important one is for instance, when $\Sigma_p = \text{Id}_p$; then the largest eigenvalue of $X'X/n$ converges to $(1 + \sqrt{r})^2$. The result was first obtained in [14], and under weaker assumptions in [30]. We cite the latter result.

Theorem 2 (Yin et al. [30]) Suppose X is an $n \times p$ data matrix, with i.i.d entries. Suppose that $\mathbf{E}(X_{i,j}) = 0$, $\mathbf{E}(X_{i,j}^2) = 1$ and $\mathbf{E}(X_{i,j}^4) < \infty$. Suppose $p/n \rightarrow r \in (0, \infty)$, as $n \rightarrow \infty$. Then, we have, if $l_1(X'X/n)$ denotes the largest eigenvalue of $X'X/n$,

$$l_1\left(\frac{X'X}{n}\right) \rightarrow (1 + \sqrt{r})^2 \quad (6)$$

As explained in [25], the existence of the fourth moment is necessary for this result to hold. Similar convergence results exist concerning the smallest eigenvalue of $\hat{\Sigma}$ (see [3] and [2, p. 635]).

More recent theoretical work concerns fluctuation behavior of the largest eigenvalues of sample covariance matrices. Following mathematical breakthroughs in the 1990s (see [27]), it has become possible to describe the limiting distribution of $l_1(X'X/n)$. In particular, the papers [9, 13, 17, 18], show the following:

Theorem 3 Suppose X is an $n \times p$ matrix with i.i.d $\mathcal{N}(0, 1)$ entries. Suppose n and p tend to infinity. Then

$$n^{2/3} \frac{l_1(X'X/n) - (1 + \sqrt{p/n})^2}{(1 + \sqrt{p/n})(1 + \sqrt{n/p})^{1/3}} \Longrightarrow \text{TW}_1 \quad (7)$$

The most important part of the result (the case p/n having a finite nonzero limit) is shown in [18]. In the previous theorem, TW_1 is a Tracy–Widom distribution. Its density is known explicitly (see [28] and [18]).

In all the previous examples, $\Sigma_p = \text{Id}_p$. The case of general Σ is starting to be understood, too, and is more relevant to techniques such as PCA. The situation where Σ has a few isolated “large” eigenvalues and all the other eigenvalues are equal is now reasonably well understood (see [5, 23]). In this case, the empirical largest eigenvalue is biased and the bias can be computed explicitly. The corresponding sample eigenvector does not converge to the population eigenvector associated with the isolated eigenvalue. The situation where Σ is truly general is more complicated. A formula for the limit of the largest eigenvalue in certain class of population covariance matrices is given in [12], along with a notion of isolation of eigenvalues. Let us finally note that as described in [4] in a slightly different context (and partially generalized in [5]), if the population largest eigenvalue is not isolated (or large) enough,

the largest empirical eigenvalue will asymptotically not be affected by its existence. In other words, in this situation, the existence of an isolated population eigenvalue is not detected by looking at the empirical largest eigenvalue, something that is potentially problematic in PCA.

Estimation of Covariance Matrices

Because of the increased availability of larger dimensional datasets, and in part because of recent results in random matrix theory indicating that in high dimensions (i.e., p large and of the same order of magnitude as n) the sample covariance matrix is not a good estimator (or proxy) for the population covariance, there has been quite a bit of activity recently in statistics to try to come up with better estimators of population covariance.

The classic paper [16] (see also [1, Section 7.8]) has shown that one could improve estimation of Σ (i.e., decrease the expected squared error of the corresponding estimator) by using linear combinations of $\hat{\Sigma}$ and any positive semidefinite matrices. These methods are often called *shrinkage* methods in statistics. More recently, Ledoit and Wolf [20] showed they could improve estimation of Σ by shrinking $\hat{\Sigma}$ toward the identity. Their method amounts to adding a certain quantity to all eigenvalues of $\hat{\Sigma}$ and keeping the same eigenvectors to get a new estimator of Σ , $\tilde{\Sigma}$. Finally, the paper [19], in the context of Markowitz portfolio optimization, proposed to use random matrix results to also shrink the eigenvalues of $\hat{\Sigma}$, this time in a nonlinear fashion; after noting the similarity between part of the histogram of eigenvalues of a sample correlation matrix computed from financial data and the density of the Marčenko–Pastur law, the authors of [19] proposed to leave unchanged the eigenvalues falling outside the Marčenko–Pastur-looking part of the histogram, and to change those falling within it to a common value. They showed that this method improved on their dataset the predictive performance of the efficient frontier they obtained when solving the Markowitz optimization problem with Σ replaced by their estimate.

Other nonlinear methods for nonlinear shrinking of eigenvalues using random matrix theory have been proposed (see [11, 24]). Improved estimation of covariance matrices is at the moment an active topic of research in statistics [6, 7, 10], and while

it is not obvious that the results obtained there are immediately applicable in the financial context, some of the insights found in these papers might prove helpful in the future.

References

- [1] Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*, Wiley Series in Probability and Statistics, 3rd Edition, Wiley-Interscience, John Wiley & Sons, Hoboken, NJ.
- [2] Bai, Z.D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review, *Statistica Sinica* **9**(3), 611–677. With comments by G. J. Rodgers and Jack W. Silverstein; and a rejoinder by the author.
- [3] Bai, Z.D. & Yin, Y.Q. (1993). Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix, *Annals of Probability* **21**(3), 1275–1294.
- [4] Baik, J., Ben Arous, G. & Pécché, S. (2005). Phase transition of the largest eigenvalue for non-null complex sample covariance matrices, *Annals of Probability* **33**(5), 1643–1697.
- [5] Baik, J. & Silverstein, J. (2006). Eigenvalues of large sample covariance matrices of spiked population models, *Journal of Multivariate Analysis* **97**(6), 1382–1408.
- [6] Bickel, P.J. & Levina, E. (2007). *Covariance Regularization by Thresholding*, Technical Report 744, Department of Statistics, UC Berkeley.
- [7] Bickel, P.J. & Levina, E. (2008). Regularized estimation of large covariance matrices, *The Annals of Statistics* **36**(1), 199–227.
- [8] El Karoui, N. Concentration of measure and spectra of random matrices: with applications to correlation matrices, elliptical distributions and beyond, *The Annals of Applied Probability* To Appear.
- [9] El Karoui, N. (2003). *On the largest eigenvalue of Wishart matrices with identity covariance when n , p and $p/n \rightarrow \infty$* . arXiv:math.ST/0309355, September 2003.
- [10] El Karoui, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices, *The Annals of Statistics* **36**(6), 2717–2756.
- [11] El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory, *The Annals of Statistics* **36**(6), 2757–2790.
- [12] El Karoui, N. (2007). Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices, *The Annals of Probability* **35**(2), 663–714.
- [13] Forrester, P.J. (1993). The spectrum edge of random matrix ensembles, *Nuclear Physics B* **402**(3), 709–728.
- [14] Geman, S. (1980). A limit theorem for the norm of random matrices, *Annals of Probability* **8**(2), 252–261.
- [15] Geronimo, J.S. & Hill, T.P. (2003). Necessary and sufficient condition that the limit of Stieltjes transforms is a Stieltjes transform, *Journal of Approximation Theory* **121**(1), 54–60.

-
- [16] Haff, L.R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix, *Annals of Statistics* **8**(3), 586–597.
 - [17] Johansson, K. (2000). Shape fluctuations and random matrices, *Communications in Mathematical Physics* **209**(2), 437–476.
 - [18] Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal component analysis, *Annals of Statistics* **29**(2), 295–327.
 - [19] Laloux, L., Cizeau, P., Bouchaud, J.-P. & Potters, M. (1999). Noise dressing of financial correlation matrices, *Physical Review Letters* **83**(7), 1467–1470.
 - [20] Ledoit, O. & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices, *Journal of Multivariate Analysis* **88**(2), 365–411.
 - [21] Marčenko, V.A. & Pastur, L.A. (1967). Distribution of eigenvalues in certain sets of random matrices, *Mathematics of the USSR-Sbornik* **72**(114), 507–536.
 - [22] Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979). *Multivariate Analysis, Probability and Mathematical Statistics: A Series of Monographs and Textbooks*, Academic Press [Harcourt Brace Jovanovich Publishers], London.
 - [23] Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, *Statistica Sinica* **17**(4), 1617–1642.
 - [24] Rao, N.R., Mingo, J., Speicher, R. & Edelman, A. (2008). Statistical Eigen-Inference from Large Wishart Matrices, *The Annals of Statistics* **36**(6), 2850–2885.
 - [25] Silverstein, J.W. (1989). On the weak limit of the largest eigenvalue of a large-dimensional sample covariance matrix, *Journal of Multivariate Analysis* **30**(2), 307–311.
 - [26] Silverstein, J.W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices, *Journal of Multivariate Analysis* **55**(2), 331–339.
 - [27] Tracy, C. & Widom, H. (1994). Level-spacing distribution and the Airy kernel, *Communications in Mathematical Physics* **159**, 151–174.
 - [28] Tracy, C. & Widom, H. (1996). On orthogonal and symplectic matrix ensembles, *Communications in Mathematical Physics* **177**, 727–754.
 - [29] Wachter, K.W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements, *Annals of Probability* **6**(1), 1–18.
 - [30] Yin, Y.Q., Bai, Z.D. & Krishnaiah, P.R. (1988). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix, *Probability Theory and Related Fields* **78**(4), 509–521.

NOUREDDINE EL KAROUI

Convexity Adjustments

The notion of convexity adjustment (or convexity correction) in fixed income markets arises when one uses prices of standard (plain vanilla) products corrected by an *adjustment* to price nonstandard products.

Many fixed income products are nonstandard with respect to aspects such as the timing, the currency, or the rate of payment. Examples of such adjustments can be found in approximations for pricing formulae of products such as in-arrears or in-advance products, quanto products, constant maturity swaps (CMS) products, or equity swaps. Despite their nonstandard features, these products are quite similar to plain vanilla ones whose price can either be directly obtained from the market or at least computed in closed form. Thus, adjustments of plain vanilla prices can be understood as corrections needed to capture the bias introduced in prices by the complexities of the nonstandard products. Taking this effect into account correctly could provide financial institutions with a competitive advantage. Examples of particular interest for practitioners include

- yield convexity adjustments;
- forward *versus* futures price adjustments;
- modified schedule or timing adjustments; and
- mismatch between currencies or quanto adjustments.

The yield convexity adjustment is somewhat unrelated to the remaining adjustments, but it is probably the “original one” in the sense that it is related to the nonlinear (and convex) relationship between bond prices and their yield to maturity. The three remaining adjustments have traditionally been separated, both by practitioners and academics, as they concern different classes of products. Various *ad hoc* rules have been proposed in the literature to calculate a variety of convexity adjustments for different products. Many of them are, however, mutually inconsistent. Nonetheless, as we explain later, the last three adjustments can be understood as corrections resulting from the fact that by using quotes of plain vanilla products to price nonstandard ones, we are taking expectations under the “wrong” martingale measure. Pelsser [14] was the first to put convexity adjustments on a firm

mathematical basis by showing that they can be interpreted as the side effect of changes in probability measures. To understand this connection, recall that in a stochastic interest rate setting, the no arbitrage price π , at time t , of any derivative paying $\Phi(T)$ at maturity T is given by

$$\begin{aligned}\pi(t, \Phi) &= \mathbb{E}_t^Q \left[e^{-\int_t^T r_u du} \Phi(T) \right] \\ &= p(t, T) \mathbb{E}_t^T [\Phi(T)]\end{aligned}\quad (1)$$

where $p(t, T)$, denotes the price at time t of a pure discount bond with maturity T and $\mathbb{E}_t^Q[\cdot]$, $\mathbb{E}_t^T[\cdot]$, denote expectation under the risk-neutral and the T -forward measure, respectively, conditional on the information available at time t . When dealing with fixed income derivatives, given the nature of the underlying asset, we are in a stochastic interest rate setting by definition, and, consequently, we are ultimately interested in computing, $\mathbb{E}_t^T[\Phi(T)]$, that is, the expected value, under the T -forward measure where the *numeraire* is $p(t, T)$, of our payoff. In the trivial cases, when our payoff turns out to be a T -forward martingale, we immediately get $\mathbb{E}_t^T[\Phi(T)] = \Phi(t)$. Unfortunately, this is *not* the case in most situations. Even if our payoff is not a martingale under the appropriate T -forward measure, it could be a martingale under *some* martingale measure, $\mathbb{Q}^\Phi$. This is obviously a restrictive assumption, but as it turns out, it includes many popular interest rate products. Suppose, therefore, our payoff is a martingale under the $\mathbb{Q}^\Phi$ martingale measure. Then we know $\Phi(t) = \mathbb{E}_t^\Phi[\Phi(T)]$, and we can define the convexity adjustment as the term CC^Φ for which the following holds:

$$\mathbb{E}_t^T[\Phi(T)] = \Phi(t) + CC^\Phi(t) \quad (2)$$

From the martingale properties of our payoff we get $\Phi(t) = \mathbb{E}_t^\Phi[\Phi(T)] = \mathbb{E}_t^T[\Phi(T)\Lambda(T)]$, where Λ denotes the Radon–Nikodym derivative (with $\Lambda(t) = 1$). Clearly, the cases where we can say something about $\mathbb{E}_t^T[\Phi(T)\Lambda(T)]$ are the cases where we can say something about convexity adjustment terms CC^Φ , which is nothing but a side effect of measure changes. For details concerning Radon–Nikodym derivatives and their relation to measure changes, we refer to [6]. In the following, we address each type of adjustment separately.

Yield Convexity Adjustments

Consider a pure discount bond with maturity T . This section is based on a particular “ T -bond”, and to ease up notation we always suppress “ T ”. That is, we simply use $p(t)$, $y(t)$, to denote the spot price and yield to maturity of the T -bond. One of the most commonly used convexity adjustments results from the nonlinear relationship between the bond’s price and its yield to maturity $p(t) = G(y(t))$. See **Bond**. The exact functional form of G depends on capitalization assumptions, but it is always a continuous and convex function. Given a concrete G , one can define the forward bond yield, $y_f(t, S)$, as the yield satisfying $f(t, S) = G(y_f(t, S))$, where $f(t, S)$ is the forward price of a contract written on the T -bond and maturing at time S . Without loss of generality, we consider $t \leq S \leq T$. One way to approximate the price at some future time S of the T -bond is to use a second-order Taylor expansion around the $y_f(t, S)$ yield, that is,

$$p(S) = f(t, S) + G'(y_f(t, S)) [y(S) - y_f(t, S)] + \frac{1}{2} G''(y_f(t, S)) [y(S) - y_f(t, S)]^2 \quad (3)$$

We can then take expectations under the S -forward measure to obtain

$$\begin{aligned} \mathbb{E}_t^S [p(S)] &\approx f(t, S) \\ &+ G'(y_f(t, S)) \mathbb{E}_t^S [y(S) - y_f(t, S)] \\ &+ \frac{1}{2} G''(y_f(t, S)) \mathbb{E}_t^S [y(S) - y_f(t, S)]^2 \end{aligned} \quad (4)$$

As $f(t, S) = \mathbb{E}^S [p(S)]$, we immediately obtain

$$\begin{aligned} \mathbb{E}_t^S [y(S)] &\approx y_f(t, S) \\ &- \frac{1}{2} \frac{G''(y_f(t, S))}{G'(y_f(t, S))} \mathbb{E}^S [y(S) - y_f(t, S)]^2 \end{aligned} \quad (5)$$

and the second term on the right-hand side (RHS) is what is understood as convexity adjustment in this context. It results from the fact that future yields of T -bonds are not today’s yields from forward contracts on T -bonds.

A common step to get the convexity adjustment in closed form is to further assume that $\mathbb{E}^S [(y(S) - y_f(t, S))^2] = \sigma^2 y(t)^2 (S - t)$. This adjustment was first proposed in [4, 9] and later sharpened in [8], but it is hard to justify on theoretical grounds. Note that $\mathbb{E}_t^S [y(S)] \neq y_f(t, S)$, so $\mathbb{E}_t^S [(y(S) - y_f(t, S))^2]$ cannot be understood as some sort of variance of future yields.

Moreover, the above derived convexity adjustment is an obvious consequence of the Jensen inequality on convex functions, which tells us that for convex G , $\mathbb{E}[G(y)] > G(\mathbb{E}[y])$. The Jensen inequality is true under any measure, so, in particular, one could also use futures contracts on T -bonds. As mentioned earlier, we could define the futures bond yield, $y_F(t, S)$, as the yield satisfying $F(t, S) = G(y_F(t, S))$ and using the fact that $F(t, S) = \mathbb{E}^Q [p(S)]$ we would get $\mathbb{E}_t^Q [y(S)] \approx y_F(t, S) - \frac{1}{2} [G''(y_F(t, S))/G'(y_F(t, S))] \mathbb{E}_t^Q [y(S) - y_F(t, S)]^2$. Once again, this is hard to interpret, but at least it would have the advantage that futures quotes may be observed directly from the market.

Finally, it is difficult to justify why the expansion should be around the forward bond yield $y_f(t, S)$ and not directly around today’s yield $y(t)$.

In our opinion, a more sensible way to study the relationship between a T -bond price and its yield to maturity is to derive it directly from the dynamics of bond prices. For any risk-neutral “ T -bond” dynamics $dp(t) = r_t p(t) dt + v_t p(t) dW_t^Q$, no arbitrage arguments allow us to obtain the *exact* relationship. For instance, in the continuously compounded yield case, we get

$$\mathbb{E}_t^Q [y(S)] = y(t) - \frac{p(t, S) - \mathbb{E}_t^Q \left[\frac{1}{2} \int_t^S v_u^2 du \right]}{T - S} \quad (6)$$

where $p(t, S)$, as above, denotes the price at time t of an S -bond. The second term on the left-hand side (LHS) can be understood as a convexity correction, and for any deterministic volatility of bond prices it can be computed in closed form. In particular, this is true for any affine model.

Forward/Futures Adjustments

A second type of convexity adjustment occurs when we are interested in quoting forward prices on any given underlying while having available futures prices on the same underlying. Recall (*see Forwards and Futures*) that the difference between futures and forward prices results from the correlation between the underlying to the financial contract and the spot interest rate. As mentioned in [12], this correlation, capitalized by the margin calls of the futures contract, leads to a more expensive (respectively cheaper) futures contract in the case of positive (respectively negative) correlation. Let us denote by $f(t, T)$ and $F(t, T)$ the forward and futures prices for contracts with maturity T on any underlying, respectively. It is a well-known fact that forward prices with maturity T are martingales under the T -forward martingale measure, while futures prices are martingales under the Q risk-neutral measure. For further reading on forward martingale measures, we refer to [1]. The difference in both prices occurs only in a stochastic interest rate setting and is therefore of particular importance when dealing with fixed income products and derivatives. For concrete examples on the difference between futures and forward term structures in a quite general setting (including any affine model), we refer to [5].

Timing Adjustments

This third cause of adjustment occurs when an interest rate derivative is structured so that it does not incorporate the natural time lag. Examples of products with modified schedule not only are obviously in-arrears products but also include the case of the CMS rate where the floating rate is in itself a swap rate with a certain maturity, usually higher than the time interval until the next payment. The market practice concerning schedule bias is to distinguish between LIBOR rates related convexity adjustments and swap rate convexity adjustments. We follow market practice and present them separately.

LIBOR Convexity Adjustments. The forward LIBOR rate $L(t, T_{i-1}, T_i)$ can be expressed in terms of discount bond prices as

$$L(t, T_{i-1}, T_i) = \frac{1}{\alpha_i} \frac{p(t, T_{i-1}) - p(t, T_i)}{p(t, T_i)} \quad (7)$$

where $\alpha_i = T_i - T_{i-1}$. Spot LIBOR rates can be defined in terms of the forward ones as $L(t, T) = L(t, t, T)$. From the above expression, one can easily see the forward LIBOR rate $L(t, T_{i-1}, T_i)$ is a martingale under the forward martingale measure $\mathbb{Q}^{T_i}$. Thus, we can price future LIBOR-dependent payments due at T_i , using equation (1) and forward LIBOR rates quoted today as we have $\mathbb{E}_t^{T_i}[L(T_{i-1}, T_i)] = L(t, T_{i-1}, T_i)$. This is the case in standard swap contracts, for instance.

However, in LIBOR-in-arrears (LIA) contracts, the LIBOR payment is due in advance, at time T_{i-1} . This leads to defining the LIA convexity correction as

$$\mathbb{E}_t^{T_{i-1}}[L(T_{i-1}, T_i)] = L(t, T_{i-1}, T_i) + CC^{\text{LIA}}(t) \quad (8)$$

In this particular case, the measure change that we are interested in occurs between two forward measures and we have

$$\begin{aligned} \mathbb{E}_t^{T_{i-1}}[L(T_{i-1}, T_i)] \\ = \mathbb{E}_t^{T_i} \left[L(T_{i-1}, T_i) \frac{1 + \alpha_i L(T_{i-1}, T_i)}{1 + \alpha_i L(t, T_{i-1}, T_i)} \right] \end{aligned} \quad (9)$$

which implies

$$\begin{aligned} CC^{\text{LIA}}(t) \\ = \alpha_i \frac{\mathbb{E}_t^{T_i}[(L(T_{i-1}, T_i))^2] - [L(t, T_{i-1}, T_i)]^2}{1 + \alpha_i L(t, T_{i-1}, T_i)} \\ = \frac{\alpha_i \text{Var}_t[L(T_{i-1}, T_i)]}{1 + \alpha_i L(t, T_{i-1}, T_i)} \end{aligned} \quad (10)$$

This is valid irrespective of the distribution that we assume for the forward LIBOR rates. Moreover, it only depends on its variance, which is the same under any equivalent martingale measure. Reference [15] provides exact LIA convexity adjustments based upon the assumption of lognormal LIBOR rates as in LIBOR market models (*see LIBOR Market Model*). In that particular case, we get $CC^{\text{LIA}}(t) = \alpha_i (L(t, T_{i-1}, T_i))^2 e^{\sigma^2(T_{i-1}-t)} / (1 + \alpha_i L(t, T_{i-1}, T_i))$, where σ represents the LIBOR volatility taken as constant. More generally, it is also possible to find exact LIA convexity adjustments in the context of affine term structure models.

Swap Convexity Adjustments. Since interest rate swaps make up a big proportion of the over-the-counter (OTC) derivatives market, convexity corrections needed for CMS are of particular interest and deserve separate attention. CMS payoffs and related products are discussed in detail in **CMS spread products**.

We start by recalling that the forward swap rate $S(t, T_n, T_N)$ quoted at time t , with start date T_n and maturity T_N is model-free and equal to

$$S(t, T_n, T_N) = \frac{p(t, T_n) - p(t, T_N)}{P(t, T_n, T_N)} \quad (11)$$

where $P(t, T_n, T_N) = \sum_{i=n}^N \alpha_i p(t, T_i)$ for α_i as before, is a portfolio of bonds, sometimes known as *swap level*. It is well known that swap rates are martingales under the so-called swap martingale measure, here denoted by Q^S , and whose *numeraire* is the swap level.

A CMS exchanges a swap rate with a certain maturity c against a fixed payment K . The difficulty lies in computing the value of each payment of the floating leg. By no arbitrage, the value at time t of the payment due at T_i is given by $\alpha_i p(t, T_i) \mathbb{E}_t^{T_i} [S(T_{i-1}, T_i, T_{i+c})]$ where, as before, $\mathbb{E}_t^{T_i}[\cdot]$ denotes conditional expectation under the T_i -forward measure. As $S(T_{i-1}, T_i, T_{i+c})$ is not a martingale under forward measures, we need a convexity adjustment to take care of this. The convexity adjustment for each payment of a CMS can be defined as

$$\mathbb{E}_t^{T_i} [S(T_{i-1}, T_i, T_{i+c})] = S(t, T_i, T_{i+c}) + CC_t^{\text{CMS}}(t) \quad (12)$$

and it relies on measure changes from the T_i -forward measure to Q^S . Concretely we get,

$$\begin{aligned} CC_t^{\text{CMS}} &= \frac{P(t, T_i, T_{i+c})}{p(t, T_i)} \\ &\times \mathbb{E}_t^S \left[S(T_{i-1}, T_i, T_{i+c}) \frac{p(T_{i-1}, T_i)}{P(T_{i-1}, T_i, T_{i+c})} \right] \\ &- S(t, T_i, T_{i+c}) \end{aligned} \quad (13)$$

This is not easy to obtain in closed form. Most CMS convexity adjustment formulas can only be obtained in an approximative way. A common practice at this

stage is to see $p(t, T_i)/P(t, T_i, T_{i+c})$ as a function G of the swap rate $S(t, T_i, T_{i+c})$.

In the case of the linear swap model (LSM) introduced in [10], G is linear and we have $p(t, T_i)/P(t, T_i, T_{i+c}) = G(S(t, T_i, T_{i+c})) = A + B(T_i)S(t, T_i, T_{i+c})$ for $A = (\sum_i \alpha_i)^{-1}$ a constant and $B(T_i) = [p(t, T_i)/P(t, T_i, T_{i+c}) - A]/S(t, T_i, T_{i+c})$ only dependent on T_i . This linearity assumption is a crude one, but it yields better results in practice than one would expect. The reason for this is that convexity adjustments only become sizable for long maturities, and, for those maturities the term structure movements tend to be parallel. Thus, they are well captured by a linear model.

Alternatively, we can consider a generic (truly convex) function G , expand it around today's swap rate $S(t, T_i, T_{i+c})$ using a Taylor expansion to then take expectations. Reference [7] is a popular paper on convexity adjustments of CMS swaps, caps, and floors that follows this line. The author analyzes three basic term structure models, starting from a standard yield curve model and continuing to more sophisticated models, which account for nonparallel shifts and continuous time compounding of interest rates. This is equivalent to assuming particular functional forms for the G function. Given a functional for G , the author then uses a first-order Taylor approximation and a lognormal distribution for the swap rates to approximations for the convexity adjustments of CMS

$$\begin{aligned} CC_t^{\text{CMS}} &\approx S(t, T_i, T_{i+c}) \frac{G'(S(t, T_i, T_{i+c}))}{G(S(t, T_i, T_{i+c}))} \\ &\times \left[S(t, T_i, T_{i+c}) e^{\sigma^2(T_{i-1}-t)} - 1 \right] \end{aligned} \quad (14)$$

where σ can be taken to be [2] implied volatility from at-the-money vanilla swaptions market quotes. For convexity adjustments of CMS caps and floors, we refer to the paper itself. For more information on popular formulation of swap rate dynamics, see **Swap Market Models**.

The method proposed in [7] includes the LSM as a special case, in which the first-order Taylor approximation is, in fact, exact. Depending on the underlying dynamics for the swap rate, the formulas may be quite evolved, but the paper [16] provides hope for nice formulas in the affine term structure setup.

Quanto Adjustments

So far, we have discussed timing adjustments. Quanto products refer to products with an underlying expressed in one currency and payment in a different currency. Quanto adjustments are adjustments that try to correct for payment in the “wrong currency” *versus* the “wrong timing”. We can have quanto (diff) swaps, swaptions, caps, floors, CMS rates, barrier options, equity swaps, and so on. For an overview of existing quanto products, see **Quanto Options**.

As explained in [14], a convexity adjustment of this type occurs when our product is a martingale under the foreign risk-neutral measure, but needs to be priced under the domestic risk-neutral measure. As an example, we shall use a diffed LIBOR contract, that is, a contract where the foreign LIBOR rate $L^f(T_{i-1}, T_i)$ is observed at T_{i-1} and is paid in domestic currency at time T_i . This contract can serve as the floating payment for a differential swap (also known as *diff swap* or *quanto swap*), and it is made against a fixed payment in domestic currency. We know that the forward LIBOR rate $L^f(t, T_{i-1}, T_i)$ is a martingale under the foreign T_i forward measure Q^{f, T_i} , so $L^f(t, T_{i-1}, T_i) = \mathbb{E}_t^{f, T_i}[L^f(T_{i-1}, T_i)]$, but we need to compute the expectation of the foreign LIBOR rate under the domestic forward measure. Under that measure we know

$$\mathbb{E}_t^{d, T_i}[L^f(T_{i-1}, T_i)] = L^f(t, T_{i-1}, T_i) + CC^{\text{diff}}(t) \quad (15)$$

and this type of convexity adjustment is given by

$$CC^{\text{diff}}(t) = \frac{\mathbb{E}_t^{f, T_i}[L^f(T, T_i)F(T_{i-1}, T_i)]}{F(t, T_i)} - L^f(t, T_{i-1}, T_i) \quad (16)$$

where $F(t, T_i)$ is the forward exchange rate (f/d), quoted at t and with delivery at T_i . The convexity correction for quanto products will typically depend on the correlation between the foreign interest rate and the exchange rate. If, in the above example, both the forward LIBOR and the forward exchange rates have lognormal distributions under Q^{f, T_i} as proposed in [11], one can easily derive the convexity adjustment formula as $CC^{\text{diff}}(t) = L^f(t, T_{i-1}, T_i) [e^{\rho_{F,L}\sigma_L\sigma_F(T_{i-1}-t)} - 1]$, where $\rho_{F,L}$ is the instant correlation between the LIBOR and the forward exchange rate, σ_L is the volatility of the LIBOR,

and σ_F is the volatility of the forward exchange rate, assumed to be constant.

References

- [1] Björk, T. (2004). *Arbitrage Theory in Continuous Time*, 2nd Edition, Oxford University Press.
- [2] Black, F. (1976). The pricing of commodity contracts, *Journal of Financial Economics* **3**(2), 167–179.
- [3] Brigo, D. & Mercurio, F. (2006). *Interest Rate Models: Theory and Practice*, 2nd Edition, Springer Finance, Heidelberg.
- [4] Brotherton-Ractcliff, R. & Iben, B. (1993). Yield curve applications of swap products, in *Advanced Strategies in Financial Risk Management*, R. Schwartz & C. Smith, eds, New York Institute of Finance, New York, pp. 400–450.
- [5] Gaspar, R.M. (2006). *Credit Risk and Forward Price Models*, EFI – The Economic Research Institute, Stockholm.
- [6] Geman, H., El-Karoui, N. & Rochet, J.-C. (1995). Changes of numeraire, changes of probability measure and option pricing, *Journal of Applied Probability* **32**, 443–458.
- [7] Hagan, P.S. (2003). Convexity conundrums: pricing CMS swaps, caps, and floors, *Wilmott magazine* March, 38–44.
- [8] Hart, Y. (1997). Unifying theory, *Risk* February, 54–55.
- [9] Hull, J.C. (2006). *Options, Futures, and other Derivative Securities*, Prentice Hall International, Inc. 6.
- [10] Hunt, P. & Kennedy, J. (2000). *Financial Derivatives in Theory and Practice*, 6th Edition, John Wiley & Sons, Chichester.
- [11] Hunt, P. & Pelsser, A. (1998). Arbitrage free pricing of quanto-swaptions, *Journal of Financial Engineering* **7**(1), 25–33.
- [12] Park, H.Y. & Chen, A.H. (1985). Differences between futures and forward prices: a further investigation of marking to market effects, *Journal of Future Markets* **5**, 77–88.
- [13] Pelsser, A. (2000). *Efficient Methods for Valuing Interest Rate Derivatives*, Springer Finance, Heidelberg.
- [14] Pelsser, A. (2003). Mathematical foundation of convexity correction, *Quantitative Finance* **3**, 59–65.
- [15] Pugachevsky, D. (2001). Forward CMS rate adjustment, *Risk* March, 125–128.
- [16] Schrager, D. & Pelsser, A. (2006). Pricing swaptions and coupon bond options in affine term structure models, *Mathematical Finance* **16**(4), 673–694.

Further Reading

References [13] and [3] are two important textbook references for fixed income products. Both books present good reviews

of the nonstandard products that lead to convexity adjustments, including all products discussed here as well as more exotic ones.

Related Articles

Affine Models; Bond; Caps and Floors; CMS Spread Products; Constant Maturity Swap;

Eurodollar Futures and Options; Forward and Swap Measures; LIBOR Market Model; LIBOR Rate; Swap Market Models; Yield Curve Construction.

RAQUEL M. GASPAR & AGATHA MURGOCI

Algorithmic Trading

Technology continues to have an increasingly significant impact on how securities are traded in today's markets. Many trading floors have been replaced by electronic trading platforms and more than a third of the trading volume in the United States can be attributed to algorithmic trading. Every large broker-dealer provides algorithmic trading services to their institutional clients in order to assist their trading. The algorithms used by institutional investors, hedge funds, and many other market participants are used to make trading decisions about the timing, price, and size of trades, with the objective of reducing risk-adjusted costs.

In a broad sense, the term *algorithmic trading* is used to describe trading in an automated manner according to a set of rules. It is often used interchangeably with statistical trading or statistical arbitrage, which may or may not be automated, but is based on signals derived from statistical analyses or models. Smart order routing, program trading, and rules-based trading are some of the other terms associated with algorithmic trading. More recently, the range of functions and activities associated with algorithmic trading has grown to include market impact modeling, execution risk analytics, cost aware portfolio construction, and the use of market microstructure effects.

In this article, we first explain the basic ideas of market impact and optimal execution from both the sell- and buy-side perspectives. We then provide an overview of the most popular algorithmic trading strategies.

Market Impact and the Order Book

The limit order book contains resting limit orders. These orders rest in the book and provide liquidity as they wait to be matched with nonresting orders, which represent a demand for liquidity. The three most common types of nonresting orders are marketable limit orders, market orders, and fill-or-kill orders.

The bid side of the limit book contains resting bids to buy a certain number of shares of stock at a certain price. The offer side contains resting offers to sell a certain number of shares of stock at a certain price.

A market order is a demand for an immediate execution of a certain number of shares at the best possible price. To get the best possible price, a market order sweeps through one side of the limit order book—starting with the best price—matching against resting orders until the full quantity of the market order is filled or the book is completely depleted.

Unlike a market order, a marketable limit order can be executed only at a specified price or better. For example, a marketable limit order to buy 100 shares at \$90.01 can match with a resting limit order to sell 200 shares at \$90.00. The trade print—the price at which the trade would take place—would be \$90.00.

The following examples illustrate how market orders to sell interact with resting limit orders to buy.

Figure 1 shows the idealized market impact of a 200-share market order to sell. The x - and y -axes display the time and price, respectively.

The bid side of the limit order book contains bids to buy a certain number of shares of stock at a certain price. Resting limit orders—orders that sit in the order book—are said to *provide liquidity* by mitigating the market impact of orders that must be filled immediately. The state of the book establishes a pretrade equilibrium (1), which is disturbed by a market order to sell 200 shares (2). Market orders must be filled immediately, and therefore represent a demand for liquidity.

As the sell order depletes the bid book by matching with limit orders to buy, it obtains an increasingly less favorable (lower) trade price, resulting in the trade print (3). Assuming no other trading activity, over time, liquidity providers replenish the bid book to (4), which is the posttrade equilibrium.

The difference between (4) and (1) is an information-based effect called *permanent market impact*. It is the market's response to information that a market participant has decided not to own 200 shares of this stock. This effect is typically modeled as immediate and linear in the total number of shares executed. Huberman and Stanzl [9] show that, if the effect were not linear and immediate, buying and selling at two different rates could produce an arbitrage profit.

The difference between (4) and (3) is called *temporary market impact*. The trader who initiated the trade is willing to obtain a less favorable fill price (3) to get his/her trade done immediately. This *cost*

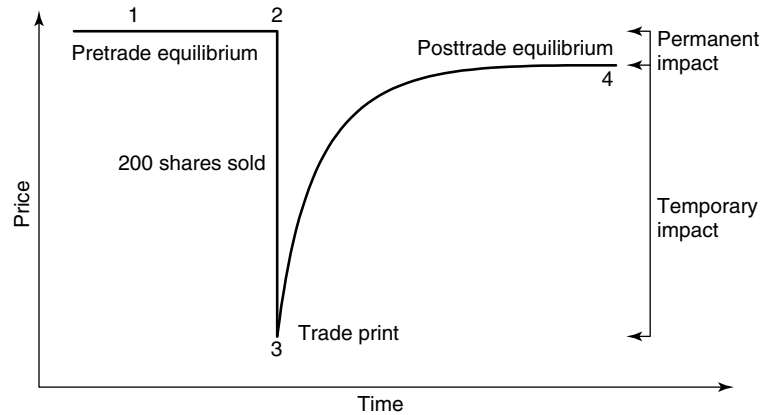


Figure 1 Idealized market impact model showing sell of 200 shares

of immediacy is typically modeled as linear or square root. Under the assumption of square root impact, with all other factors held constant, a trade of 200 shares executed over the same period of time as a trade of 100 shares would have square root of two times more temporary impact per share.

Figure 2 shows what would happen if the same trader were willing to wait some time between trades. The trade print from the previous figure is shown as a reference point (1). As in Figure 1, a pretrade equilibrium (2) is disturbed by a 100-share market order to sell (3). As the market order depletes the bid book by matching with limit orders to buy, it obtains a fill price (4). Over time (5), liquidity providers refill the bid book with limit orders to buy. However, the new posttrade equilibrium (6) is lower

than the pretrade equilibrium because it incorporates the information of the executed market order.

Our trader then places another market sell order for 100 shares (6) and obtains a trade print (7). Over time, the temporary impact—(8) minus (7)—decays and results in a new posttrade equilibrium (8). As the permanent impact is assumed to be linear and immediate, the posttrade equilibrium is shown to be the same for one order of 200 shares as it is for two orders of 100 shares each.

Optimal Execution

While our trader waits between trades (5), he/she incurs price risk—the risk that his/her execution will

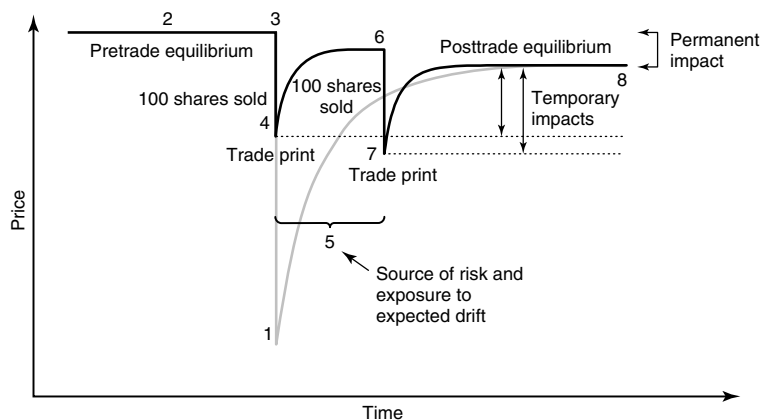


Figure 2 Idealized market impact model showing two sells of 100 shares each

be less favorable due to the random movement of prices. In this context, a *shortfall* is the difference between the effective execution price and the arrival price—the prevailing price at the start of the execution period. If we use the variance of *shortfalls* as a proxy for risk, a trader's aversion to risk establishes a risk/cost trade-off. In the first scenario, he/she pays a higher cost—the difference between (8) and (1)—to eliminate risk. In the second scenario, he/she pays a lower cost—the average of the differences between (8) and (4), and (8) and (7)—but takes on a greater dispersion of shortfalls associated with the waiting time between trades (5). This is the trade-off considered in the seminal paper of Almgren and Chriss [1].

Risk aversion increases a trader's sense of urgency and makes it attractive to pay some premium to reduce risk. The premium the trader pays is in the form of higher temporary market impact. All other factors held constant, a higher expected temporary market impact encourages slower trading, while a higher expected risk or risk aversion encourages faster trading.

Risk aversion embodies the notion that people dislike risk. For a risk-averse agent, the utility of a fair game, $u(G)$, is less than the utility of having the expected value of the game, $E(u(G))$, with certainty. The degree of risk aversion may be captured by the risk aversion parameter λ , which is used to translate risk into a *certain dollar cost equivalent*—the smallest certain dollar amount that would be accepted instead of the uncertain payoff from the fair game. For an agent with quadratic utility, the certain dollar cost equivalent is given by $E(G) - \lambda \text{Var}(G)$. Hence, his/her degree of risk aversion is characterized by the family of risk/return pairs with the same constant trade-off between the expected return and risk. An annualized target return and standard deviation imply a risk aversion, and may be translated to a risk aversion parameter of the type used in some optimal execution algorithms.

Another factor that influences the decision to trade more quickly or more slowly is the expectation of price change. For the purpose of execution, a *positive alpha* is an expectation of profits per share per unit time for unexecuted shares. A faster execution captures more of the profits associated with this expectation of price change. A *negative alpha* is the expectation of losses per share per unit time for unexecuted shares. A slower execution incurs less of

the losses associated with this expectation of price change.

For example, a trader has positive alpha if he/she expects prices to move lower while he/she is executing his/her sell orders. He/she may choose to front-weight his/her trade schedule—execute more rapidly at the beginning of the execution period—to obtain better execution prices. Similarly, a seller who believes that prices are moving higher may back-weight his/her trade schedule or delay the execution.

The general form of the optimal execution problem involves finding of the best trade-off between the effects of risk, market impact, and alpha by minimizing risk-adjusted costs relative to a prespecified benchmark. Common benchmarks are volume-weighted average price (VWAP) and arrival price (the price prevailing at the beginning of the execution period).

The first formulations of this problem go back to the seminal papers of Bertsimas and Lo [4] and Almgren and Chriss [1]. Assuming a quadratic utility function, a general formulation of this problem takes the form

$$\min_{x_t} [E(C(x_t)) + \lambda \text{Var}(C(x_t))] \quad (1)$$

where $C(x_t)$ is the cost of deviating from the benchmark. The solution is given by the trade schedule x_t that represents the number of shares that remains to buy/sell at time t . The trader's optimal trade schedule is a function of his/her level of risk aversion, $\lambda \geq 0$, which determines his/her urgency to trade, and dictates the preferred trade-off between execution cost and risk.

In the following two subsections, we describe the sell-and buy-side perspectives of the typical arrival price optimal execution models.

The Sell-side Perspective

The typical optimal execution model uses arrival price as a benchmark and balances the trade-off between market impact, price risk, and opportunity cost. Alpha is assumed to be greater than or equal to zero, which means that delaying execution may carry an associated opportunity cost, but does not carry an expectation of profit. The optimal strategy lies somewhere between two extremes: (i) trade everything immediately at a known cost or (ii) reduce market

impact by spreading the order into smaller trades over a longer horizon at the expense of increased price risk and opportunity cost.

Bertsimas and Lo [4] proposed an algorithm for the optimal execution problem that finds the minimum expected cost of trading over a fixed period of time for a risk neutral trader, $\lambda = 0$, facing an environment where price movements are assumed to be serially uncorrelated.

Almgren and Chriss [1] extended this concept using quadratic utility to embody the trade-off between the expected cost and price risk. The more aggressive (passive) trade schedules incur higher (lower) market impact costs and lower (higher) price risk. Similar to classical portfolio theory, as λ varies the resulting set of points $(Var(\lambda), E(\lambda))$ traces out the *efficient frontier of optimal trading strategies*. The two extreme cases $\lambda = 0$ and $\lambda \rightarrow \infty$ correspond to the minimum impact strategy—trading at a constant rate throughout the execution period—and the minimum variance strategy—a single execution of the entire target quantity at the start of the execution period.

Let us consider selling X shares, that is, we want $x_0 = X$ and $x_T = 0$. Under the assumptions that asset prices follow an arithmetic Brownian motion, permanent impact is immediate and linear in total shares executed, and temporary impact is linear in the rate of trading, the solution of the Almgren and Chriss model is

$$x_t = X \frac{\sinh(\kappa(T - t))}{\sinh(\kappa T)} \quad (2)$$

where $\kappa = \sqrt{\frac{\lambda\sigma^2}{\eta}}$. Here, σ and η represent stock volatility and linear temporary market impact cost.

Note that the solution is effectively a decaying exponential $X \exp(-\kappa t)$ adjusted such that $x_T = 0$. It does not depend on the permanent market impact, consistent with the discussion in the previous section. The urgency of trading is embodied in κ . This parameter determines the speed of liquidation independent of the order size X . For a higher risk aversion parameter or volatility—for example, representing increased perceived risk—the speed of trading increases as well. We also see that for a higher expected temporary market impact cost, the speed of trading decreases.

Impact Models

An impact model is used to predict changes in price due to trading activity. This expectation of price change may be used to inform execution and portfolio construction decisions. Several well-known models have been proposed. For example, see Hasbrouck [8], Lillo *et al.* [12], and Almgren *et al.* [3].

Almgren *et al.* use a proprietary data set obtained from Citigroup's equity trading desk in which a trade's direction (buyer or seller initiated) is known. Note that for most public data sets, trade direction is not available and has to be estimated by a classification algorithm. Classification errors in algorithms such as Lee and Ready [11], and Ellis *et al.* [5] introduce a bias that produces an overestimate of the true trading cost.

In Almgren *et al.*, trades serve as a proxy for trading imbalance. The authors assume that, some time after the complete execution of a parent order, only permanent impact remains. This allows them to separate impact into its temporary and permanent components.

The model parameters can then be calculated from a regression, giving the following results. First, permanent impact cost is linear in trade size and volatility. Second, temporary impact cost is linear in volatility and roughly proportional to the square root—Almgren *et al.* find a power 3/5—of the fraction of volume represented by one's own trading during the period of execution. Hence, for a given rate of trading, a less volatile stock with large average daily volume has the lowest temporary impact costs.

The Buy-side Perspective

Optimal execution algorithms have less value to a typical portfolio manager if analyzed separately from the corresponding returns earned by his/her trading strategy. In fact, high transaction costs are not bad *per se*—they could simply prove to be necessary for generating superior returns. At present, the typical sell-side perspective of algorithmic trading does not take expectation of profits or the client's portfolio objectives into account. Needless to say, this is an important component of execution.

The decisions of the trader and the portfolio manager are based on different objectives. The trader decides on the timing of the execution, breaking

large parent orders into a series of child orders that, when executed over time, represent the correct trade-off between opportunity cost, market impact, and risk. The trader sees only the trading assets, whereas the portfolio manager sees the entire portfolio, which includes both, the trading assets and the static—nontrading—positions.

The portfolio manager's task is to construct a portfolio by optimizing the trade-off between the opportunity cost, market impact, and risk for the full set of trading and nontrading assets. In general, the optimal execution framework described by Almgren and Chriss is not appropriate for the portfolio manager.

Engle and Ferstenberg [6] proposed a framework that unites these objectives by combining optimal execution and classical mean–variance optimization models. In their model, trading takes place at discrete time intervals as the portfolio manager rebalances his/her portfolio holdings w_t at times $t = 0, 1, \dots, T$ subject to changing expected returns, μ_t , and risk (as measured by the covariance matrix of returns), Ω_t , until he/she reaches the portfolio that reflects his/her final view $w_T = \frac{1}{2\lambda} \Omega_T^{-1} \mu_T$. The joint dynamic optimization problem has the form

$$\begin{aligned} \max_{\{w_t\}} \left\{ \sum_{t=1}^T (w_t^T \mu_t - \lambda w_t^T \Omega_t w_t) \right. \\ \left. - \sum_{t=1}^T \{ \Delta w_t^T \tau_t + (w_t - w_{t-1}) \mu_t \right. \\ \left. + \lambda (w_t - w_{t-1}) \Omega_t (w_t - w_{t-1}) \right\} \\ \left. + 2\lambda \sum_{t=1}^T (w_t - w_{t-1}) \Omega_t w_t \right\} \quad (3) \end{aligned}$$

where $\tau_t = \tau_t(\Delta w_t)$ is the temporary market impact function (for simplicity of exposition, we ignore permanent impacts). This is a dynamic programming problem that has to be solved by numerical techniques.

Each one of the three terms in the objective function above has an intuitive interpretation. The first term represents the standard mean–variance optimization problem. The second term corresponds to the optimal execution problem. The third term is the covariance between the remaining shares to be traded and the final position. In the single asset case, the third term is positive (negative) for buying

(selling) orders, which implies that risk is reduced (increased). If this term is ignored, which occurs when portfolio allocation and optimal execution are performed separately, then the measurement of total risk is biased.

Popular Algorithmic Trading Strategies

A small number of execution strategies have become *de facto* standards and are offered by most technology providers, banks, and institutional broker/dealers. However, even among these standards, the large number of input parameters makes it difficult to compare execution strategies directly.

Typically, a strategy is motivated by a *theme* or *style* of trading. The objective is to minimize either absolute or risk-adjusted costs relative to a *benchmark*. For strategies with mathematically defined objectives, an *optimization* is performed to determine how to best use the strategy to maximize a trader's or portfolio manager's utility. A *trade schedule*—or *trajectory*—is planned for strategies with a target quantity of shares to execute. The *order placement* engine—sometimes called the *micro-trader*—translates from a strategy's broad objectives to individual orders. User-defined *input parameters* control the trade schedule and order placement strategy.

In this section, we review some of the most common algorithmic trading strategies.

Volume-weighted Average Price

Six or seven years ago, the VWAP execution strategy represented the bulk of algorithmic trading activity. Currently, it is second in popularity only to arrival price. The appeal of benchmarking to VWAP is that the benchmark is easy to compute and intuitively accessible.

The typical parameters of a VWAP execution are the start time, the end time, and the number of shares to execute. Additionally, optimized forms of this strategy require a choice of risk aversion.

The most basic form of VWAP trading uses a model of the fractional daily volume pattern over the execution period. A trade schedule is calculated to match this volume pattern. For example, if the execution period is one day, and 20% of a day's volume is expected to be transacted in the first hour,

a trader using this basic strategy would trade 20% of his/her target accumulation or liquidation in the first hour of the day. Since the daily volume pattern has a U shape—with more trading in the morning and afternoon and less in the middle of the day—the volume distribution of shares executed in a VWAP pattern would also have this U shape.

VWAP is an ideal strategy for a trader who meets all of the following criteria:

- his/her trading has little or no alpha during the execution period;
- he/she is benchmarked against the VWAP;
- he/she believes that market impact is minimized when his/her own rate of trading represents the smallest possible fraction of all trading activity; and
- he/she has a set number of shares to buy or sell.

Deviation from these criteria may make VWAP strategies less attractive. For example, market participants who trade over the course of a day and have strong positive alpha may prefer a front-weighted trajectory, such as those that are produced by an arrival price strategy.

The period of a VWAP execution is most typically a day or a large fraction of a day. Basic VWAP models predict the daily volume pattern using a simple historical average of fractional volume. Several weeks to several months of data are commonly used. However, this forecast is noisy. On any given day, the actual volume pattern deviates substantially from its historical average, complicating the strategy's objective of minimizing its risk-adjusted cost relative to the VWAP benchmark. Some models of fractional volume attempt to increase the accuracy of volume pattern prediction by making dynamic adjustments to the prediction based on observed trading results throughout the day.

Several variations of the basic VWAP strategy are common. The ideal VWAP user (as defined earlier) can lower his/her expected costs by increasing his/her exposure to risk relative to the VWAP benchmark. For example, assuming an alpha of zero, placing limit orders throughout the execution period and catching up to a target quantity with a market order at the end of the execution period will lower the expected cost while increasing risk. This is the highest risk strategy. Continuously placing small market orders in the fractional volume pattern is the lowest risk strategy,

but has a higher expected cost. For a particular choice of risk aversion, somewhere between the highest and lowest risk strategies, is a compromise optimal strategy that perfectly balances risk and costs.

For example, a risk-averse VWAP strategy might place one market order of 100 shares every 20 s, whereas a less risk-averse strategy might place a limit order of 200 shares, and, 40 s later, place a market order for the difference between the desired fill of 200 and the actual fill (which may have been smaller). The choice of the average time between market orders in a VWAP execution implies a particular risk aversion.

For market participants with a positive alpha, a frequently used rule-of-thumb optimization is compressing trading into a shorter execution period. For example, a market participant may try to capture more profits by doing all of his/her VWAP trading in the first half of the day instead of taking the entire day to execute.

In another variant of VWAP—*guaranteed VWAP*—a broker commits capital to guarantee his/her client the VWAP price in return for a pre-determined fee. The broker takes on a risk that the difference between his/her execution and VWAP will be greater than the fee he/she collects. If institutional trading volume and individual stock returns were uncorrelated, the risk of guaranteed VWAP trading could be diversified away across many clients and many stocks. In practice, managing a guaranteed VWAP book requires some complex risk calculations that include modeling the correlations of institutional trading volume.

Time-weighted Average Price

The time-weighted average price (TWAP) execution strategy attempts to minimize market impact costs by maintaining an approximately constant rate of trading over the execution period. With only a few parameters—start time, end time, and target quantity—TWAP has the advantage of being the simplest execution strategy to implement. As with VWAP, optimized forms of TWAP may require a choice of risk aversion. Typically, the VWAP or arrival price benchmarks are used to gauge the quality of a TWAP execution. TWAP is hardly ever used as its own benchmark.

The most basic form of TWAP breaks a parent order into small child orders and executes these child orders at a constant rate. For example, a parent order

of 300 shares with an execution period of 10 min could be divided into three child orders of 100 shares each. The child orders would be executed at the 3:20, 6:40, and 10:00 min marks. Between market orders, the strategy may place limit orders in an attempt to improve execution quality.

An ideal TWAP user has almost the same characteristics as an ideal VWAP user, except that he/she believes that the lowest *trading rate*—not the lowest *participation rate*—incurs the lowest market impact costs.

TWAP users can benefit from the same type of optimization as VWAP users by placing market orders less frequently, and using resting limit orders to attempt to improve execution quality.

Participation

The participation strategy attempts to maintain a constant fractional trading rate. That is, its own trading rate as a fraction of the market's total trading rate should be constant throughout the execution period. If the fractional trading rate is maintained exactly, participation strategies cannot guarantee a target fill quantity.

The parameters of a participation strategy are the start time, end time, fraction of market volume the strategy should represent, and max number of shares to execute. If the max number of shares is specified, the strategy may complete execution before the end time. Along with VWAP and TWAP, participation is a popular form of nonoptimized strategies, though some improvements are possible with optimization.

VWAP and arrival price benchmarks are often used to gauge the quality of a participation strategy execution. The VWAP benchmark is particularly appropriate because the volume pattern of a perfectly executed participation strategy is the market's volume pattern during the period of execution. An ideal user of participation strategies has all of the same characteristics as an ideal user of VWAP strategies, except that he/she is willing to forego certain execution to maintain the lowest possible fractional participation rate.

Participation strategies do not use a trade schedule. The strategy's objective is to participate in volume as it arises. Without a trade schedule, a participation strategy cannot guarantee a target fill quantity. The most basic form of participation strategies waits for

trading volume to show up on the tape, and follows this volume with market orders. For example, if the target fractional participation rate is 10%, and an execution of 10 000 shares is shown to have been transacted by other market participants, a participation strategy would execute 1000 shares in response.

Unlike a VWAP trading strategy, which for a given execution may experience large deviations from an execution period's actual volume pattern, participation strategies can closely track the actual—as opposed to the predicted—volume pattern. However, close tracking has a price. In the earlier example, placing a market order of 1000 shares has a larger expected market impact than slowly following the market's trading volume with smaller orders. An optimized form of the participation strategy amortizes the trading shortfall over some period of time. Specifically, if an execution of 10 000 shares is shown to have been transacted by other market participants, instead of placing 1000 shares all at once, a 10% participation strategy might place 100 share orders over some period of time to amortize the shortfall of 1000 shares. The result is a lower expected shortfall, but a higher dispersion of shortfalls.

Market-on-close

The market-on-close strategy is popular with market participants who either want to minimize risk-adjusted costs relative to the closing price of the day or want to manipulate—*game*—the close to create the perception of a good execution. The ideal market-on-close user is benchmarked to the close of the day and has low or negative alpha. The parameters of a market-on-close execution are the start time, the end time, and the number of shares to execute. Optimized forms of this strategy require a risk-aversion parameter.

When market-on-close is used as an optimized strategy, it is similar in its formulation to an arrival price strategy. However, with market-on-close, a back-weighted trade schedule incurs less risk than a front-weighted one. With arrival price, an infinitely risk averse trader would execute everything in the opening seconds of the execution period. With market-on-close, an infinitely risk averse trader would execute everything at the closing seconds of the day. For typical levels of risk aversion, some trading would take place throughout the execution period. As with arrival price optimization, positive alpha

increases urgency to trade and negative alpha encourages delayed execution.

In the past, market-on-close strategies were used to manipulate—or *game*—the close, but this has become less popular as the use of VWAP and arrival price benchmarks have increased. Gaming the close is achieved by executing rapidly near the close of the day. The trade print becomes the closing price or very close to it, and hence shows little or no shortfall from the closing price benchmark. The true cost of the execution is hidden until the next day when temporary impact dissipates and prices return to a new equilibrium.

Arrival Price

The arrival price strategy—also called the *implementation shortfall* strategy—attempts to minimize risk-adjusted costs using the arrival price benchmark. Arrival price optimization is the most sophisticated and popular of the commonly used algorithmic trading strategies.

The ideal user of arrival price strategies has the following characteristics:

- he/she is benchmarked to the arrival price;
- he/she is risk averse and knows his/her risk aversion parameter;
- he/she has high positive or high negative alpha; and
- he/she believes that market impact is minimized by maintaining a constant rate of trading over the maximum execution period while keeping trade size small.

Most implementations are based on some form of the risk-adjusted cost minimization introduced by Almgren and Chriss [1]. In the most general terms, an arrival price strategy evaluates a series of trade schedules to determine which one minimizes risk-adjusted costs relative to the arrival price benchmark. As discussed in the section on optimal execution, under certain assumptions, this problem has a closed form solution.

The parameters in an arrival price optimization are alpha, number of shares to execute, start time, end time, and a risk aversion parameter. For buyers (sellers), positive (negative) alpha encourages faster trading. For both buyers and sellers, risk encourages faster trading, while market impact costs encourage slower trading.

For traders with positive alpha, the feasible region of trade schedules lies between the immediate execution of total target quantity and a constant rate of trading throughout the execution period.

A more general form of arrival price optimization allows for both the buyers and sellers to have either positive or negative alpha. For example, under the assumption of negative alpha, shares held long and scheduled for liquidation are—without considering one's own trading—expected to go up in price over the execution period. This would encourage a trader to delay execution or stretch out trading. Hence, the feasible region of solutions that account for both positive and negative alpha includes back-weighted as well as front-weighted trade schedules.

Other factors that necessitate back-weighted trade schedules in an arrival price optimization are expected changes in liquidity and expected crossing opportunities. For example, an expectation of a cross later in the execution period may provide enough cost savings to warrant taking on some price risk and the possibility of a compressed execution if the cross fails to materialize. Similarly, if market impact costs are expected to be lower later in the execution period, a rational trader may take on some risk to obtain this cost savings.

A variant of the basic arrival price strategy is *adaptive arrival price*. A favorable execution may result in a windfall in which an accumulation of a large number of shares takes place at a price significantly below the arrival price. This can happen by random chance alone. Almgren and Lorenz [2] demonstrated that a risk-averse trader should use some of this windfall to reduce the risk of the remaining shares. He/she does this by trading faster and thus incurring a higher market impact. Hence, the strategy is adaptive in that it changes its behavior based on how well it is performing.

Crossing

Though crossing networks have been around for some time, their use in algorithmic trading strategies is a relatively recent development. The idea behind crossing networks is that large limit orders—the kind of orders that may be placed by large institutional traders—are not adequately protected in a public exchange. Simply displaying large limit orders in the open book of an electronic exchange may

leak too much information about institutional traders' intentions. This information is used by prospective counterparties to trade more passively in the expectation that time constraints will force traders to replace some or all of the large limit orders with market orders. In other words, information leakage encourages *gaming* of large limit orders. Crossing networks are designed to limit information leakage by making their limit books opaque to both their clients and the general public.

A popular form of cross is the *midquote cross*, in which two counterparties obtain a midquote fill price. The midquote is obtained from a reference exchange, such as the NYSE or other public exchange. Regulations require that the trade is then printed to a public exchange to alert other market participants that it has taken place. The cross has no market impact but both the counterparties pay a fee to the crossing network. These fees are typically higher than the fees for other types of algorithmic trading because the market impact savings are significant while the fee is contingent on a successful cross.

More recently, crossing networks have offered their clients the ability to place limit orders in the crossing networks' dark books. Placing a limit order in a crossing network allows a cross to occur only at a certain price. This makes crossing networks much more like traditional exchanges, with the important difference that their books are opaque to market participants.

To protect their clients from price manipulation, crossing networks implement antigaming logic. As previously explained, opaqueness is itself a form of antigaming, but there are other strategies. For example, some crossing networks require orders to be above a minimum size or to remain in the network longer than a prespecified minimum time. Other networks will cross only orders of similar size. This prevents traders from ping-pong—sending small orders to the network to determine which side of the network's book has an order imbalance.

Another approach to antigaming prevents crosses from taking place during periods of unusual market activity. The assumption is that some of this unusual activity is caused by traders trying to manipulate the spread in the open markets to get a better fill in a crossing network.

Some networks also attempt to limit participation by active traders, monitoring their clients' activities

to see if their behavior is more consistent with normal trading than with gaming.

There are several different kinds of crossing networks. A *continuous crossing network* constantly sweeps through its book in an attempt to match buy orders with sell orders. A *discrete crossing network* specifies points in time when a cross will take place, say every half hour. This allows market participants to queue up in the crossing network just prior to a cross instead of committing resting orders to the network for extended periods of time. Some crossing networks allow scraping—a one time sweep to see if a single order can find a counterparty in the crossing network's book—while others allow only resting orders.

In *automated* crossing networks, resting orders are matched according to a set of rules, without direct interaction between the counterparties. In *negotiated* crossing networks, the counterparties first exchange indications of interest, then negotiate price and size *via* tools provided by the system.

Some traditional exchanges now allow the use of *invisible orders*, resting orders that sit in their order books but are not visible to market participants. These orders are also referred to as *dark liquidity*. The difference between these orders and those placed in a crossing network is that traditional exchanges offer no special antigaming protection.

Private dark pools are collections of orders that are not directly available to the public. For example, a bank or pension manager might have enough order flow to maintain an internal order book that, under special circumstances, is exposed to external scraping by a crossing network or *crossing aggregator*.

A *crossing aggregator* charges a fee for managing a single large order across multiple crossing networks. Order placement and antigaming rules differ across networks, making this task fairly complex. A crossing aggregator may also use information about historical and real-time fills to direct orders. For example, failure to fill a small resting buy order in a crossing network may betray information of a much larger imbalance in the network's book. This makes the network a more attractive destination for future sell orders. In general, the management of information across crossing networks should give crossing aggregators higher fill rates than exposure to any individual network.

Crossing lends itself to several optimization strategies. Longer exposure to a crossing network not only

increases the chances of an impact-free fill, but also increases the risk of a large and compressed execution if an order fails to obtain a fill. Finding an optimal exposure time is one type of crossing optimization. A more sophisticated version of this approach is solving for a *trade-out*, a schedule for trading shares out of the crossing network into the open markets. As time passes and a cross is not obtained, the strategy mitigates the risk of a large, compressed execution by slowly trading parts of the order into the open markets.

Other Algorithms

Two other algorithms are typically included in the mix of standard algorithmic trading offerings. The first is *liquidity seeking* where the objective is to soak up available liquidity. As the order book is depleted, trading slows down. As the order book is replenished, trading speeds up.

The second algorithm is *financed trading*. The idea behind this strategy is to use a sale to finance the purchase of a buy with the objective of obtaining some form of hedge. This problem has all of the components of a full optimization. For example, if, after a sell, a buy is executed too quickly, it will obtain a less favorable fill price. On the other hand, executing a buy *leg* too slowly increases the tracking error between the two components of the hedge and increases the dispersion of costs required to complete the hedge.

What is Next?

The average trade size for IBM, as reported in the Trade and Quote (TAQ) database, declined from 650 shares in 2004 to 240 shares in 2007. Falling trade sizes are evidence of the impact of algorithmic trading. Large, infrequent portfolio rebalancing and trading are being replaced by *small delta continuous trading*.

The antithesis of the small delta continuous trading approach is embodied in the idea of *lazy portfolios*, in which portfolios are rebalanced infrequently to reduce market impact costs. The first argument against lazy portfolios is that as time passes, the weights drift further away from optimal target holdings, in both the alpha and risk dimensions. Second, the use of an optimizer after long holding periods tends to produce large deviations from

current holdings. When executed—often relatively quickly—these deviations result in significant market impact costs.

Engle and Ferstenberg [6] show that to correctly measure risk, we must take both the existing positions and unexecuted shares into account. This idea unites execution risk with portfolio risk. Portfolio construction and optimal execution are similarly united by incorporating market impact costs directly into the portfolio construction process.

Ideally, the portfolio manager would like to solve a problem similar in nature to the multi-period consumption-investment problem [13], that, in addition, takes market impact costs and changing probability distributions for a large universe of securities into account. This *dynamic portfolio* or *small delta continuous trading* problem represents the next step in the evolution of institutional money management. However, it presents some mathematical and computational challenges. As has been pointed out by Sneddon [14], dynamic portfolio problem differs in several important ways from the classical multiperiod consumption-investment problem. First, the return probability distributions change throughout time. Second, the objective functions for active portfolio management do not depend on the predicted alpha/risk, but rather on realized return/risk. Finally, the dynamics of the model may be far more complex. Grinold [7] provides an elegant and analytically tractable but greatly simplified model. Kolm and Maclin [10] describe a full-scale simulation-based framework that incorporates realistic constraints and a transaction cost model.

Other efforts of ongoing research in algorithmic trading are extending market microstructure and optimal execution models to futures, options, and fixed income products. These initiatives follow the dominant theme of algorithmic trading, the creation of a unified view, an all-encompassing framework for the entire trading process, including modeling, portfolio construction, risk analytics, and execution across all tradeable asset classes.

References

- [1] Almgren, R. & Chriss, N. (2000). Optimal execution of portfolio transactions, *Journal of Risk* 3(2), 5–39.
- [2] Almgren, R. & Lorenz, J. (2007). Adaptive arrival price, in *Algorithmic Trading III: Precision, Control, Execution*, B.R. Bruce, ed, Institutional Investor Journals.

-
- [3] Almgren, R., Thum, C., Hauptmann, E. & Li, H. (2005). Equity market impact, *Risk* **18**(7), 57–62.
- [4] Bertsimas, D. & Lo, A.W. (1998). Optimal control of liquidation costs, *Journal of Financial Markets* **1**, 1–50.
- [5] Ellis, K., Michaely, R. & O'Hara, M. (2000). The accuracy of trade classification rules: Evidence from Nasdaq, *Journal of Financial and Quantitative Analysis* **35**(4), 529–551.
- [6] Engle, R.F. & Ferstenberg, R. (2007). Execution risk, *Journal Portfolio Management* **33**, 34–44.
- [7] Grinold, R.C. (2007). Dynamic portfolio analysis, *Journal of Portfolio Management* **34**(1), 16–26.
- [8] Hasbrouck, J. (1991). Measuring the information content of stock trades, *Journal of Finance* **46**(1), 179–207.
- [9] Huberman, G. & Stanzl, W. (2004). Price manipulation and quasi-arbitrage, *Econometrica* **72**(4), 1247–1275.
- [10] Kolm, P. & Maclin, L. (2009). *Small Delta Continuous Trading for Dynamic Portfolios*, Courant Institute, New York University.
- [11] Lee, C. & Ready, M. (1991). Inferring trade direction from intraday data, *Journal of Finance* **46**, 733–746.
- [12] Lillo, F., Farmer, J.D. & Mantegna, R.N. (2003). Master curve for price-impact function, *Nature* **421**, 129–130.
- [13] Merton, R. (1969). Lifetime portfolio selection under uncertainty: The continuous-time case, *Review of Economics and Statistics* **51**, 241–257.
- [14] Sneddon, L. (2005). *The Dynamics of Active Portfolios*, Westpeak Global Advisors.

PETTER N. KOLM & LEE MACLIN

Autocall

An equity-linked autocallable note, or autocall, is a legal instrument in which the notional engaged by the customer is automatically reimbursed when the underlying stock reaches a strike at a specified recalling date or at maturity if such an event does not occur (when the recall appends a coupon is also paid to give the client a return on his/her capital higher than the one he/she would get from an investment in a bond). Equity-linked autocallable notes are examples of hybrid products in the sense that the equity and interest rate movements need to be jointly modeled.

Product Description

A note is a legal wrapper issued by a financial institution. Usually, the capital engaged by the client is guaranteed at maturity and the note also delivers the payoff of a derivative instrument. If T_0 is the issue date, T the maturity, and $B(T_0, T)$, the discount factor between T_0 and T , at T_0 , one unit of the capital guaranteed at maturity is worth $B(T_0, T)$ (if we do not consider the credit risk associated to the issuer). Hence, for an issue at par, the price of the derivative product should be $1 - B(T_0, T)$ (called by practitioner *the available*). Thus, a fixed maturity note can be constructed with any type of payoff as long as the price of the derivative product matches the available.

When buying a note, the goal of the investor is to have a return on his/her capital that is higher than the riskless interest rate one, if the market ends at maturity in his/her expected configuration. However, the problem the client faces with fixed maturity notes is that his/her capital is blocked up until maturity. Autocallable notes solve that issue. The terms of a typical equity-linked autocallable note are listed below:

- At T_0 , the client invests 100.
- The product maturity is T and we consider T_i , $1 \leq i \leq N$ the autocallable dates with $T_0 < T_1 < \dots < T_i < \dots < T_N = T$.
- If S is the underlying (a stock or an index), K_i , $1 \leq i \leq N$ a set of strikes, C_i , $1 \leq i \leq N$ a set of coupon, the payoff is as follows:

- at T_i , $1 \leq i < N$, if the note has not been recalled at a previous autocallable date and if $S_{T_i} \geq K_i$, then the note is terminated (“automatically recalled” in the terminology) and the client receives $100 + C_i$ and,
- at T , if the note has not been recalled at a previous autocallable date and if $S_{T_N} \geq K_N$, the client receives $100 + C_N$; otherwise, he receives 100.

Usually, the autocallable dates have a fixed frequency, the strikes are equal, and the coupons have a linear progression: $C_i = iC$ with C that is higher than the riskless interest rate on the interval $\Delta T = T_{i+1} - T_i$, $0 \leq i < N$. Thus, in a bullish market, the investor can have a higher return on his/her capital than the interest rate one and his/her capital is available as soon as his/her target is reached. This explains why autocallable notes are very popular in the equity markets and have been massively traded in the past years.

Adapted Equity Rate Modeling

Compared to fixed maturity note, the modeling challenge with autocallable notes is the uncertain maturity of the product. With fixed maturity notes, the natural way of hedging the guaranteed capital is to trade a LIBOR-linked interest rate swap (see **LIBOR Rate**) from T_0 to T , where the seller of the note receives the fixed leg (the value of the LIBOR leg is $1 - B(T_0, T)$) and offset the interest rate risk on the capital). With an autocallable note, we can trade the same swap but we need to be able to cancel the remaining part of it at any callable date T_i if $S_{T_i} \geq K_i$. In a deterministic interest rate framework, the hedge would be to wait for the callable date to trade the corresponding remaining swap. Obviously, because the evolution of the interest rates is also a risk factor, this is not acceptable. A joint model on the equity and on the interest rate is required.

Regarding the specific interest rate volatility risk, it is a first-order risk in a sense that an autocallable product is affected by the global level of the rate curve much more than it is by the shape of the distribution of forward start swaps or by the autocorrelation of the rate curve. In other words, a one-factor model on the rate curve is sufficient and a natural choice is an extended Vasicek model (see **Term**

Structure Models) whose term structure of volatility has to be calibrated on the set of swaptions of maturity T_i written on the swap from T_i to T for all $i, 1 \leq i < N$, to be consistent with the autocallable note definition. This calibration set is a very classical one and is usually referred to as the *coterminal calibration set* (for the sum of the maturity of the swaptions and of the length of the corresponding swap is constantly equal to $T - T_0$). Technically, this calibration can be done prior to the equity calibration as it is just associated to the dates of the contract (and possibly to its set of strikes if one wishes to chose swaption strikes not equal to the “at money” strike to integrate a part of the swaption smile risk).

Regarding the equity, we propose a simple representation using only a term structure of volatility to show the impact of the rate factor on the equity volatility. Specifically, we use a two-factor Brownian model, one factor for the interest rate and the other for the Equity. Moreover, we suppose that the Equity is not paying dividends (*see Dividend Modeling*) and that the repo rate (*see Swaps*) is null. Under the risk neutral measure Q , this can be written as

$$\frac{dB(t, T)}{B(t, T)} = r_t dt + \Gamma(t, T) dW_t^{Q,r} \quad (1)$$

$$\frac{dS_t}{S_t} = r_t dt + \sigma_t^S dW_t^{Q,S} \quad (2)$$

$$d\langle W^r, W^S \rangle_t = \rho dt \quad (3)$$

The last equation does not depend on Q as change of measures affects neither the correlation between the Brownian motions nor their volatilities. Moreover, for the first equation, we suppose that the volatility $\Gamma(t, T)$ has been calibrated according to the coterminal calibration set of the product. Since the equity is not paying fixed, discrete dividends, its forward $F(t, T)$ is a martingale under the forward neutral measure Q_T , which is defined as

$$\frac{dQ_T}{dQ}(T) = \frac{\exp\left(-\int_{T_0}^T r_s ds\right)}{B(T_0, T)} \quad (4)$$

The expression of the forward, in that case, is $F(t, T) = S_t/B(t, T)$; hence, we have

$$\frac{dF(t, T)}{F(t, T)} = \sigma^S(t, T) dW_t^{Q_T,S} \quad (5)$$

and also, by combining equations (1) and (2)

$$\frac{dF(t, T)}{F(t, T)} = \sigma_t^S dW_t^{Q_T,S} - \Gamma(t, T) dW_t^{Q_T,r} \quad (6)$$

Hence,

$$\begin{aligned} \int_0^T |\sigma^S(t, T)|^2 dt &= \int_0^T |\sigma_t^S|^2 dt + \int_0^T |\Gamma(t, T)|^2 dt \\ &\quad - 2\rho \int_0^T |\sigma_t^S| |\Gamma(t, T)| dt \end{aligned} \quad (7)$$

Moreover, we have

$$\begin{aligned} E_t^Q[B(t, T)(S_T - K)^+] \\ &= E_t^Q[B(t, T)(F(T, T) - K)^+] \\ &= B(t, T) E_t^{Q_T}[(F(T, T) - K)^+] \end{aligned} \quad (8)$$

which leads to the generalization of the Black and Scholes formula in the sense that, in a setting with stochastic interest rates, the implied volatility Σ_T of the Black and Scholes framework is given by

$$\Sigma_T = \sqrt{\frac{1}{T} \int_0^T |\sigma^S(t, T)|^2 dt} \quad (9)$$

Thus, since the interest rate model has been calibrated independently to the equity model and supposing that the term structure of implied volatility is observed, it is possible to calibrate the term structure of the volatility of the stock, σ_t^S .

From the relation on the volatilities of this simple framework, it is important to note that, depending on the sign of the correlation of the rate factor with the equity factor and of the value of the volatility of the bond, the stock volatility can be either higher or lower than the implied volatility. This does not affect the probability of recalling the note at the first recalling date (which is not conditional) but the probability of recalling the note at a further recalling date conditionally of not having recalled the note at a prior recalling date. Furthermore, the addition of a rate factor, by extending the dimension of the model, leads to a Vega hedging of larger dimension since the swaptions used in the interest rate calibration set should now be considered as hedging instruments, and to a more complex Gamma hedging (*see Gamma Hedging*). In particular, for the latter, the “cross Gamma”, the second derivative with respect to the spot of the equity and to the spot of

the rate, cannot be hedged by any liquid instrument in the market. Therefore, it is necessary to adopt an approach by robustness, applying the same stochastic control arguments as for uncertain volatility models (see **Uncertain Volatility Model**), in order to control the evolution of the value of the hedged portfolio including the autocallable note. Practically, this can be done by replacing the instantaneous correlation between the two factors by an estimation of its upper bound as the cross Gamma, in the case of the autocallable note, is positive.

Advanced Modeling

While the equity/rate issue is truly the specificity of autocallable notes, two other specificities of the equity market need to be taken into consideration for such products, namely, the smile modeling and the dividends.

Regarding the smile issue, the difference between the price of an up binary option in the Black and Scholes framework with a fixed volatility and the price of the same up binary option in the presence of a smile is equal to the Vega of the call of same strike time the opposite of the local skew, $\partial\Sigma/\partial K$, of the implied volatility (this result is a direct application of the fact that an up binary option is bounded by two call spreads). Hence, in the equity market, the presence of a skew increases the price of the up binary option. Applying this result to the autocallable note means that the probability of recalling the note at the closest recalling date is higher if the volatility skew is considered than it is in a setting like the one we proposed earlier and, as a consequence, the term structure of the rate sensitivity is affected. However, while the risk associated to the skew of volatility is therefore important to model, the calibration of a two-factor model including this risk is more complex than what we proposed in the previous part and requires the use of numerical tools such as PDE resolutions (see **Forward-Backward Stochastic Differential Equations (SDEs); Partial Differential Equations; Finite Element Methods**). In particular, it is not possible to use the local volatility formula (see **Local Volatility Model**) as it is specific to a monofactor model.

The dividend risk is not so much of an issue in the case of the autocallable note defined in the section Product Description. However, it could become more

important in structures where the customer chose to risk a loss on his capital by selling a Down and In Put in exchange of higher coupons in case of an early termination. Nevertheless, under the forward neutral measure Q_T , all assets of the form $x_t/B(t, T)$, with x_t a spot process, are martingales (we have used this result in the previous part where S_t is a spot process). Yet, with the addition of fixed discrete dividends, the forward of the equity is including discount factor terms from each dividend dates to the considered maturity. Hence, it is no longer a martingale and, in the presence of stochastic interest rates, the calibration of its volatility is more complex than what we have proposed earlier and also requires the use of advanced numerical tools because of the **Convexity Adjustments** due to the fixed dividends.

Extensions

The payoff we have presented is the most basic one. To increase the value of the coupons, a common practice is to consider a basket of stocks as underlying and to condition the recall on the spot of the worst performer at each recalling date. While this does not change the analysis we have made, this extends the dimension of the problem and can require the integration of other sources of risk such as the equity correlation skew (see **Basket Options**). Moreover, if the stocks are not in the same economy, the number of rate factors also increases and, as explained in **Quanto Options**, quanto drift terms lead to more complex equations.

Another classical extension of the autocallable note is to condition the coupon on a rate instrument. Typically, upon recall, the coupon paid would be equal to a swap rate or a LIBOR rate plus a fixed spread. In that case, it is usually necessary to adapt the rate factor to have correct convexity adjustments (see **CMS Spread Products** for the explanation of the convexity adjustment for constant maturity swaps and **Bermudan Swaptions and Callable Libor Exotics** for correlated issues). Moreover, in that case, the impact of the joint equity/rate modeling on the price of the autocallable note is far more important than it is in the fixed coupons case.

There are a few more extensions of the autocallable note that have been proposed in the recent

years, but the payoffs we have mentioned represent the vast majority of traded contracts. Yet, in all cases, the modeling challenge lies in the integration of stochastic interest rates in the equity

framework, which defines autocallable notes as hybrid products.

CHARLES-HENRI ROUBINET

Municipal Bonds

Municipal bonds are bonds issued by local and state governments to raise funds for public works with interest payment(s) that are generally exempt from federal taxes, and often from state and local taxes as well. An investor who buys a bond will receive interest payments at a stated yield rate until the bond reaches maturity, at which time the principle, or *par*, will be returned. Not all municipal bonds are tax exempt. Those intended for purposes not essential to the public good, such as those issued to finance sports arenas, for example, are taxable. Typically though, the terms *municipal* and *tax exempt* are synonymous.

There are two main types of municipal bonds: *general obligation* and *revenue*. Both are used to raise funds for public works, but they differ in how they generate the proceeds to service the interest and/or principal payment(s). General obligation bonds use tax revenue to pay back bondholders, while revenue bonds use money generated from the works themselves; a toll road, for instance, would pay the bond interest using monies raised by tolls. Neither is necessarily a more secure investment. General obligation municipal bonds, which tie interest payments to taxes, are rated based on the creditworthiness of their issuers. Since these are local or state governments, they are assessed by comparing the strength of their tax base to their debt service requirements. Revenue bonds, in contrast, are rated based on their *debt service coverage ratio*, which divides the amount of interest to be paid by the amount of money available, as estimated by analysis of past and future data. Moody's, Standard & Poor's, and Fitch Ratings are the large commercial ratings agencies that establish the basis for most assessments of municipal bond ratings.

Revenue bonds generally yield more than general obligation bonds do, though neither is considered risk-free; taxes cannot be raised without limit to pay back general obligation investors during a financial crisis, nor can revenue bonds be guaranteed to always cover their debt service. If a bond's credit quality falls, its yield will increase and its price will drop, which will make it more attractive to investors. Both types can be resold in the secondary market, though not without considerable cost.

Bond insurance, purchased by the issuer to cover both the principal and interest in case of default,

can also be a factor in the rating of particular bond. A bond backed by a bond insurance firm carries the firm's investment rating, though that rating will change if the insuring firm's rating is downgraded. The yields on insured bond are generally somewhat lower than they otherwise would be, but investors have an extra safeguard against the risk of default, which has the ancillary benefit of making these bonds easier to resell in the secondary market. It is important for investors to know what the rating of a particular bond, or bonds of that type, would be if it were uninsured.

Another factor which investors consider when evaluating a municipal bond is its coupon rate, meaning the rate at which it pays interest until it matures. A bond's coupon rate can be fixed or it may be reset periodically; this latter type is referred to as a *floating-rate* or *variable-rate* issue.

Several specialized subsets of municipal bonds exist, with particular revisions to key features. For instance, municipal bonds may be issued as "refunded" bonds, which are backed by US Treasury bonds held in escrow. Municipalities that issued bonds at higher interest rates in the past can issue new bonds, "refunded bonds," at lower rates to redeem the principal of these older bonds; the proceeds from the sale of the new bonds are used to buy US Treasury securities, which back the coupon payments for the original bonds, freeing them of any default risk. These refunded bonds typically have shorter maturities and corresponding low interest-rate risks. Because their yields are backed by funds held in escrow and are therefore risk-free, these are highly valued bonds, almost always carrying an AAA rating.

Municipal notes are another category of municipal bonds. Municipal notes have maturities less than three years and are issued in anticipation of some future municipal revenue, either from taxes or bonds.

Municipal zeros, another special type of municipal bond, are issued at a discount from their *par* value, which refers to the amount of their principal. When they reach maturity, they are redeemed for their full par value, with the difference between the two rates equaling a specified compounded annual yield. Because municipal zeros pay no interest, potential investors must carefully consider the creditworthiness of their issuers. Issuers also retain a fuller freedom to *call* these bonds, meaning that they can collect these bonds before they mature, at a price discounted from par.

Some municipal bonds are issued with *put* provisions, meaning that purchasers can tender the bond back to the issuer at par, which offers security for investors, though not without trade-offs. The total return on these bonds is usually lower, and in the case that interest rates decline, they will appreciate less than comparable long-term bonds.

Many other specialized types of municipal bonds attract different investors. *Dedicated tax-backed bonds* or *structure/asset bonds* meet their debt service with certain revenues, such as sales taxes, dedicated to repayment of the debt. Other types include *letter of credit back bonds*, *moral obligation bonds*, and *tax-allocation bonds*.

Several aspects influence municipal bond pricing. Municipal bonds can be priced at a discount, at par, or at premium. The bond's features, including its maturity, call features, and credit quality, influence its price. A *call risk* refers to the possibility that the issuer will redeem a bond before its maturity date, meaning that purchasers will not earn the subsequent interest. Issuers typically may redeem bonds at stipulated dates, starting at 10 years from the date of issue, and at 5-year intervals thereafter, at prices increasing from par, though housing bonds can have more irregular call features. A substantial drop in interest rates may cause an issuer to call a municipal bond.

Municipal bond pricing can vary because of the disparities between broker markups and the general lack of transparency in the pricing process. In general, smaller lots, those with values less than \$25 000, are more difficult to sell and so are marked up disproportionately by brokers, and their prices can be unpredictable. Larger lots generally sell at narrower price ranges. Other price factors include whether the bond is well known and widely traded, and whether it is insured. Much of the price information available though still applies generally to very large lots, purchased wholesale by institutions, not by individual investors.

Municipal bonds can be bought from brokerage firms, both full service and discount, including those who specialize in municipal bonds, in addition to banks. All of these offer online-purchasing options, though generally with both markups and "transaction fees" added on. Buying bonds as they first become available is known as *buying at issue*, which has several attractive features. During this short time, bonds are sold at a uniform price, often at par, and

dealers anxious to sell will offer good prices, which results in attractive yields.

While buying municipal bonds, it must be borne in mind that a key distinction exists between premium and discount bonds. Premium bonds yield more than discount bonds, assuming an equality of credit quality and maturity, because their yield to maturity, the amount they will earn with interest over their lifespan, will be equal to their actual yields. In contrast, discount bonds have yields, which include a taxable capital gain because they are sold at prices below their maturity value. Only coupon interest is exempt from federal taxes, which lowers discount bonds' actual yield below their quoted yield to maturity. Premium bonds have a higher coupon value, meaning that they produce more interest income, making them better suited to investors who want their short-term municipal bond holdings to yield an income to reinvest. Others looking to hold onto their bonds to maturity and use them to fund long-term investments will be better served by discount bonds.

While municipal bonds are generally free from federal taxes, which is their most attractive feature, state or local taxes may still apply. Out-of-state buyers may be exempt from these taxes. Other tax considerations include the alternative minimum tax, which applies to bonds for "nonessential" projects issued after August 7, 1986, but typically only for buyers with large incomes. This is also true of *de minimus* taxes, which applies to discount bonds. Taxability of a bond at the local, state, or federal level is not the only criterion to determine its potential value to an investor; depending on which tax bracket the investor is subject to, a taxable bond may still offer a higher yield.

Municipal Bond Arbitrage

In recent years, many bond traders, investment banks, and hedge funds have created leveraged bond portfolios that utilize what we may call municipal bond arbitrage strategies. The origin of municipal arbitrage is the generally cheaper pricing of long maturity municipal bonds relative to similar fixed income investments. There are two approaches: those that use tender option bond programs to augment the yield spread earned by hedging municipal carry trades, and those that use basis trading strategies to exploit volatility in the relative value of municipals. Both

types of arbitrage impact the municipal market as a whole by increasing liquidity and improving price discovery.

Municipal arbitrage is similar to other types of fixed income arbitrage; it targets pricing inefficiency, hedges against the fluctuation of interest rates, and uses leverage. However, unlike its now infamous cousin, mortgage-backed security arbitrage, opportunities in municipal arbitrage, though not without their own risks, arise from the inherent structure of the municipal market, rather than from the sheer complexity of modeling risk in that market.

The Structure of Municipal Bond Arbitrage Strategy

Municipal bond arbitrage trades are typically hedged against interest rates with LIBOR swaps [2]. The trader makes a semiannual fixed payment at a market rate set at the time of the swap, and receives a variable payment set at the quarterly LIBOR reference rate. As interest rates fluctuate, so does the swap's value.

The price value of a basis point is calculated by comparing a one basis point change in interest rates across the yield curve to the corresponding change in the value of an interest-hedged position. This is also known as the *swap's dollar duration*. The hedge ratio is the ratio of the dollar durations of the swap position and the original bond position.

As his/her starting point is a municipal carry trade, the trader also seeks to generate income by locking in the municipal yield premium; when financing and hedging costs are subtracted from this is, it is known as the *net spread*. The municipal *versus* LIBOR yield ratio is thus a supplementary source of profit or loss.

Hedging

The basis of a hedged position is the difference in value between the hedge and the security that is hedged by it. The greater their dissimilarity, the greater the basis risk—in other words, the risk that the hedge position will diverge in value from the hedged position. The arbitrageur attempts to isolate the basis risk of a dubiously priced security by hedging against other risks, as in the interest-rate swaps previously described. Yield curve risk can be hedged by taking a bond position of the same tenor.

Tax risk, issuer default risk, and liquidity risk should also be considered, though they are simply considered part of the basis in calculating the municipal/LIBOR yield ratio.

The arbitrageur must also determine a suitable hedging ratio once the hedge itself is chosen. This is typically done by balancing the dollar durations of one basis point of the bond and one basis point of the hedge. At the same time, the arbitrageur proceeds under the assumptions that the yield on the bond and the yield on the hedge will not remain symmetrical and that this yield ratio itself is influenced by interest-rate changes. What the arbitrageur must attempt to anticipate is the hedge beta, or the change in municipal yields that accompanies any change in taxable yields. Traders use several methods to estimate hedge beta. Ratio-based hedges use the yield ratio itself, supposing that it will revert to a mean or to the overall market. Regression-based hedges differ in that they account for the usual inverse correlation between yield ratios and interest rates. Alternately, some traders consider the interest-rate variance as the key component of the position and hedge accordingly. It is also possible to hedge against the callability of longer municipal bonds (which is typically priced into quoted yields).

Tender Option Bonds, Variable-rate Demand Obligations, and Leverage

Most municipal bond arbitrage strategies incorporate leverage, which is enabled by tender option bond programs (TOBs) that securitize bonds or bond portfolios and thus allow them to be financed at what amounts to tax-exempt rates [1]. Leverage is achieved in the taxable bond market by means of repurchase transactions, in which the owner of a security sells it to a second party while agreeing to repurchase it later at a higher price. If a lender provides similar means to purchase a municipal bond, the interest he receives will still be fully taxable. However, borrowers cannot deduct this interest cost, since they are purchasing a tax-exempt security. As a result, borrowers will generally receive a lower yield from the lender even though they are financing at a higher rate.

In order to finance the municipal bond purchase, the tender option bond trusts can separate the bond into a tax-efficient component and a second component that is exposed to tax risk (along with liquidity

risk and the other risks that account for its yield curve premium) by creating the following:

1. A short-term, variable-rate instrument backed by a long-term municipal bond.
2. An instrument that passes the upside and downside of this collateral to an investor who seeks leveraged exposure to a municipal bond.

Tender option bond programs are sponsored by most major municipal bond dealers. All of these programs allow customers to structure leveraged municipal bond position by issuing short-term money-market securities collateralized by long-term municipal bonds (sometimes called *floaters*) in tandem with securities that represent a leveraged position in the collateral (sometimes called *residuals*).

The fate of the underlying bonds is largely determined by the residual holders, who generally hold the right to require the floater holders to tender their securities with a 1-day notice, allowing the former to terminate the trust and liquidate the underlying bonds. Floaters have priority access to payments from the trust, which allows them to take advantage of a rise in short-term rates (leaving, perhaps, little income to the residuals).

As with variable-rate demand obligations (VRDOs), the rate on the floaters is set by the remarketing agent at a level that allows the bond to be remarketed at par. When the underlying bonds are deposited, the trust becomes the legal owner. The trustee administers the trust and its payments on behalf of the beneficial owners, that is, the holders of the floaters and the residuals. TOBs are tax partnerships in order to preserve the underlying bonds as tax exempt, which means that there are specific guidelines that govern gain sharing and termination and tend to discourage active trading of the underlying bonds. Both types of tender option bond securities are only available as private placements.

Some companies are able to efficiently finance municipal bonds without TOBs; the tax code stipulates that interest costs are deductible if tax-exempt obligations are less than 2% of the taxpayer's assets. (Municipal bond broker-dealers can deduct interest costs in this manner by classifying a portion of their inventory as equity-funded rather than as debt-funded).

Furthermore, some tender option bond programs offer pooled trusts to increase leverage, new tools for hedging such as credit default swaps are available, and nontraditional participants (e.g., foreign taxpayers) are entering the market through municipal ratio swaps. Even market participants who do not use arbitrage strategies should be aware of these developments because of their potential to impact the shape of the market as a whole.

References

- [1] Fabozzi, F.J. & Feldstein, S.G. (2008). *The Handbook of Municipal Bonds*, Wiley, New York.
- [2] Namvar, E.. (2006). *An Introduction to Municipal Bond Arbitrage*, University of California, Irvine, Working Paper.

Further Reading

- Fabozzi, F.J. (2001). *The Handbook of Fixed Income Securities*, 6th Edition, McGraw-Hill, New York.
- Fabozzi, F.J. (2007). *Bond Markets, Analysis, and Strategies*, 6th Edition, Pearson, New Jersey.
- Thau, A. (2001). *The Bond Book*, 2nd Edition, McGraw-Hill, New York.

Related Articles

Bond; Bond Options; LIBOR Rate; Swaps.

ETHAN NAMVAR

Asset–Liability Management

Asset–liability management is a major concern for large financial institutions. Assets must be invested to achieve sufficient returns to pay liabilities and achieve goals subject to uncertainties, policy, taxes, and legal constraints. This article discusses asset and liability management with portfolio complexities, transaction costs, liquidity, taxes, investor preferences (including downside risk control, policy constraints, and other constraints), uncertain returns, and the timing of returns and commitments. Applications to the management of investment portfolios for large financial institutions, including pension funds and insurance companies, and for individual life-cycle planning are discussed. The approach recommended is discrete-time, multiperiod stochastic programming. For most practical purposes, such models provide a superior alternative to other approaches, such as mean–variance, simulation, control theory, and continuous-time finance. These models utilize investor preferences in a simple, understandable way. We present the Russell–Yasuda Kasai and the InnoALM models, both of which had a major impact on the practice of asset–liability management (ALM).

The use of scenario-based, stochastic programming optimization models in discrete time provides an approach to asset–liability modeling over time. The models provide a tool to think about, organize, and do calculations concerning how one should choose asset mixes over time to achieve goals and cover liabilities. Risk and return are balanced to achieve period-by-period goals and pay liabilities. The models force diversification and the consideration of extreme scenarios to protect investors from the effects of tail outcomes and also do well in normal times. They will not let individuals or institutions get into situations in which extreme, but plausible, scenarios would lead to truly disastrous consequences, such as losing half or more of one’s assets. Because these models force consideration of all relevant scenarios, the common practice of assuming that low-probability scenarios will not occur is avoided. Hence, the disasters that frequently follow from this error are avoided.

In discrete-time, multiperiod stochastic programming models, one typically maximizes a concave,

risk-averse utility function composed of the discounted expected wealth in the final period less a risk measure composed of a risk-aversion index times the sum of convex penalties for target violations relating to investor goals of various types in various periods. The convexity means that the larger the target violation, the larger the penalty cost. Hence, risk is measured as the nonattainment of investor goals, and this risk is traded off against expected returns. This approach is similar to a mean–variance preference structure, except that it is based on final wealth and the risks are downside risks that are measured across periods and investor goals. Discrete scenarios that represent the possible returns and other random parameter outcomes in various periods are generated from econometric and other models, such as those related to market dangers with increasing risk and from expert modeling. Mean-return estimation and the inclusion of extreme events are important for model success. The scenario approach has a number of advantages:

- Normality or lognormality, which is not an accurate representation of actual asset prices (*see Stylized Properties of Asset Returns; Heavy Tails*), especially for losses, need not be assumed.
- Tail events can be easily included; studies show that downside probabilities estimated from actual option prices are 10–1000 times fatter than log-normal (*see Heavy Tails; Stylized Properties of Asset Returns*).
- Scenario-dependent correlations between assets can be modeled and used in the decision-making process so that “normal” and “crisis” economic times (with higher and differing signed correlations) can be considered separately.
- The exact scenario that will occur and the probabilities and values of all the scenarios need not be accurately determined to provide model performance that is superior to that of other models and strategies, such as mean–variance or portfolio insurance.
- Most of the natural, practical aspects of asset–liability applications can be modeled well in the multiperiod stochastic programming approach. Model output is easy to understand by nonexperts such as pension fund trustees. The models can be tested *via* simulation and statistical methods and considerable independent evidence

demonstrates their superiority to other standard approaches and strategies.

- The approach protects investors from large market losses by considering the effects of extreme scenarios while accounting for other key aspects of the problem.
- Determining whether investment positions are truly diversified and of the right size across time is crucial to protect against extreme scenarios and ensure that the results will be good in normal times and avoid disasters.

Procedures for Scenario Generation

There are many methods to estimate scenarios. Abaffy *et al.* [1, 10] survey scenario estimation and aggregation methods that represent a larger number of scenarios by a smaller number. See [31] for further references and discussion. Scenarios are a means to describe and approximate possible future economic environments. They are represented as discrete probabilities of specific events. Together, all the scenarios represent the possible evolution of the future world. The basic idea is to have a set of T period scenarios of the form $S^T = (S_1, S_2, \dots, S_T)$, where $s_t \in S_t$ are the possible outcomes of all random problem elements and where s_t occurs in period t with probability $p_t(s_t)$.

For asset–liability modeling, the most important parts of the distribution are the means and the left tail. The mean drives the returns and the left tail the losses. We cannot include all possible scenarios but rather focus on a discrete set that well approximates the possible important events that could happen. With S^T total scenarios, we can include those we want. Once a scenario is included, the problem must react to what would be the consequences of that scenario. This important and flexible feature of stochastic programming modeling is not usually available in other approaches. The inclusion of such extreme scenarios means that the model must react to the possibility of that scenario occurring. This inclusion is one of the ways a stochastic programming overall model would have helped mitigate the 1998 losses and collapse of LTCM (see **Long-Term Capital Management**). The model would not have let them hold such large positions. We frequently aggregate scenarios to pick the best N out of the S^T so that the modeling effort is manageable. The generation of good scenarios that well represent the future

evolution of the key parameters is crucial to the success of the modeling effort.

Among other things, scenarios should consider the following:

- mean reversion of asset prices;
- volatility clumping, in which a period of high volatility is followed by another period of high volatility;
- volatility increases when prices fall and decreases when they rise;
- trending of currency, interest rates, and bond prices;
- ways to estimate mean returns;
- ways to estimate fat tails; and
- ways to eliminate arbitrage opportunities or minimize their effects.

The true distribution P is approximated by a finite number of points $(w^1, \dots, w^S)$ with positive probability p^s for each scenario s . The sum of all scenario probabilities is 1.

Economic variables and actuarial predictions drive the liability side, whereas economic variables and sentiment drive financial markets and security prices. Hence, estimating scenarios for liabilities may be easier than for assets because there often are mortality tables, actuarial risks, legal requirements, such as pension or social security rules, well-established policies, and so on. Such scenarios may come from simulation models embedded into the optimization models that attempt to model the complex interaction between the economy, financial markets, and liability values. We can use the following classification:

1. There can be complete knowledge of the exact probability distribution. This knowledge usually comes from a theoretical model, but it is possible to use historical data or an expert's experience.
2. There can be a known parametric family based on a theoretical model whose parameters are estimated from available and possibly forecasted data.
3. Scenarios can be formed by sample-moment information that aggregates large numbers of scenarios into a smaller, easier set or generates scenarios from assumed probability distributions. This idea was used in the five-period Russell–Yasuda Kasai insurance model in 1989 [6]. Høyland and Wallace (HW) [17] have expanded and refined the idea for multiperiod

problems and they have generated scenarios that match some moments of the true distribution. Typically it is the first four moments.

4. The idea is to assume that the past will repeat in the sense that events that occurred with various probabilities will reoccur with the same probabilities but not necessarily in the same order. Despite missing fat tails, the method has had some success. This idea can be implemented by using the raw data or through procedures, such as vector autoregressive modeling or bootstrapping, that sample from the past data. Using subsets of past data is a viable scenario generation method. Grauer and Hakansson [13] have successfully used the idea in many asset-only studies. When no reliable data exist, one can use an expert's forecasts. Examples include [24, 28] or governmental regulations.
5. The Hochreiter and Pflug (HP) [16] approach generates scenario trees using a multidimensional facility location problem that minimizes a Wasserstein distance measure from the true distribution. Kolmogorov–Smirnov (KS) distance approximation is an alternative approach, but it does not take care of tails and higher moments. The HP method combines a good approximation of the moments and the tails.

Whatever method is used to generate the scenarios, in relying on the meshing of decision-maker subjective estimates, expert judgment, and empirical estimation, it is crucial to validate the estimated distributions and to make sure that the decision maker has not defined the range too narrowly. Perhaps reflecting on the distribution by asking what would make the value be outside the range and then assessing the probability would help expand the range and make the probability assessment more realistic.

Assessing Scenarios

The philosophy proposed here for assessing scenarios is as follows. Markets are understandable most (over 95%) of the time, but real asset prices have fat tails because extreme events occur much more frequently than lognormal or normal distributions indicate. According to Keim and Ziemba [20], much of asset returns are *not* predictable. Hence, we must have ways to combine conventional models, options pricing, and so on, which are accurate most of the

time with the irrational unexplainable aspects that occur once in a while. Whether the extreme events are predictable is not the key issue—what is *crucial* is that we consider that they can happen in various levels with various chances.

Even apart from the instability due to speculation, there is the instability due to the characteristic of human nature that a large proportion of our positive activities depend on spontaneous optimism rather than mathematical expectations, whether moral or hedonistic or economic. Most, probably, of our decision to do something positive, the full consequences of which will be drawn out over many days to come, can only be taken as the result of animal spirits, a spontaneous urge to action rather than inaction, and not as the outcome of a weighted average of quantitative benefits multiplied by quantitative probabilities [21].

Human behavior is a main factor in how markets act. Indeed, sometimes markets act quickly, violently with little warning. Ultimately, history tells us that there will be a correction of some significant dimension. I have no doubt that human nature being what it is, that it is going to happen again and again [14].

As a working background hypothesis, we assume that markets are basically efficient most of the time, so we have to deal with that plus the anomaly and behavioral finance departures from efficiency.

Dynamic and Liability Aspects

Figure 1 shows the time flow of assets arriving and liability commitments leaving for institutions, such as insurance companies, pension funds, banks, and individuals.

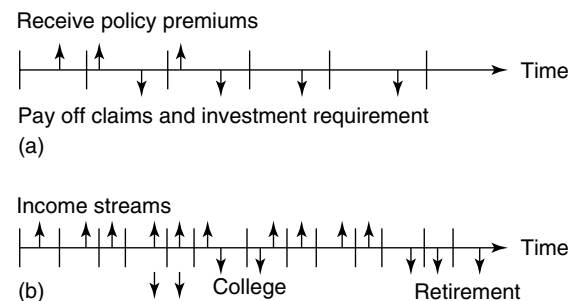


Figure 1 Time flow of assets: (a) institutions and (b) individuals

Ziemba and Mulvey [32] suggest the following risk ladder with various levels of details, aggregation, and model decisions.

- *Rung 5*: Total integrated risk management.
- *Rung 4*: Dynamic asset and liability management.
- *Rung 3*: Dynamic asset-only.
- *Rung 2*: Static asset-only portfolios.
- *Rung 1*: Pricing single securities.

This article concerns mostly Rung 3 and, especially, Rung 4. Rung 5 represents overall company-wide models that involve all aspects of a business; see [26].

The stochastic programming approach, which considers the following aspects, is ideally suited to analyze such problems. It is argued that it has the following elements.

- Multiple time periods; possible use of end effects—steady state after decision horizon adds one more decision period to the model; the trade-off is to have either an end-effects period or a larger model with one less period. The Russell–Yasuda Kansai model uses end effects while the InnoALM model does not.
- The scenarios should be consistent with economic and financial theory for asset returns, interest rates, and bond prices with anomaly and behavioral finance scenarios included.
- Discrete scenarios for random elements—returns, liabilities, and currencies.
- Scenario-dependent correlation matrices so that correlations change for extreme scenarios. This was first implemented in the InnoALM model. Since bond prices rise during stock market crashes, these assets are negatively correlated then but they are positively correlated otherwise. Thus, capturing this effect with one correlation matrix will not work. One simply needs multiple correlation matrices for scenarios sets. The LTCM and October 27–28, 1997 crash examples discussed in [31] remind one of the importance of this.
- The use of various forecasting models that can handle fat tails.
- Institutional, legal, and policy constraints.
- Model derivatives and illiquid assets.
- Model the transaction costs.
- Use expressions of risk in terms understandable to decision makers item.

- Simple, easy-to-understand concave, piecewise linear, risk-averse utility functions that maximize long-run expected profits, net of expected discounted penalty costs for shortfalls, are suggested. One pays more penalty for shortfalls as they increase (highly preferable to Value-at-Risk (VaR)).
- Model as constraints or penalty costs for target violations in the objective function. This allows for multiple goals through targets within periods and across periods. However, one must be able to specify the relative importance of these targets within periods and across periods.
- Adequate reserves, cash levels, and regularity requirements must be maintained.
- Realistic multiperiod problems can now be solved on modern workstations and personal computers by using large-scale linear programming and stochastic programming algorithms.
- The model makes one diversify—the key for keeping out of trouble.

Some possible approaches to model situations with such events are as follows:

- Simulation models have very large output to understand but are very useful as a check. Stochastic programming (SP) models are useful because they can easily perform stress tests and what if calculations.
- Mean–variance models are static in nature and useful for such applications, but they are not very useful with liquidity or other constraints or for multiperiod problems or with liabilities, and so on.
- Expected log models yield very risky short-term strategies that do not diversify well; fractional Kelly with downside constraints is excellent for risky investment betting [23, 30].
- Many stochastic control models, although theoretically interesting, give hair-trigger and bang-bang policies in which one is 100% stocks and then 0% stocks one second later (*see Stochastic Control; Uncertain Volatility Model*). See, for example, the Brennan and Schwartz model [3] in [32] and the Rudolf and Ziemba model [27]. The question is how to constrain the asset weight changes to be practical. Possibly, the best work with this approach has been done by Campbell and Viceira [4], who used the approach successfully to analyze long-term asset-only allocation decisions in

which the power of the technique dominates the limitations of the model (which they acknowledge). Among other conclusions, they show the following:

1. The riskless asset for a long-term investor is not Treasury-bills; rather, it is an inflation-indexed bond since it delivers a predictable stream of real income.
2. A safe labor income stream is equivalent to a large position in the risk-free asset, allowing the investor to hold much more in risky assets; then, this is reversed to some extent. Fixed commitments are negative income.
3. Risky investments are extremely attractive to young households because they have large relatively safe human wealth relative to their financial wealth.
4. Business owners should have less equity exposure since their income stream is correlated with the stock market.
5. Wealthy investors should be more risk averse since more of their consumption stream depends on their financial success.
6. Since stocks are mean reverting, that is, they have lower risk over longer time horizons, investors can time the market over longer horizons; since the equity risk premium is time varying, the optimal strategic allocation mix changes over time. These are useful rules of thumb for many investors derived from a theoretically sound framework. The goal in SP-ALM models is to tailor the asset-allocation mix for the particular institution or investor given their consumption and other goals, taxes, preferences, uncertainties, transactions costs, liquidity, and so on.
7. Stochastic programming models with decision rules have policy prescriptions, such as fixed mix or buy and hold; the decision rules are intuitively appealing but are suboptimal and usually lead to nonconvex difficult optimization modeling.

Stochastic Programming is Superior to Fixed Mix

Despite their good results, fixed-mix and buy-and-hold strategies do not use new information from return outcomes in their asset allocation (see **Fixed Mix Strategy**). Indeed, their asset allocation is fixed

independent of all new information. Hence, a stochastic programming model that reacts to such information should be superior. For an example of this, see [31].

Fleten *et al.* [11] compared two versions, namely, multistage stochastic linear programming and fixed-mix, of a portfolio model for the Norwegian life insurance company Gjensidige NOR. They found that the multiperiod stochastic programming model dominated the fixed-mix approach. However, the degree of dominance was much smaller out of sample than in sample (Figure 2).

Why does the SP advantage decrease so much? The answer seems to be that out of sample, the random input data are structurally different from those in sample, so the stochastic programming model

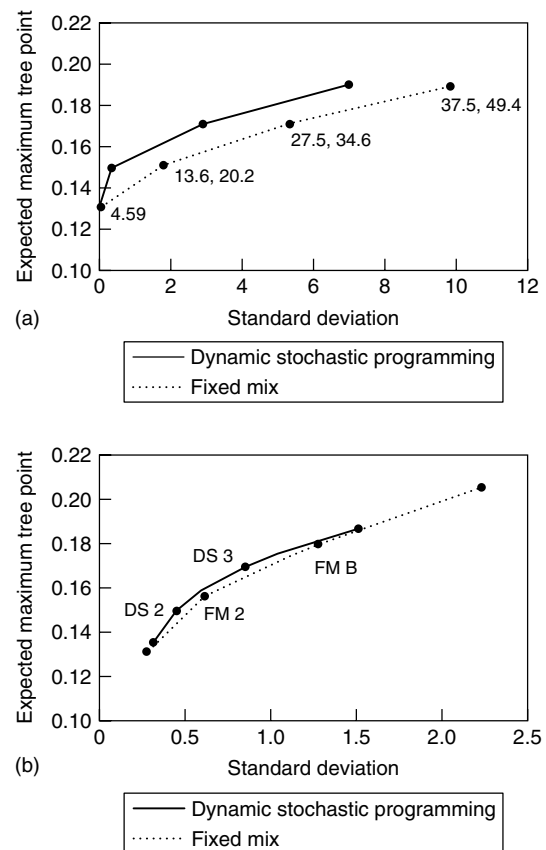


Figure 2 Advantage of stochastic programming over fixed-mix model: (a) in-sample results and (b) out-of-sample results [Reproduced from Fleten *et al.*, 2002. © Elsevier.]

loses its advantage in optimally adapting to the information available in the scenario tree. In addition, the performance of the fixed-mix approach improves because the asset mix is updated at each stage through its volatility pumping procedure.

The Russell–Yasuda Kasai Model

The Russell–Yasuda Kasai model was the first large-scale multiperiod stochastic programming model implemented for a major financial institution (see [15]). The model was designed while the author was working as a consultant to the Frank Russell Company in Tacoma, Washington, from 1989 to 1991. Under the direction of Research Head Andy Turner, the team of David Cariño, Taka Eguchi, David Myers, Celine Stacy, and Mike Sylvanus at Russell in Tacoma, Washington, implemented the model for the Yasuda Fire and Marine Insurance Company in Tokyo.

The Russell–Yasuda Kasai model was designed to satisfy the following need, as articulated by Kunihiro Sasamoto, director and deputy president of Yasuda Kasai:

The liability structure of the property and casualty insurance business has become very complex, and the insurance industry has various restrictions in terms

of asset management. We concluded that existing models, such as Markowitz mean–variance, would not function well and that we needed to develop a new asset–liability management model.

The Russell–Yasuda Kasai model is now at the core of all asset–liability work for the firm. We can define our risks in concrete terms, rather than through an abstract, in business terms, measure like standard deviation. The model has provided an important side benefit by pushing the technology and efficiency of other models in Yasuda forward to complement it. The model has assisted Yasuda in determining when and how human judgment is best used in the asset–liability process ([5], p. 49)

The full model is described in [6, 7]. Figure 3 shows the decision-making process. The Russell–Yasuda Kasai model led to other Russell models (Table 1).

InnoALM: The Innovest Austrian Pension Fund Financial Planning Model

Siemens Oesterreich, a part of the global Siemens Corporation, is the largest privately owned industrial company in Austria. Its businesses, with revenues of €2.4 billion in 1999, include information and communication networks, information and communication products, business services, energy and traveling technology, and medical equipment. Its pension fund,

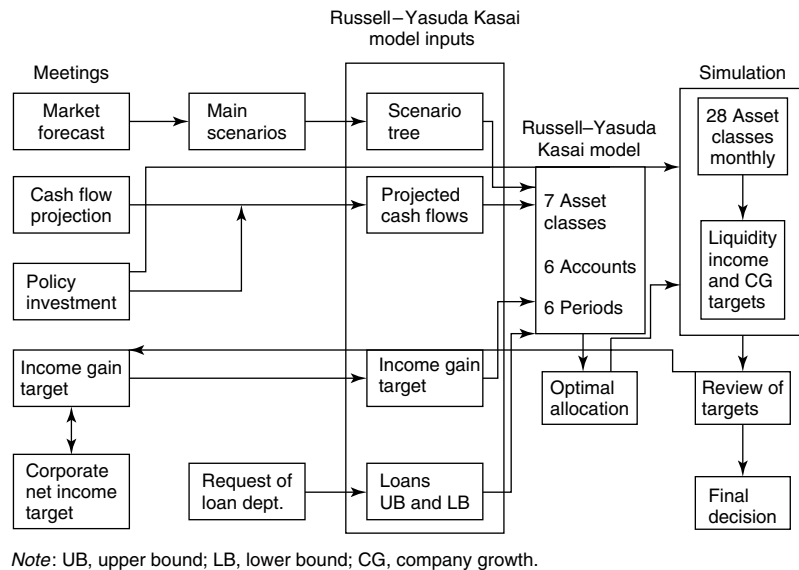


Figure 3 Yasuda Kasai’s asset–liability decision-making process [Reproduced from Cariño, 1994. © INFORMS.]

established in 1998, is the largest corporate pension plan in Austria and is a DCP. More than 15 000 employees and 5000 pensioners are members of the pension plan, which had €510 million in assets under management as of December 1999.

Innovest Finanzdienstleistungs, the investment manager of the Siemens pension plan, was rated as the best plan in Austria. Various uncertain aspects, possible future economic scenarios, stocks, bonds, and other investments, transaction costs, liquidity, currency aspects, liability commitments over time, Austrian pension fund law, and company policy suggested that a good way to approach this asset–liability problem was *via* a multiperiod stochastic linear programming model. This model has innovative features, such as state-dependent correlation matrices, fat-tailed asset-return distributions, simple computational schemes, and output. InnoALM was developed in six months in 2000, with Geyer and Ziemba serving as consultants and with assistance from Herold and Kontriner from Innovest.

The liability side of the Siemens pension plan consists of employees for whom Siemens is contributing DCP payments and retired employees who receive pension payments. Contributions are based on a fixed fraction of salaries, which varies across employees. Active employees are assumed to be in steady state; thus, employees are replaced by a new employee with the same qualification and sex so that there are a constant number of similar employees. Newly employed staff start with less salary than retired staff, which implies that total contributions grow less rapidly than individual salaries.

The set of retired employees is modeled using Austrian mortality and marital tables. Widows receive 60% of the pension payments. Retired employees receive pension payments after reaching age 65

for men and 60 for women. Payments to retired employees are based on the individually accumulated contribution and the fund performance during active employment. The annual pension payments are based on a discount rate of 6% and the remaining life expectancy at the time of retirement. These annuities grow by 1.5% annually to compensate for inflation. Hence, the wealth of the pension fund must grow by 7.5% a year to match liability commitments. Another output of the computations is the expected annual net cash flow of plan contributions minus payments. Because the number of pensioners is rising faster than plan contributions, these cash flows are negative and the plan is declining in size.

The model determines the optimal purchases and sales for each of N assets in each of T planning periods. Typical asset classes used at Innovest are US, Pacific, European, and emerging market equities and US, UK, Japanese, and European bonds. The objective is to maximize the concave risk-averse utility function “expected terminal wealth” less convex penalty costs subject to various linear constraints. The effect of such constraints is evaluated in the examples that follow, including Austria’s limits of 40% maximum in equities, 45% maximum in foreign securities, and 40% minimum in Eurobonds. The convex risk measure is approximated by a piecewise linear function, so the model is a multiperiod stochastic linear program. Typical targets that the model tries to achieve (and is penalized for if it does not) are for growth of 7.5% a year in wealth (the fund’s assets) and for portfolio performance returns to exceed benchmarks. Excess wealth is placed into surplus reserves, and a portion of the excess is paid out in succeeding years.

Table 1 Russell business engineering models

Model	Type of application	Year delivered	Number of scenarios	Computer hardware
Russell–Yasuch (Tokyo)	Property and casualty insurance	1991	256	IBM RISC 6000
Mitsubishi Trust (Tokyo)	Pension consulting	1994	2000	IBM RISC 6000 with parallel processors
Swiss Bank Corp. (Basle)	Pension consulting	1996	8000	IBM UNIX2
Daido Life Insurance Company (Tokyo)	Life insurance	1997	25 600	IBM PC
Banca Fideuram (Rome)	Assets only (personal)	1997	10 000	IBM UNIX2 and PC
Consulting Clients	Assets only (institutional)	1998	Various	IBM UNIX2 and PC

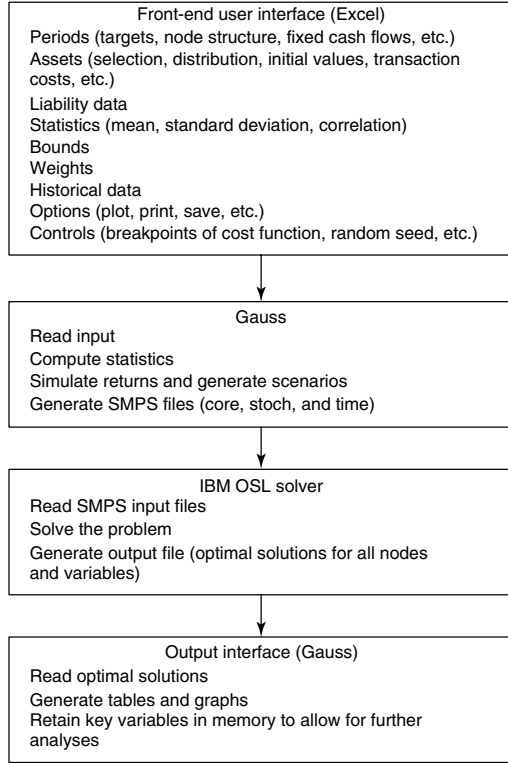


Figure 4 Elements of InnoALM [Reproduced from Geyer and Ziemba, 2008. © INFORMS.]

The elements of InnoALM are described in Figure 4. The interface to read in data and problem elements uses Excel. Statistical calculations use the program Gauss, and these data are fed into the IBM OSL solver, which generates the optimal solution to the stochastic program. The output uses Gauss to generate various tables and graphs and retains key variables in memory to allow for future modeling calculations.

Formulation of InnoALM as a Multistage Stochastic Linear Programming Model

The nonnegative decision variables are wealth (after transactions) W_{it} , and purchases P_{it} and sales S_{it} for each asset ($i = 1, \dots, N$). Purchases and sales are in periods $t = 0, \dots, T - 1$. Purchases and sales are scenario dependent except for $t = 0$.

Wealth accumulates over time for a T period model according to

$$W_{i0} = W_i^{\text{init}} + P_{i0} - S_{i0}, \quad t = 0 \quad (1)$$

$$\tilde{W}_{i1} = \tilde{R}_{i1} W_{i0} + \tilde{P}_{i1} - \tilde{S}_{i1}, \quad t = 1 \quad (2)$$

$$\tilde{W}_{it} = \tilde{R}_{it} \tilde{W}_{i,t-1} + \tilde{P}_{it} - \tilde{S}_{it}, \quad t = 2, \dots, T - 1 \quad (3)$$

and

$$\tilde{W}_{iT} = \tilde{R}_{iT} \tilde{W}_{i,T-1}, \quad t = T \quad (4)$$

where W_i^{init} is the prespecified initial value of asset i . There is no uncertainty in the initialization period $t = 0$. Tildes denote scenario-dependent random parameters or decision variables. Returns are associated with time intervals. $\tilde{R}_{it}$ ($t = 1, \dots, T$) are the (random) gross returns for asset i between $t - 1$ and t . The scenario generation and statistical properties of returns are discussed below.

The budget constraints are

$$\sum_{i=1}^N P_{i0}(1 + tcp_i) = \sum_{i=1}^N S_{i0}(1 - tcs_i) + C_0 \quad t = 0 \quad (5)$$

and

$$\sum_{i=1}^N \tilde{P}_{it}(1 + tcp_i) = \sum_{i=1}^N \tilde{S}_{it}(1 - tcs_i) + C_t, \quad t = 1, \dots, T - 1 \quad (6)$$

where tcp_i and tcs_i denote asset-specific linear transaction costs for purchases and sales, and C_t is the fixed (nonrandom) net cashflow (inflow if positive).

Portfolio weights can be constrained over linear combinations (subsets) of assets or individual assets via

$$\sum_{i \in U} \tilde{W}_{it} - \theta_U \sum_{i=1}^N \tilde{W}_{it} \leq 0 \quad (7)$$

and

$$-\sum_{i \in L} \tilde{W}_{it} + \theta_L \sum_{i=1}^N \tilde{W}_{it} \leq 0, \quad t = 1, \dots, T - 1 \quad (8)$$

where θ_U is the maximum percentage and θ_L is the minimum percentage of the subsets U and L of assets

$i = 1, \dots, N$ included in the restrictions. θ_U 's, θ_L 's, U 's and L 's may be time dependent.

Risk is measured as a weighted discounted convex function of target violation shortfalls of various types in various periods. In a typical application, the deterministic wealth target $\bar{W}_t$ is assumed to grow by 7.5% in each year. The wealth targets are modeled via

$$\sum_{i=1}^N (\tilde{W}_{it} - \tilde{P}_{it} + \tilde{S}_{it}) + \tilde{M}_t^W \geq \tilde{W}_t, \quad t = 1, \dots, T \quad (9)$$

where $\tilde{M}_t^W$ are nonnegative wealth-target shortfall variables. The shortfall is penalized using a piecewise linear convex risk measure. Stochastic benchmark goals can also be set by the user and are similarly penalized for underachievement. The benchmark target $\tilde{B}_t$ is scenario dependent. It is based on stochastic asset returns and fixed asset weights α_i defining the benchmark portfolio

$$\tilde{B}_t = W_0 \sum_{j=1}^t \sum_{i=1}^N \alpha_i \tilde{R}_{ij} \quad (10)$$

The corresponding shortfall constraints are

$$\sum_{i=1}^N (\tilde{W}_{it} - \tilde{P}_{it} + \tilde{S}_{it}) + \tilde{M}_t^B \geq \tilde{B}_t \quad t = 1, \dots, T \quad (11)$$

where $\tilde{M}_t^B$ is the benchmark-target shortfall. These shortfalls are also penalized with a piecewise linear convex risk measure. If total wealth exceeds the target, a fraction $\gamma = 10\%$ of the exceeding amount is allocated to a reserve account and invested in the same way as other available funds. However, the wealth targets at future stages are adjusted. Additional nonnegative decision variables $\tilde{D}_t$ are introduced and the wealth-target constraints become

$$\begin{aligned} & \sum_{i=1}^N (\tilde{W}_{it} - \tilde{P}_{it} + \tilde{S}_{it}) - \tilde{D}_t + \tilde{M}_t^W \\ & = \bar{W}_t + \sum_{j=1}^{t-1} \gamma \tilde{D}_{t-j}, \quad t = 1, \dots, T-1 \end{aligned} \quad (12)$$

where $\tilde{D}_1 = 0$.

Since pension payments are based on wealth levels, increasing these levels increases pension payments. The reserves provide security for the pension plan's increase in pension payments at each future stage.

The pension plan's objective function is to maximize the expected discounted value of terminal wealth in period T net of the expected discounted penalty costs over the horizon from the convex risk measures $c_k(\cdot)$ for the wealth- and benchmark-targets, respectively,

$$\begin{aligned} & \text{Max} E \left[d_T \sum_{i=1}^N \tilde{W}_{iT} - \lambda \sum_{t=1}^T d_t w_t \right. \\ & \quad \left. \times \left(\sum_{k \in \{W, B\}} v_k c_k(\tilde{M}_t^k) \right) \right] \end{aligned} \quad (13)$$

Expectation is over T period scenarios S_T . The discount factors d_t are related to the interest rate r by $d_t = (1+r)^{-t}$. Usually r is taken to be the three- or six-month Treasury-bill rate. v_k 's are weights for the wealth- and benchmark-shortfalls and the w_t 's are weights for the weighted sum of shortfalls at each stage normalized via

$$\sum_{k \in \{W, B\}} v_k = 1 \quad \text{and} \quad \sum_{t=1}^T w_t = T \quad (14)$$

Such concave objective functions with convex risk measures date back to 1986 [22], were used in the Russell–Yasuda model [7] and are justified in an axiomatic sense in [25]. Nontechnical decision makers find the increasing penalty for target violations a good approach and easy to understand. Allocations are based on optimizing the stochastic linear program.

Applications

Geyer and Ziemba [12] consider four asset classes (European stocks, US stocks, European bonds, and US bonds) with five periods (six stages): twice 1 year, twice 2 years, and 4 years (10 years in total). They assumed discrete compounding, so the mean return for asset i in simulations is $\mu_i = \exp(\bar{y}_i) - 1$ where $\bar{y}_i$ is the mean, based on log returns. Using a 100–5–5–2–2 node structure, they generated 10 000

scenarios. Initial wealth is equal to 100 units, and the wealth target is assumed to grow at an annual rate of 7.5%. They used a risk-aversion index of $R_A = 4$, and the discount factor is equal to 5%, which roughly corresponds with a simple static mean–variance model to a standard 60/40 stock/bond pension fund mix [19].

Assumptions about the statistical properties of returns measured in nominal euros are based on a sample of monthly data from January 1970 for stocks and 1986 for bonds to September 2000. Summary statistics for monthly and annual log returns are shown in Table 2. The US and European equity means for the longer period 1970–2000 were much lower and slightly less volatile than those for the period 1986–2000. The monthly stock returns were nonnormal and negatively skewed. Monthly stock returns were fat tailed, whereas monthly bond returns were close to normal (the critical value of the Jarque–Bera test for $\alpha = 0.01$ is 9.2).

For long-term planning models such as InnoALM, with its one-year review period, however, properties of monthly returns are less relevant. Table 2 shows statistics for annual returns. Although average returns and volatilities remained about the same, one year of data is lost when annual returns are computed and the distributional properties changed significantly. There was negative skewness, but no evidence existed for fat tails in annual returns, except for European stocks (1970–2000) and US bonds.

The mean returns from this sample are comparable to the 1900–2000 101-year mean returns estimated by Dimson [9]. Their estimate of the nominal mean equity return was 12.0% for the United States and 13.6% for Germany and the United Kingdom (the simple average of the two countries means). They estimated mean of bond returns of 5.1% for the United States and 5.4% for Germany and the United Kingdom.

Assumptions about means, standard deviations, and correlations for the applications of InnoALM are listed in Table 3 and are based on the sample statistics presented in Table 4. Projecting future rates of returns from past data is difficult. The equity means from the 1970 to 2000 period are used because the 1986–2000 period had an exceptionally high performance of stocks that is not assumed to prevail in the long run.

The correlation matrices in Table 3 for the three different regimes are based on the regression approach of Solnik *et al.* [29]. Moving average estimates of correlations among all assets are functions of standard deviations of US equity returns. The estimated regression equations are then used to predict the correlations in the three regimes shown in Table 3. Results for the estimated regression equations appear in Table 4. Three regimes are considered, and the assumption is that 10% of the time, equity markets are extremely volatile; 20% of the time, markets are characterized by high volatility; and 70% of the time, markets are normal. The 35% quantile of US equity return volatility defines “normal”

Table 2 Statistical properties of asset returns

	European stocks		US stocks		European bonds	US bonds
Returns	1/70–9/00	1/86–9/00	1/70–9/00	1/86–9/00	1/86–9/00	1/86–9/00
<i>Monthly</i>						
Mean (%) ^(a)	10.60	13.30	10.70	14.80	6.50	7.20
Standard deviation (%) ^(a)	16.10	17.40	19.0	20.200	3.70	11.3
Skewness	−0.90	−1.43	−0.72	−1.04	−0.50	0.52
Kurtosis	7.05	8.43	5.79	7.09	3.25	3.30
Jarque–Bera test	302.60	277.30	151.90	155.6	7.70	8.50
<i>Annual</i>						
Mean (%)	11.10	13.30	11.0	15.20	6.50	6.90
Standard deviation (%)	17.20	16.20	20.10	18.40	4.80	12.10
Skewness	−0.53	−0.10	−0.23	−0.28	−0.20	−0.42
Kurtosis	3.23	2.28	2.56	2.45	2.25	2.26
Jarque–Bera test	17.40	3.90	6.20	4.20	5.00	8.70

Reproduced from Geyer and Ziemba, 2008. © INFORMS

^(a) Annualized

Table 3 Mean, standard deviation, and correlation assumptions

Asset class	European stocks	US stocks	European bonds	US bonds
<i>Normal periods (70% of the time)</i>				
US stocks	0.755			
European bonds	0.334	0.286		
US bonds	0.514	0.780	0.333	
Standard deviation	14.6%	17.3%	3.3%	10.9%
<i>High volatility (20% of the time)</i>				
US stocks	0.786			
European bonds	0.171	0.100		
US bonds	0.435	0.715	0.159	
Standard deviation	19.2%	21.1%	4.1%	12.4%
<i>Extreme periods (10% of the time)</i>				
US stocks	0.832			
European bonds	−0.075	−0.182		
US bonds	0.315	0.618	−0.104	
Standard deviation	21.7%	27.1%	4.4%	12.9%
<i>Average period</i>				
US stocks	0.769			
European bonds	0.261	0.202		
US bonds	0.478	0.751	0.255	
Standard deviation	16.4%	19.3%	3.6%	11.4%
<i>All periods</i>				
Mean	10.6%	10.7%	6.5%	7.2%

Reproduced from Geyer and Ziemba, 2008. © INFORMS

Table 4 Regression equations relating asset correlations and US stock return volatility

Correlation	Constant	Slope with respect to US stock volatility	<i>t</i> -Statistic of slope	<i>R</i> ²
European stocks–US stocks	0.62	2.7	6.5	0.23
European stocks–European bonds	1.05	−14.4	−16.9	0.67
European stocks–US bonds	0.86	−7.0	−9.7	0.40
US stocks–European bonds	1.11	−16.5	−25.2	0.82
US stocks–US bonds	1.07	−5.7	−11.2	0.48
European bonds–US bonds	1.10	−15.4	−12.8	0.54

Reproduced from Geyer and Ziemba, 2008. © INFORMS

periods. “Highly volatile” periods are based on the 80% volatility quantile and “extreme” periods on the 95% quantile. The associated correlations reflect the return relationships that typically prevailed during those market conditions. The correlations show a distinct pattern across the three regimes. Correlations among stocks tend to increase as stock return volatility rises, whereas the correlations between stocks and bonds tend to decrease. European bonds may serve as a hedge for equities during extremely volatile periods because bond and stock returns, which are usually positively correlated, are then negatively correlated.

This is a major reason using scenario-dependent correlation matrices is a major advance over the sensitivity of using one correlation matrix as is usually assumed.

Optimal portfolios were calculated for seven cases—with and without mixing of correlations and with normal, *t*-, and historical distributions. The “mixing” cases NM, TM, and HM use mixing correlations. Case NM assumes normal distributions for all assets. Case HM uses the historical distributions of each asset. Case TM assumes *t*-distributions with five degrees of freedom for stock returns, whereas bond returns are assumed to have normal

Table 5 Optimal initial asset weights at Stage 1 by case

Case	European stocks	US stocks	European bonds	US bonds
Single-period, mean–variance optimal weights (average periods)	34.8%	9.6%	55.6%	0.0%
NA: No-mixing (average periods) normal distributions	27.2	10.5	62.3	0.0
HA: No-mixing (average periods) historical distributions	40.0	4.1	55.9	0.0
TA: No-mixing (average periods) t -distributions for stocks	44.2	1.1	54.7	0.0
NM: Mixing correlations normal distributions	47.0	27.6	25.4	0.0
HM: Mixing correlations historical distributions	37.9	25.2	36.8	0.0
TM: Mixing correlations t -distributions for stocks	53.4	11.1	35.5	0.0
TMC: Mixing correlations historical distributions; constraints on asset weights	35.1	4.9	60.0	0.0

Reproduced from Geyer and Ziemba, 2008. © INFORMS

distributions. The “average” cases NA, HA, and TA use the same distribution assumptions, with no mixing of correlation matrices. Instead, the correlations and standard deviations used in these cases correspond to an “average” period in which 10, 20, and 70% weights are used to compute the averages of correlations and standard deviations in the three different regimes. Comparisons of the average (A) cases and mixing (M) cases are mainly intended to investigate the effect of mixing correlations. TMC maintains all assumptions of case TM but uses Austria’s constraints on asset weights (see Table 5). Eurobonds must be at least 40% and equity at most 40%, and these constraints are binding. Tables 5 and 6 exhibit a distinct pattern: the mixing-correlation cases initially assign a much lower weight to European bonds than the average-period cases. Single-period, mean–variance optimization, and average-period cases (NA, HA, and TA) suggest an approximate 45%/55% stock/bond mix. The mixing-correlation cases (NM, HM, and TM) imply a 65%/35% stock/bond mix. Investing in US bonds is not optimal at Stage 1 in any of the cases, an apparent result of the relatively high volatility of US bonds.

Table 6 shows that the distinction between A and M cases becomes less pronounced over time. European equities, however, still have a consistently higher weight in the mixing cases than in the no-mixing cases. This higher weight is mainly at the expense of Eurobonds. In general, the proportion of equities at the final stage is much higher than in the first stage. This result may be explained by the fact that the expected portfolio wealth at later stages is far above the target wealth level (206.1 at Stage 6), and the higher risk associated with stocks

is less important. The constraints in case TMC lead to lower expected portfolio wealth throughout the horizon and to a higher shortfall probability than in any other case. Calculations show that the initial wealth would have to be 35% higher to compensate for the loss in terminal expected wealth stemming from those constraints. In all cases, the optimal weight of equities is much higher than the historical 4.1% in Austria.

The expected terminal wealth levels and the shortfall probabilities at the final stage shown in Table 6 make the difference between mixing and no-mixing cases even clearer. The mixing-correlation cases yield higher levels of terminal wealth and lower shortfall probabilities.

If the level of portfolio wealth exceeds the target, the surplus, $\bar{D}_j$, is allocated to a reserve account. The reserves in t are computed from $\sum_{j=1}^t \bar{D}_j$ and are shown in Table 6 for the final stage. These values are in monetary units given an initial wealth level of 100. They can be compared with the wealth target 206.1 at Stage 6. Expected reserves exceed the target level at the final stage by up to 16%. Depending on the scenario, the reserves can be as high as 1800. Their standard deviation (across scenarios) ranges from 5 at the first stage to 200 at the final stage. The constraints in case TMC lead to a much lower level of reserves compared with the other cases, which implies, in fact, less security against future increases in pension payments.

Optimal allocations, expected wealth, and shortfall probabilities are mainly affected by considering mixing correlations, but the type of distribution chosen has a smaller impact. This distinction is primarily the result of the higher proportion allocated to equities,

Table 6 Final stage results

Case	European stocks	US stocks	European bonds	US bonds	Expected terminal wealth	Expected reserves at Stage 6	Probability of target shortfall
NA	34.3%	49.6%	11.7%	4.4%	328.9	202.8	11.2%
HA	33.5	48.1	13.6	4.8	328.9	205.2	13.7
TA	35.5	50.2	11.4	2.9	327.9	202.2	10.9
NM	38.0	49.7	8.3	4.0	349.8	240.1	9.3
HM	39.3	46.9	10.1	3.7	349.1	235.2	10.0
TM	38.1	51.5	7.4	2.9	342.8	226.6	8.3
TMC	20.4	20.8	46.3	12.4	253.1	86.9	16.1

Reproduced from Geyer and Ziemba, 2008. © INFORMS

if different market conditions are taken into account by mixing correlations.

The results of any asset-allocation strategy crucially depend on the expected returns. Geyer and Ziemba investigated the effect by parameterizing the forecasted expected equity returns. Assume that an econometric model forecasts that the future return for US equities is between 5 and 15%. The mean of European equities is adjusted accordingly so that the ratio of equity means and the mean bond returns shown in Table 6 is maintained. Geyer and Ziemba retain all other assumptions of case NM (normal distribution and mixing correlations). As expected, the results are sensitive to the choice of the mean return; see [8] and [18, 19]. If the mean return for US stocks is assumed to equal the long-run mean of 12%, as estimated in [9], the model yields an optimal weight for equities of 100%. A mean return for US stocks of 9%, however, implies an optimal weight of less than 30% for equities.

Positive effects on the pension fund performance induced by the stochastic, multiperiod planning approach will be realized only if the portfolio is dynamically rebalanced, as implied by the optimal scenario tree [12]. Figure 5 compares the cumulated monthly returns obtained from the rebalancing strategy for the two cases with a buy-and-hold strategy that assumes that the portfolio weights on January 1992 were fixed at the optimal TM weights throughout the test period. In comparison to the buy-and-hold strategy or the performance using TA results, for which rebalancing does not account for different correlation and volatility regimes, rebalancing on the basis of the optimal TM scenario tree provided a substantial gain.

The model, developed in 2000, proved to be very useful. In 2006, Konrad Kontriner (Member of the

Board) and Wolfgang Herold (Senior Risk Strategist) of Innovest stated:

The InnoALM model... has become the only consistently implemented and fully integrated proprietary tool for assessing pension allocation issues within Siemens AG worldwide. [...] The key elements that make InnoALM superior to other consulting models are the flexibility to adopt individual constraints and target functions in combination with the broad and deep array of results, which allows to investigate individual, path dependent behavior of assets and liabilities as well as scenario based and Monte-Carlo like risk assessment of both sides. [...] The implementation of a scenario based asset allocation model will lead to more flexible allocation restraints that will allow for more risk tolerance and will ultimately result in better long term investment performance.

Conclusions

Some key points concerning the stochastic programming approach to asset-liability management are as follows:

- *Point 1*: Means are by far the most important part of the distribution of returns, especially the direction. One must estimate future means well or one can quickly travel in the wrong direction, which usually leads to losses or underperformance, or complete disaster if one is overlevered.
- *Point 2*: Mean–variance models are useful as a basic guideline when one is in an assets-only situation. Professionals adjust means using mean-reversion, James–Stein, or truncated estimators and constrain output weights. Asset positions should not be changed unless the advantage of the change is significant. Mean–variance analysis should not be used with liabilities and other

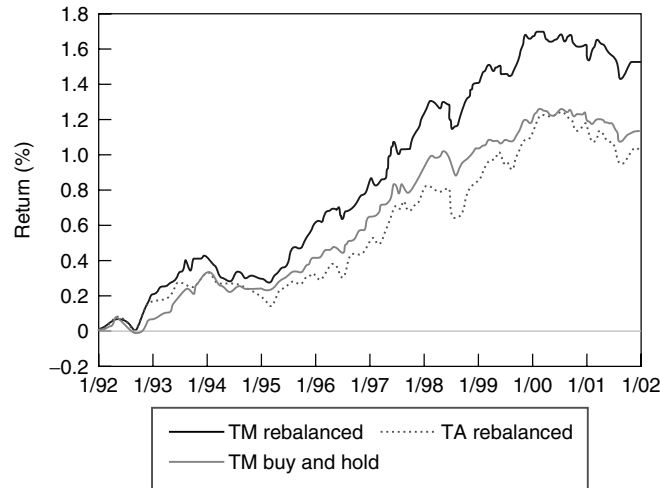


Figure 5 Cumulative monthly returns for different strategies, 1992–2002 [Reproduced from Geyer and Ziemba, 2008. © INFORMS.]

major market imperfections, except as a first test analysis.

- *Point 3:* Trouble arises when one overbets and a bad scenario occurs. Thus, overbetting should not be done when there is any possibility of a bad scenario occurring, unless the bet is protected by some type of hedge or stop loss.
- *Point 4:* Trouble is exacerbated when the expected diversification does not hold in the scenario that occurs. Thus, one must use scenario-dependent correlation matrices because simulations around historical correlation matrices are inadequate for extreme scenarios.
- *Point 5:* When a large decline in the stock market occurs, the positive correlation between stocks and bonds fails and they become negatively correlated. Thus, when the mean of the stock market is negative, bonds are more attractive, as is cash.
- *Point 6:* Stochastic programming scenario-based models are useful when one wants to look at aggregate overall decisions, with liabilities, liquidity, taxes, policy, legal, and other constraints, and have targets and goals one wants to achieve. It, thus, pays to make a complex stochastic programming model when a lot is at stake and the essential problem has many complications.
- *Point 7:* Other approaches, such as continuous-time finance, decision-rule-based stochastic programming, control theory, and so on, are useful

for problem insights and theoretical results. However in actual use, they may lead to disaster unless modified. Black and Scholes [2] option pricing theory infers that one can hedge perfectly with lognormal assets, which can lead to overbetting. Fat tails and jumps arise frequently and can occur without warning. Thus, one should be careful of the assumptions, including implicit ones, of theoretical models. The results should be used with caution no matter how complex and elegant the math or how smart or famous the author. Remember, one has to be very smart to lose millions and even smarter to lose billions.

- *Point 8:* One should not be concerned with getting all the scenarios exactly right when using stochastic programming models. Instead, one should worry about having the problem periods laid out reasonably and make sure the scenarios basically cover the means, the tails, and the chance of what could happen. If the current situation has never occurred before, use one that is similar to add scenarios. For a crisis in Brazil, use Russian crisis data, for example. The results of stochastic programming will give one good advice when times are normal and keep one out of severe trouble when times are bad. Those using stochastic programming models may lose 5, 10, or 15%, but they will not lose 50, 70, or 95%, as some investors and hedge funds have. Thus, if the

scenarios are more or less accurate and the problem elements are reasonably modeled, stochastic programming will give good advice. One may slightly outperform in normal markets, but one will greatly outperform in bad markets when other approaches may blow up.

- *Point 9:* Advances in computing power and modeling expertise have made stochastic programming modeling much less expensive to implement. Such models, which are still complex and require approximately six months to develop and test, cost a couple of hundred thousand dollars. A small team can make a model for a complex organization quite quickly at fairly low cost compared with what is at stake.
- *Point 10:* Eventually, as more disasters occur and more successful stochastic programming models are built and used, they will become popular. Thus, the ultimate goal is to have them in regulations, such as VaR. Although VaR does more good than harm, its safety is questionable in many applications. Conditional VaR is an improvement, but for most people and organizations, the nonattainment of goals is more than proportional (i.e., convex) in the nonattainment.

References

- [1] Abaffy, J., Bertocchi, M., Dupačová, J. & Moriggia, V. (2000). On generating scenarios for bond portfolios, *Bulletin of the Czech Economic Society* February, 3–27.
- [2] Black, F. & Scholes, M. (1973). The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.
- [3] Brennan, M.J. & Schwartz, E.S. (1998). The use of Treasury bill futures in strategies asset allocation program, in *Worldwide Asset and Liability Modeling*, W.T. Ziemba & J.M. Mulvey, eds, Cambridge University Press, pp. 205–228.
- [4] Campbell, J.Y. & Viceira, L.M. (2002). *Strategic Asset Allocation: Portfolio Choice for Long-term Investors*, Oxford University Press.
- [5] Cariño, D.R., Kent, T., Myers, D.H., Stacy, C., Sylvanus, M., Turner, A.L., Watanabe, K. & Ziemba, W.T. (1994). The Russell-Yasuda Kasai model: an asset/liability model for a Japanese insurance company using multistage stochastic programming, *Interfaces* **24**, 29–49.
- [6] Cariño, D., Myers, R. & Ziemba, W.T. (1998). Concepts, technical issues and uses of the Russell-Yasuda Kasai financial planning model, *Operations Research* **46**, 450–462.
- [7] Cariño, D. & Ziemba, W.T. (1998). Formulation of Russell-Yasuda Kasai financial planning model, *Operations Research* **46**, 433–449.
- [8] Chopra, V.K. & Ziemba, W.T. (1993). The effect of errors in mean, variance and co-variance estimates on optimal portfolio choice, *Journal of Portfolio Management* **19**, 6–11.
- [9] Dimson, E., Marsh, P. & Staunton, M. (2002). *Triumph of the Optimists*. Princeton University Press.
- [10] Dupačová, J., Consigli, G. & Wallace, S.W. (2000). Scenarios for multistage stochastic programs, *Annals of Operations Research* **100**, 25–53.
- [11] Fleten, S.-E., Høyland, K. & Wallace, S. (2002). The performance of stochastic dynamic and fixed mix portfolio models, *European Journal of Operational Research* **140**(1), 37–49.
- [12] Geyer, A. & Ziemba, W.T. (2008). The innovest Austrian pension fund financial planning model InnoALM, *Operations Research* **56**(4), 797–810.
- [13] Grauer, R.R. & Hakansson, N.H. (1998). On naive approaches to timing the market: the empirical probability assessment approach with an inflation adapter, in *World Wide Asset and Liability Modeling*, W.T. Ziemba, & J.M. Mulvey, eds, Cambridge University Press, pp. 149–181.
- [14] Greenspan, A. (1998). *Speech before the Committee on Banking and Social Services*, US House of Representatives.
- [15] Henriques, D.B. (1991). A better way to back your assets, *New York Times* March, Section 3 (Business), 11.
- [16] Hochreiter, R. & Pflug, G.C. (2003). *Scenario Generation for Stochastic Multi-stage Decision Processes as Facility Location Problems*, Technical report. Department of Statistics and Decision Support Systems, University of Vienna.
- [17] Høyland, K. & Wallace, S.W. (2001). Generating scenario trees for multistage decision problems, *Management Science* **47**, 295–307.
- [18] Kallberg, J.G. & Ziemba, W.T. (1981). Remarks on optimal portfolio selection, in *Methods of Operations Research*, G. Bamberg & O. Optiz, eds, Gunn and Hain, Oelgeschlager, Vol. 44, pp. 507–520.
- [19] Kallberg, J.G. & Ziemba, W.T. (1983). Comparison of alternative utility functions in portfolio selection problems, *Management Science* **29**, 1257–1276.
- [20] Keim, D. & Ziemba, W.T. (eds) (2000). *Security Market Imperfections in Worldwide Equity Markets*, Cambridge University Press.
- [21] Keynes, J.M. (1938). *Investment Policy Report on the Chest Fund*, Kings College, Cambridge, UK.
- [22] Kusy, M.I. & Ziemba, W.T. (1986). A bank asset and liability management model, *Operations Research* **34**(3), 356–376.
- [23] MacLean, L. & Ziemba, W.T. (2006). Capital growth theory and practice, in *Handbook of Asset and Liability Modeling*, S.A. Zenios & W.T. Ziemba, eds, On theory and methodology, North Holland, Amsterdam, Vol. 1, pp. 429–475.

- [24] Markowitz, H.M. & Perold, A. (1981). Portfolio analysis with factors and scenarios, *Journal of Finance* **36**, 871–877.
- [25] Rockafellar, T. & Ziemba, W.T. (2000). *Modified Risk Measures and Acceptance Sets*. Working Paper, University of Washington.
- [26] Rosen, D. & Zenios, S.A. (2006). Enterprise-wide asset and liability management: issues, institutions, and models, in *Theory and Methodology, Handbook of Asset and Liability Modeling*, S.A. Zenios & W.T. Ziemba, eds, North-Holland, Amsterdam, Vol. 1, pp. 1–23.
- [27] Rudolf, M. & Ziemba, W.T. (2004). Intertemporal surplus management, *Journal of Economic Dynamics and Control* **28**(4), 975–990.
- [28] Shapiro, J.F. (1988). Stochastic programming models for dedicated portfolio selection, in *Mathematical Models for Decision Support*, G.B. Mitra, ed., ASI Series, Springer-Verlag, Berlin, Vol. F48, pp. 587–611.
- [29] Solnik, B., Bouccelle, C. & Le Fur, Y. (1996). International market correlation and volatility, *Financial Analysts Journal* **52**, 17–34.
- [30] Thorp, E.O. (2006). The Kelly criterion in blackjack, sports betting and the stock market, in *Handbook of Asset and Liability Modeling*, S.A. Zenios & W.T. Ziemba, eds, On theory and methodology, North Holland, Amsterdam, Vol. 1, 385–428.
- [31] Ziemba, W.T. (2007). The Russell Yasuda Kasai, InnoALM and related models for pensions, insurance companies and high net worth individuals, in *Handbook of Asset and Liability Management*, S.A. Zenios & W.T. Ziemba, eds, Applications and Case Studies, North Holland, Amsterdam, The Netherlands, Vol. 2, pp. 861–962.
- [32] Ziemba, W.T. & Mulvey, J.M. (eds) (1998). *World Wide Asset and Liability Modeling*, Cambridge University Press.

Further Reading

- Kallberg, J.G., White, R. & Ziemba, W.T. (1982). Short term financial planning under uncertainty. *Management Science* **XXVIII**, 670–682.
- Ziemba, W.T. (2003). *The Stochastic Programming Approach to Asset Liability and Wealth Management*, AIMR, Charlottesville, Virginia.
- Ziemba, R.E.S. & Ziemba, W.T. (2007). *Scenarios for Risk management and Global Investment Strategies*, John Wiley & Sons, Chichester, England.

Related Articles

Expected Shortfall; Fixed Mix Strategy; Kelly Problem; Mean–Variance Hedging; Mutual Funds; Performance Measures; Risk–Return Analysis; Stochastic Control; Stress Testing; Stylized Properties of Asset Returns; Value-at-Risk.

WILLIAM T. ZIEMBA

Passport Options

A passport option is a call option on a traded account, giving the buyer the right to choose his/her strategy (within some limits) and to keep all the gains generated from this strategy while letting the seller cover any losses. In contrast to other exotic options, the value of the traded account (also termed the *wealth of the buyer*) is treated as a security underlying an option. Since the buyer of this option retains profits but cannot make losses, it can be thought of as an option achieving the best total payoff for the buyer. For this reason, the passport option is sometimes called the *perfect trader option* since if a trader buys such an option, he/she cannot lose money beyond the option premium.

The passport option was first sold in foreign exchange markets in the mid-1990s and was later extended to other assets including bonds and interest rate futures. The first study on passport options was done by Hyer *et al.* [10] in 1997 and was heralded by the funds management industry to be a breakthrough in providing derivative-based tools for principal protection of actively managed funds. The resulting contract acts like a stop-loss, limiting the buyer's downside to the premium paid.

Definition and Model

The buyer of a passport option pays a premium and is allowed to switch between short and long positions in an asset over its life. The risky asset in which positions may be taken is denoted by P . The absolute value of the number of units of asset P held at time t , q_t , must be less than or equal to some constant L . We can scale L out of the problem, and hence the restriction becomes $|q| \leq 1$. Gains or losses from this resulting trading strategy are accumulated, and when the option expires at T , the buyer is given any profits from the strategy, but does not pay for any losses.

If we denote the wealth at time t , generated by the trading strategy q_t , by X_t , the passport option payoff at T is

$$\max[0; X_T] = X_T^+ \quad (1)$$

The payoff of the passport option is contingent on the strategy q_t undertaken by the buyer, and thus we cannot define a unique price simply by taking

the discounted risk-neutral expectation of the payoff in the usual way (*see*, for instance, **Risk-neutral Pricing**). We need to assume that the buyer attempts to maximize profits, and thus the seller charges the maximum price over all allowable strategies. This leads to the formulation for the option price:

$$V(t, P_t, X_t) = \sup_{|q_t| \leq 1} e^{-r(T-t)} \mathbb{E}_t X_T^+ \quad (2)$$

where $\mathbb{E}_t$ denotes risk-neutral expectation conditional on information at time t . Note that if the optimal strategy is not followed by the buyer, then the seller has charged more than necessary to hedge and will in fact superhedge or super-replicate.

To define a useful model, we must elaborate on how the wealth or the value of the traded account is generated. The value of this account X_t corresponding to strategy q_t follows

$$dX_t = q_t dP_t + \mu(X_t - q_t P_t) dt \quad (3)$$

with initial value X_0 . The interpretation is that the buyer invests in q units of the risky asset P and the remainder earns interest rate μ (which could be different from the riskless rate of r). The dynamics of the risky asset P will be taken to be an exponential Brownian motion for the time being (we generalize this later):

$$dP_t = \nu P_t dt + \sigma P_t dW_t \quad (4)$$

where ν is the drift (r if a nondividend paying stock, $r - \delta$ if a dividend paying stock, or $r - f$ if foreign currency), σ is the volatility of the asset, and W is a standard Brownian motion.

The Hamilton–Jacobi–Bellman (HJB) equation (*see Expected Utility Maximization*) for the stochastic control (*see Stochastic Control*) problem (2) can be written as

$$\sup_{|q_t| \leq 1} \left[-rV + \dot{V} + V_P P \nu + (\mu X + q P (\nu - \mu)) V_X + \frac{1}{2} \sigma^2 P^2 (V_{PP} + q^2 V_{XX} + 2q V_{PX}) \right] = 0 \quad (5)$$

with boundary condition $V(T, p, x) = x^+$. The above equality holds for $q_t = q_t^*$, where q_t^* attains the supremum in equation (2). This is a nonlinear parabolic partial differential equation (PDE) in two space variables.

The analysis can be simplified if we make the additional assumption that $\mu = \nu = r$, which is the so-called symmetric case. This is assuming that money in the bank earns the riskless rate and the asset grows at rate r . Under this assumption, an explicit price for the passport option and an expression for the optimal strategy q_t^* can be found. In this case, the value of the traded account when discounted by the riskless rate becomes a martingale under the risk-neutral pricing measure. This is the key point that is used in the mathematics to simplify the pricing.

Note that under the symmetric case, the stock may not pay dividends or be a foreign currency. These cases fit into the nonsymmetric framework and must be priced *via* numerical solution of the PDE. Such problems were studied by Andersen *et al.* [2] and Ahn *et al.* [1]. For this article, we continue to be in the symmetric framework to enable us to illustrate the main points of passport pricing.

Pricing

There are two main techniques used in the study of passport option pricing: the partial differential equation (see **Partial Differential Equations**) approach and the probabilistic approach. When the model setup falls into the symmetric case, a closed-form solution is obtainable for the passport option under exponential Brownian motion *via* both of these techniques. There are two issues in solving the pricing problem: to identify the optimal strategy and to calculate the value *via* the expression in equation (2).

Optimal Strategy

We first give some justification for the optimal strategy. Under the symmetric framework, the optimal strategy is found to be given as

$$q_t^* = -\text{sgn}(X_t) \quad (6)$$

where

$$\text{sgn}(y) = \begin{cases} 1, & y > 0 \\ -1, & y \leq 0 \end{cases} \quad (7)$$

This implies the buyer should be long up to the maximum allowed when behind ($X_t \leq 0$) and short up to the maximum when ahead ($X_t > 0$). A solution of this form is called *bang-bang* as it

switches between the two extremes. This solution can be shown to be optimal in a variety of ways. Hyer *et al.* [10], Andersen *et al.* [2], and Ahn *et al.* [1] reduce the dimension of the HJB equation by writing $V(t, P, X) = P\phi(t, z)$ with $z = X/P$ and then from this it was deduced that the strategy is optimal. In contrast, Shreve and Vecer [12] (see also Vecer and Shreve [13]) make the same transformation and then apply a stochastic comparison theorem to achieve the result. Interestingly, the form of the optimal strategy in equation (6) remains unchanged if we generalize the asset price dynamics to allow for diffusion models (under the assumption that the diffusion coefficient is nondecreasing in S). This is shown in [7] using stochastic coupling techniques.

The Option Price

We also follow Henderson and Hobson [7] (and Henderson [5] as well as Delbaen and Yor [3]) in deriving a representation for the price of the passport option using probabilistic methods. The reader is referred to these sources for more technical details and assumptions. As mentioned above, probabilistic techniques allow us to use more general asset prices, namely, P may follow a diffusion. The discounted value of the traded account and the discounted asset price are denoted by $G_t = e^{-rt} X_t$ and $S_t = e^{-rt} P_t$, respectively. Then if

$$dS_t = S_t \sigma(S_t, t) dW_t \quad (8)$$

we have $dG_t = q_t dS_t$, and (under appropriate assumptions) G_t is a martingale under the risk-neutral measure. Using Tanaka's formula and the Skorokhod lemma, the passport option price in equation (2) can be represented as

$$\begin{aligned} \sup_{|q_t| \leq 1} e^{-r(T-t)} \mathbb{E}_t X_T^+ &= \sup_{|q_t| \leq 1} e^{rt} \mathbb{E}_t G_T^+ \\ &= \sup_{|q_t| \leq 1} e^{rt} \left[G_t^+ + \frac{1}{2} \mathbb{E}_t L_{t,T}^G(0) \right] \\ &= \sup_{|q_t| \leq 1} e^{rt} \left[\frac{1}{2} \mathbb{E}_t (M_T^* - |G_t|)^+ \right. \\ &\quad \left. + G_t^+ \right] \end{aligned} \quad (9)$$

where $L_{t,T}^G(0)$ is the local time of G over time t to T (see **Local Times**) and $M_z^* =$

$\sup_{u \leq z} \int_t^u -q \operatorname{sgn}(G) dS$. Using the knowledge of the optimal strategy, the price is reduced to

$$\frac{1}{2} \mathbb{E}_t \left(e^{rt} \sup_{t \leq r \leq T} S_r - P_t - |X_t| \right)^+ + X_t^+ \quad (10)$$

The representation (10) holds for diffusion models with the assumption that the diffusion coefficient $S\sigma(S, t)$ is nondecreasing in S . Under the exponential Brownian motion model where $\sigma(S, t) = \sigma$, a constant, the passport price can be calculated explicitly to be

$$\begin{aligned} X_t^+ + \frac{1}{2} \Big[& P_t(N(d) - N(d - \sigma\sqrt{T-t})) \\ & + \sigma\sqrt{T-t}(N'(d) - dN(d)) \\ & - |X_t|N(d - \sigma\sqrt{T-t}) \Big] \end{aligned} \quad (11)$$

where N is the cumulative normal distribution function and $d = \frac{-\ln(1+|X_t|/P_t) + \frac{1}{2}\sigma^2(T-t)}{\sigma\sqrt{T-t}}$. The price in equation (11) was obtained by Andersen *et al.* [2] by solving a rescaled one-dimensional PDE, by Shreve and Vecer [12] using probabilistic techniques involving a change in measure and a direct derivation of the optimal strategy *via* Hajek's theorem [4], and by Henderson [5] and Henderson and Hobson [7] using the probabilistic arguments summarized here.

Related Options and Further Work

The representation of the passport option price in equation (10) shows that the price is closely related to the price of a lookback option (*see Lookback Options*). From this knowledge, we can deduce that the passport option has a high premium, since lookback options are expensive. This should be anticipated from the fact that a passport option gives the buyer downside protection and upside potential. As with lookback options, the high price of a passport option is its main drawback. The variants discussed below may be cheaper, but each of these limits the buyers trading differently.

Variants on the passport option have been developed and these options on a traded account allow the buyer to switch between positions α_t and β_t where $\alpha_t \leq \beta_t$ in the asset S . A passport option corresponds to the choice $\alpha_t = -1$ and $\beta_t = 1$. Other variants introduced in [12] are the vacation call [$\alpha_t = 0, \beta_t =$

1] and vacation put [$\alpha_t = -1, \beta_t = 0$]. They show that many options are a special case of options on a traded account, including European and American calls and puts and Asian options. Lipton [11] also discussed some similarities between the passport option and Asian options. (*See* articles **American Options** and **Call Options** for descriptions of European and American options and **Asian Options** for descriptions of Asian options.)

Passport options on futures and forwards contracts can be used to develop new hedging tools for commodity exposures. Typical hedging involves the sale or the purchase of futures or forwards, or vanilla options on the commodity. (*See*, for instance, **Commodity Risk; Commodity Forward Curve Modeling**.) These are European in nature and cannot be used to capture daily or weekly movements in future curves. A hedging strategy that allows for speculation on the upside, while retaining a downside hedge, involves a passport option on additional trading as well as a standard forward or futures hedge. This may appeal to the hedger, as they will not be any worse off than if they had hedged passively, but have upside also. Compared to buying a put option, this new strategy is expensive. However, after adjusting for cost, the new strategy has a higher probability of being in the money and thus is a better hedge against the worst outcomes. These ideas were developed in the work of Henderson *et al.* [9]. Further results on passport options are derived in [6] where coupling techniques and the mean comparison theorem of Hayek [4] are used to give option price comparison results. Henderson and Hobson [8] study pricing of passport options under stochastic volatility models. Delbaen and Yor [3] extend the probabilistic pricing argument to obtain a new theorem about martingales, study the case with nonzero interest rate *via* hitting times for Bessel processes, and solve the symmetric case when trading is permitted at discrete times. For results concerning discrete trading, see also [1].

References

- [1] Ahn, H., Penaud, A. & Wilmott, P. (1999). Various passport options and their valuation, *Applied Mathematical Finance* 6(4), 275–292.
- [2] Andersen L., Andreasen, J. & Brotherton-Ratcliffe, R. (1998). The passport option, *Journal of Computational Finance* 1(3), 15–36.
- [3] Delbaen, F. & Yor, M. (2002). Passport options, *Mathematical Finance* 12, 299–328.

- [4] Hajek, B. (1985). Mean stochastic comparison of diffusions, *Zeitschrift Wahrscheinlichkeitstheorie Verwandte Gebiete* **68**, 315–329.
- [5] Henderson, V. (1999). *A Probabilistic Approach to Passport Options*, PhD Thesis, University of Bath.
- [6] Henderson, V. (2000). Price comparison results and super-replication: an application to passport options, *Applied Stochastic Models in Business and Industry* **16**(4), 297–310.
- [7] Henderson, V. & Hobson, D. (2000). Local time, coupling and the Passport option, *Finance and Stochastics* **4**(1), 69–80.
- [8] Henderson, V. & Hobson, D. (2001). Passport options with stochastic volatility, *Applied Mathematical Finance* **8**(2), 97–119.
- [9] Henderson, V., Hobson, D. & Kentwell, G. (2002). A new class of commodity hedging strategies: a passport options approach, *International Journal of Theoretical and Applied Finance* **5**, 3.
- [10] Hyer, T., Lipton-Lifschitz, A. & Pugachevsky, D. (1997). Passport to success, *Risk* **10**(9), 127–131.
- [11] Lipton, A. (1999). Similarities via self-similarities, *Risk* September, 101–105.
- [12] Shreve, S. & Vecer, J. (2000). Options on a traded account: vacation calls, vacation puts and passport options, *Finance and Stochastics* **4**, 255–274.
- [13] Vecer, J. & Shreve, S. (2000). Upgrading your passport, *Risk* July, 81–83.

Related Articles

American Options; Asian Options; Call Options; Commodity Forward Curve Modeling; Commodity Risk; Expected Utility Maximization; Local Times; Lookback Options; Partial Differential Equations; Risk-neutral Pricing; Stochastic Control.

VICKY HENDERSON

Ruin Theory

The earliest ruin problem was formulated by de Moivre around 1730. A gambler has an initial fortune of u monetary units (an integer). Each time he plays (heads or tails in coin tossing, black or red at the roulette, *etc.*), his fortune goes one up or down according to whether he wins or loses. If the fortune becomes 0, the gambler is ruined. He decides *a priori* to stop gambling if his fortune reaches $v \geq u$. If the probability of a win is p , what is then $\psi_v(u)$, the probability of getting ruined as opposed to reach the target v ?

The classical ruin problem of insurance mathematics is rather similar: u is the initial reserve of the insurance company, and incoming premiums make the reserve go up, payments to insurance holders make it go down. One then considers $\psi(u)$, the ruin probability defined as the probability that the reserve ever becomes nonpositive (note that this is a one-barrier problem, whereas the gambler's ruin problem involves two barriers).

A general mathematical formulation of this and related types of problems is a stochastic process $\{R(t)\}$ (corresponding to the fortune of the gambler or the insurance company at time t) and the stopping times

$$\begin{aligned}\tau_v(u) &= \inf\{t : R(t) \notin (0, v) | R(0) = u\} \\ &= \inf\{t : R(t) \leq 0 \text{ or } R(t) \geq v | R(0) = u\} \\ \tau(u) &= \inf\{t : R(t) \leq 0 | R(0) = u\}\end{aligned}\quad (1)$$

In broad terms, ruin theory is concerned with deriving properties of these stopping times. In particular, in the two-barrier setting one typically has $\psi_v(u) < \infty$ and could ask for $\mathbb{P}(R(\tau_v(u)) \geq v)$ or $\mathbb{P}(R(\tau_v(u)) \leq 0)$. In the one-barrier setting, the main object of interest is the probability $\psi(u) = \mathbb{P}(\tau(u) < \infty)$ of eventual ruin. Other quantities of interest are the distribution of $\tau(u)$, needed to say something about the finite horizon ruin probability $\psi(u, T) = \mathbb{P}(\tau(u) \leq T)$, and the deficit $-R(\tau(u))$ after ruin (*see Gerber–Shiu Function*).

Ruin problems occur also in finance. A simple example is an equity default swap (*see Equity Swaps*) that pays a certain amount if a security price drops by a certain amount (say 50%) in a time period $[0, T]$. The price of the equity default

swap then involves the risk-neutral probability (*see Risk-neutral Pricing*) that this happens, which can easily be represented in the form $\psi(u, T)$. A few more examples occur in barrier options (*see Barrier Options*). For example, a digital down-and-in option will pay the amount 1 if the security price $Z(t)$ goes below L before T , but is above $K > L$ at time T ($Z(T) > K$). A standard down-and-in option will pay instead $(Z(T) - K)^+$ under the same conditions.

Explicit Solutions

In the rest of the article, we assume that $\{R(t)\}_{t \geq 0}$ has stationary independent increments (though there is a considerable literature on extensions such as Markovian regime, *see Regime-switching Models*, jumps at the epochs of a renewal process, *etc.*). Beyond the simple random walk in the gambler's ruin problem, key examples are Brownian motion (with drift μ and variance constant σ^2 , say) and the Cramér–Lundberg risk model

$$R(t) = u + ct - \sum_{i=1}^{N(t)} V_i \quad (2)$$

where c is the premium rate, N is the Poisson process of times of arrivals of claims, and $V_1, V_2, \dots$ are the claim sizes.

The gambler's ruin problem is easily solvable: conditioning upon the outcome of the first play gives

$$\psi_v(u) = p\psi_v(u+1) + (1-p)\psi_v(u-1) \quad (3)$$

when $0 < u < v$. Subject to the obvious boundary conditions $\psi_v(0) = 1$, $\psi_v(v) = 0$, this set of equations has a unique solution that can be seen to be

$$\psi_v(u) = \frac{[(1-p)/p]^v - [(1-p)/p]^u}{[(1-p)/p]^v - 1} \quad (4)$$

(for $p = 1/2$, this needs to be modified to $\psi_v(u) = 1 - u/v$).

What with other models? The simplest example is Brownian motion with drift μ and variance σ^2 , where conditioning upon $R(h)$ and letting $h \downarrow 0$ gives $\psi_v(u) = \mu\psi'_v(u) + \sigma^2\psi''_v(u)/2$. Together with $\psi_v(0) = 1$, $\psi_v(v) = 0$, this differential equation is easily solved to get

$$\psi_v(u) = \frac{e^{-\gamma u} - e^{-\gamma v}}{1 - e^{-\gamma v}} \quad \text{where } \gamma = 2\mu/\sigma^2 \quad (5)$$

for $\mu > 0$ and $\psi_v(u) = 1 - u/v$ for $\mu = 0$. A more elegant and general approach uses martingales (see **Martingales**). The Wald exponential martingale is

$$W(t) = W_\alpha(t) = \exp \{-\alpha R(t) - t\kappa(\alpha)\} \quad (6)$$

where $\kappa(\alpha) = \log \mathbb{E} \{-\alpha R(1)\}$ (the so-called Lévy exponent, see **Lévy Processes**). A particularly important choice of α is γ , the nonzero solution of $\kappa(\gamma) = 0$ and going under names such as the adjustment coefficient, the Cramér–Lundberg root and so on. Indeed, for this case $W(t) = \exp \{-\alpha R(t)\}$, and since $R(\tau_v(u)) = v$ when $R(\tau_v(u)) \geq v$ (and similarly $R(\tau_v(u)) = 0$ when $R(\tau_v(u)) \leq 0$) for Brownian motion, optional stopping gives

$$\begin{aligned} e^{-\gamma u} &= W(0) = \mathbb{E} W(\tau_v(u)) \\ &= \psi_v(u)e^0 + (1 - \psi_v(u))e^{-\gamma v} \end{aligned} \quad (7)$$

In the Brownian case, $\kappa(\alpha) = -\alpha\mu + \alpha^2\sigma^2/2$, and so $\gamma = 2\mu/\sigma^2$. Inserting in equation (7) and solving for $\psi_v(u)$, equation (5) follows easily. For the following, note that

$$\kappa(\alpha) = \lambda(\mathbb{E}e^{-\alpha V_i} - 1) + \alpha c \quad (8)$$

for the Cramér–Lundberg model (here γ is only explicit for a few examples, in particular exponential V_i , cf. equation (11) below).

For further generalizations, basically one needs to be able to say something about the overshoot $R(\tau_v(u)) - v$, given $R(\tau_v(u)) \geq v$ and the undershoot $-R(\tau_v(u))$ given $R(\tau_v(u)) \leq 0$. A basic example is the Cramér–Lundberg model with the V_i exponential, say, at rate μ . Then by the memoryless property of the exponential distribution, $-R(\tau_v(u))$, given $R(\tau_v(u)) \leq 0$ is again exponential(μ) so that

$$\mathbb{E}[\exp \{\gamma R(\tau_v(u))\} | R(\tau_v(u)) \leq 0] = \frac{\mu}{\mu + \gamma} \quad (9)$$

Since $R(\tau_v(u)) = v$ when $R(\tau_v(u)) \geq v$ (the process has no upward jumps), it follows that

$$\begin{aligned} e^{-\gamma u} &= W(0) = \mathbb{E} W(\tau_v(u)) \\ &= \psi_v(u) \frac{\mu}{\mu + \gamma} + (1 - \psi_v(u))e^{-\gamma v} \end{aligned} \quad (10)$$

However, if λ is the Poisson intensity, then

$$\kappa(\alpha) = \lambda \left(\frac{\mu}{\mu + \alpha} - 1 \right) + \alpha c \implies \gamma = \lambda/c - \mu \quad (11)$$

Solving equation (10) for $\psi_v(u)$ yields an explicit expression for $\psi_v(u)$, and letting $v \rightarrow \infty$ we then get one of the classical formulas in ruin theory,

$$\psi(u) = \frac{\lambda}{c\mu} \exp \{-(\lambda/c - \mu)u\} \quad (12)$$

in the exponential case.

A generalization of equation (10) is to let the V_i be phase-type, say with initial vector $\mathbf{v}$, phase generator $\mathbf{T}$ and exit rates $\mathbf{t}$. Then ([1 p. 227])

$$\psi(u) = \mathbf{v}_+ \exp \{(\mathbf{T} + \mathbf{t}\mathbf{v}_+)u\} \quad \text{where } \mathbf{v}_+ = -\lambda\mathbf{v}\mathbf{T}^{-1}/c \quad (13)$$

The derivation involves some additional steps, but again, the crucial feature is that one can control the undershoot: the distribution of $-R(\tau_v(u))$, given $R(\tau_v(u)) \leq 0$ is determined by the phase in which 0 is downcrossed. Thus we have a finite set of unknowns, which motivates the matrix form of equation (13).

Bounds and Symptotics

Formula (13) is more or less the end of the road with regard to explicit solutions, even for the simple Cramér–Lundberg model. Therefore, inequalities and approximations play a major role in the area.

The celebrated *Lundberg's inequality* states that

$$\psi(u) \leq e^{-\gamma u} \quad (14)$$

A simple proof of equation (14) (and many other inequalities and approximations) is to use change of measure: for each θ with $\kappa(\theta) < \infty$, consider the risk process with λ changed to $\lambda_\theta = \lambda \mathbb{E}e^{-\theta V_i}$ and the distribution of V_i given by $\mathbb{E}_\theta e^{\alpha V_i} = \mathbb{E}e^{(\alpha-\theta)V_i} / \mathbb{E}e^{-\theta V_i}$. Then

$$\begin{aligned} \psi(u) &= \mathbb{P}(\tau(u) < \infty) \\ &= \mathbb{E}_\theta \exp \{\theta[R(\tau(u)) - u] - \theta\kappa(\theta)\tau(u)\} \end{aligned} \quad (15)$$

Taking $\theta = \gamma$ and bounding $R(\tau(u))$ below by 0 immediately yields equation (14). A more elaborate argument involving the $\mathbb{P}_\theta$ -limit distribution of $R(\tau(u))$ gives the other main result in the area,

going under the name of the Cramér–Lundberg approximation,

$$\psi(u) \sim Ce^{-\gamma u}, u \rightarrow \infty, \text{ where } C = \frac{1 - \kappa'(0)/c}{\kappa'(\gamma)/c - 1} \quad (16)$$

The results (14) and (17), both require a light-tailed assumption, namely, the existence of γ (and for equation (17) in addition that $\kappa'(\gamma) < \infty$). With heavy tails, the situation is completely different. Assume that V is subexponential (see **Heavy Tails in Insurance; Heavy Tails**; for example, the tail is regularly varying, $\mathbb{P}(V > v) = L(v)/v^\alpha$ with $\alpha > 1$ and $L(\cdot)$ slowly varying). Then

$$\psi(u) \sim \frac{\lambda/\mathbb{E}V}{(c - \lambda/\mathbb{E}V)\mathbb{E}V} \int_u^\infty \mathbb{P}(V > y) dy, u \rightarrow \infty \quad (17)$$

a result that is often associated with the names (in alphabetical order) of Borovkov, Cohen, Embrechts, Pakes, Veraverbeke, and von Bahr.

Finite Horizon Ruin Probabilities

The density or the distribution function of $\tau(u)$ is essentially only explicit for Brownian motion: if $\sigma^2 = 1$, then

$$\begin{aligned} \mathbb{P}_\mu(\tau(u) \in dt) \\ = \frac{u}{\sqrt{2\pi}t^{3/2}} \exp\left\{-\mu u - \frac{1}{2}\left(\frac{u^2}{t} + \mu^2 t\right)\right\} \end{aligned} \quad (18)$$

However, if one is satisfied with transforms, the situation is somewhat better, at least for the Cramér–Lundberg model: with exponential claims, one has

$$\mathbb{E}[e^{-a\tau(u)}; \tau(u) < \infty] = e^{-\theta u}(1 - \theta/\mu) \quad (19)$$

where θ is the largest root of $\kappa(\theta) = ac$ or, equivalently, of the quadratic $\theta^2 + (\lambda/c - \mu + ac)\theta - ac\mu$. For general claim size distributions, one has to go to double transforms to get a reasonably simple for-

mula:

$$\begin{aligned} \int_0^\infty e^{bu} \mathbb{E}[e^{-a\tau(u)}; \tau(u) < \infty] du \\ = \frac{-ac/r(a) - \kappa(b)/b}{\kappa(b) + ac} \end{aligned} \quad (20)$$

where $r(a)$ is the smallest solution of $-ac = \kappa(r(a))$.

The asymptotics of the ruin time $\tau(u)$ is, however, well understood. In the light-tailed setting, $\tau(u)/u$ has limit (conditional upon $\tau(u) < \infty$) $1/\kappa'(\gamma)$ as $u \rightarrow \infty$, and there is further a central limit theorem for $(\tau(u) - u/\kappa'(\gamma))/u^{1/2}$. In the subexponential case, there is again a limit for $\tau(u)/c(u)$ for some suitable constants $c(u) \rightarrow \infty$, but the limit is random (e.g., Pareto in the regularly varying case where $c(u) = u$), not constant as for light tails. These statements translate easily into approximations for

$$\begin{aligned} \psi(u, t) = \mathbb{P}(\tau(u) \leq t) \\ = \psi(u) \mathbb{P}(\tau(u) \leq t | \tau(u) < \infty) \end{aligned} \quad (21)$$

Further approximations worth mentioning are the corrected diffusion approximation and the saddle-point approximation. For these, as well as the remaining topics discussed in this article, see Asmussen [1]. Some additional standard monographs in the area are Bühlmann [2], Gerber [3], and Grandell [4].

References

- [1] Asmussen, S. (2000). *Ruin Probabilities*, World Scientific, Singapore.
- [2] Bühlmann, H. (1976). *Mathematical Methods in Risk Theory*, Springer-Verlag.
- [3] Gerber, H.U. (1979). *An Introduction to Mathematical Risk Theory*, SS. Huebner Foundation Monographs, University of Pennsylvania.
- [4] Grandell, J. (1990). *Aspects of Risk Theory*, Chapman & Hall.

Related Articles

Gerber–Shiu Function; Heavy Tails in Insurance; Lévy Processes; Solvency.

SØREN ASMUSSEN

Cramér–Lundberg Estimates

Cramér–Lundberg estimates and bounds are used to estimate ruin probabilities of the surplus in an insurance risk model. The studies of insurance risk models can be dated back to 1909, when F. Lundberg presented his paper “On the theory of risk” at the Sixth International Congress of Actuaries (see [14]). In the paper, a shifted compound Poisson process was used to model the surplus of an insurance company. The model was further investigated by H. Cramér (see [2]). Because of the pioneering work of Lundberg and Cramér on the shifted compound Poisson model, the shifted compound Poisson model is often referred to as the Cramér–Lundberg risk model/process or the classical compound Poisson risk model these days. For a comprehensive coverage of insurance risk models, see **Ruin Theory** and **Insurance Risk Models**.

In the Cramér–Lundberg risk model, the number of claims from an insurance portfolio of an insurance company follows a Poisson process $N(t)$ (see **Poisson Process**) with intensity λ , the successive claim amounts X_n , $n = 1, 2, \dots$, are positive, independent, and identically distributed random variables with common distribution function $P(x)$, $x > 0$, and they are assumed to be independent of the number of claims process $N(t)$. The aggregate claims from the insurance portfolio up to time t are thus given by $S(t) = \sum_{n=1}^{N(t)} X_n$; this is a compound Poisson process with mean $E\{S(t)\} = \lambda t E(X)$ and variance $Var\{S(t)\} = \lambda t E(X^2)$, where X is a generic copy of X_n 's (see **Accumulated Claims**). Suppose that the company has initial surplus $u \geq 0$ and receives premiums continuously at annual rate c . The company's surplus process is written as $U(t) = u + ct - S(t)$. In order to ensure that the premium incomes cover the insurance claims over time, it is generally assumed that $c > \lambda E(X)$. Let $\theta = [c - E\{S(1)\}]/E\{S(1)\}$. Then c may be written as $c = (1 + \theta)\lambda E(X)$. Thus, θ is the risk premium for the expected claim of one monetary unit per unit time and is called the *relative security loading*.

Several quantities associated with the surplus process are of central interest. Among them are the first time that the surplus becomes negative or the company is insolvent $T = \inf\{t; U(t) < 0\}$ and

the probability that the ruin event occurs $\psi(u) = Pr\{T < \infty\}$. Thanks to the independent increments property of the Poisson process, it can be shown that the probability of ruin $\psi(u)$ satisfies the defective renewal equation

$$\psi(u) = \frac{1}{1 + \theta} \int_0^u \psi(u - x) dP_e(x) + \frac{1}{1 + \theta} [1 - P_e(u)] \quad (1)$$

where $P_e(x)$ is the equilibrium distribution of $P(x)$ and is given by

$$P_e(x) = \frac{\int_0^x [1 - P(y)] dy}{E(X)} \quad (2)$$

As the probability of ruin is the solution of the defective renewal equation (1), methods in renewal theory apply. The negative solution $-\kappa$, when it exists, of the following equation

$$\int_0^\infty e^{-zx} dP_e(x) = 1 + \theta \quad (3)$$

or its equivalent form

$$cz + \lambda \tilde{p}(z) - \lambda = 0 \quad (4)$$

where $\tilde{p}(z)$ is the Laplace–Stieltjes transform of $P(x)$, plays a key role in the analysis of time-of-ruin related quantities of the risk model. Equation (3) or (4) is referred to as the Cramér–Lundberg equation or condition, and κ the Lundberg adjustment coefficient. It is straightforward to show that the function $e^{\kappa u} \psi(u)$ satisfies a proper renewal equation and thus the key renewal theorem applies. As a result, one has the Cramér–Lundberg estimate or approximation of ruin probabilities

$$\psi(u) \sim \frac{\theta E(X)}{E\{Xe^{\kappa X}\} - (1 + \theta)E(X)} e^{-\kappa u}, \text{ as } u \rightarrow \infty \quad (5)$$

Moreover, a simple exponential upper bound, called the *Lundberg Bound* or *Lundberg inequality*, is obtainable:

$$\psi(u) \leq e^{-\kappa u}, \quad u \geq 0 \quad (6)$$

The results (5) and (6) are two of the most celebrated results in ruin theory (see **Ruin Theory**).

Expected Discounted Penalty Function

Two other important quantities related to the time of ruin are the surplus immediately before ruin $U(T-)$ and the deficit at ruin $|U(T)|$, where $T-$ is the left limit of T . In the seminal paper [7], Gerber and Shiu introduced the expected discounted penalty function (see **Gerber–Shiu Function**) to provide a unified treatment of the time of ruin T , the surplus $U(T-)$ and the deficit $|U(T)|$. The function is defined as

$$\begin{aligned} \Psi(u) \\ = E \left[e^{-\delta T} w(U(T-), |U(T)|) \mid I(T < \infty) \mid U(0) = u \right] \end{aligned} \quad (7)$$

where $w(x, y)$ is the penalty for a surplus amount of x immediately before ruin and a deficit amount of y at ruin, and δ is interpreted as either a force of interest or a Laplace transform variable. The probability of ruin $\psi(u)$, the joint distribution of $U(T-)$ and $|U(T)|$, and more, are the special cases of the expected discounted penalty function. There have been many papers analyzing the expected discounted penalty function for various insurance risk models. It has been shown that the expected discounted penalty function $\Psi(u)$ satisfies a defective renewal equation. As a result, a Cramér–Lundberg estimate for the expected discounted penalty function $\Psi(u)$ is obtainable.

Distributional Generalizations

It is easy to verify that the solution of the defective renewal equation (1) is the (right) tail of a compound geometric distribution. Hence, the results for the probability of ruin may also apply to compound geometric distributions. If $G(x)$ is the distribution function of a compound geometric distribution with geometric parameter ϕ and the secondary distribution $F_Y(x)$, one has the Cramér–Lundberg estimate

$$1 - G(x) \sim C e^{-\kappa x}, \quad x \rightarrow \infty \quad (8)$$

with

$$C = \frac{1 - \phi}{\kappa \phi E \{ Y e^{\kappa Y} \}} \quad (9)$$

and the Lundberg inequality

$$1 - G(x) \leq e^{-\kappa x}, \quad x \geq 0 \quad (10)$$

where κ is the Lundberg adjustment coefficient when the equilibrium $P_e(x)$ is replaced by $F_Y(x)$. Further generalizations include the refinements and generalizations of the exponential bound to new worse than used (NWU) bounds and distributional generalizations of the Cramér–Lundberg estimate and the Lundberg inequality to other compound distributions. See [4, 10] and [17], and references therein.

Further Extensions

The Cramér–Lundberg risk model has been extended in several ways. Dufresne and Gerber [3] added a Brownian motion to the Cramér–Lundberg risk model:

$$U(t) = u + ct - S(t) + \sigma W(t) \quad (11)$$

where $\{W(t)\}$ is the standard Brownian motion. The additional diffusion term may be interpreted as the future uncertainty of the aggregate claims, the future uncertainty of the premium incomes, or the fluctuation of the investment of the surplus. In this situation, the expected discounted penalty function $\Phi(u)$ satisfies the defective renewal equations. Thus, a Cramér–Lundberg estimate similar to equation (4) exists for $\Phi(u)$ and the result can be found in [15].

The model (11) is a special case of Lévy processes (see **Lévy Processes**). A number of papers discuss the probability of ruin for this class of surplus processes. See [9, 11] and [18]. More recently, the expected discounted penalty function for general Lévy processes has been investigated in [5]. Generally, the results for Lévy surplus processes are very similar to those for the model (11).

Another extension is to replace the Poisson process by a renewal process. The resulting risk model is often called the *Sparre Andersen model* as it was first investigated by Sparre Andersen [1]. In this case, the expected discounted penalty function defined by equation (7) still satisfies a defective renewal equation. The defective renewal equation structure implies that Lundberg estimates and bounds hold in this more general situation. For further discussion of these issues, see **Ruin Theory** [8, 12, 16], and references therein.

Financial Applications

The expected discounted penalty function can be used to price American perpetual options. Assume that the price of a risky asset follows a jump-diffusion process under a risk-neutral probability measure: $A(t) = A(0)\exp\{ct + \sigma W(t) - S(t)\}$, $t \geq 0$ where the components in this model are the same as those in equation (11) except that the choice of c is such that the present value process of $A(t)$ is a martingale (see **Itô's Formula**). Consider an American perpetual put option with time- t payoff $\max\{K - A(t), 0\}$, where K is the strike price. The optimal exercise boundary is constant owing to the stationary and independent increment property of the process. Let φ be the constant and the time of exercise thus is

$$T = \inf\{t : A(t) < \varphi\} \quad (12)$$

If we define the initial surplus $u = \ln[A(0)/\varphi]$ and the penalty function $w(x, y) = \max\{K - \varphi e^{-y}, 0\}$, the exercise time T in equation (12) is the time of ruin for the model (11) and the price of the American perpetual put option is the same as the expected discounted penalty function for the model (11). See [6] for details. Further applications include the pricing of perpetual catastrophe equity put options. A catastrophe equity put option is an option contract that allows an insurer to issue a fixed number of convertible preferred shares at a predetermined price to the option writer, when the accumulated catastrophic loss amount of the insurer or a loss index exceeds a prespecified level. For details, see [13].

References

- [1] Andersen, E.S. (1957). On the collective theory of risk in the case of contagion between claims, *Transactions XVth International Congress of Actuaries* **2**, 219–229.
- [2] Cramér, H. (1930). *On the Mathematical Theory of Risk*, Skandia Jubilee Volume, Stockholm.
- [3] Dufresne, F. & Gerber, H. (1991). Risk theory for the compound Poisson process that is perturbed by diffusion, *Insurance: Mathematics and Economics* **12**, 9–22.
- [4] Embrechts, P., Maejima, M. & Teugels, J. (1985). Asymptotic behaviour of compound distributions, *ASTIN Bulletin* **15**, 45–48.
- [5] Garrido, J. & Morales, M. (2006). On the expected discounted penalty function for the Levy risk processes, *North American Actuarial Journal* **10**(4), 196–216.
- [6] Gerber, H. & Landry, B. (1998). On the discounted penalty at ruin in a jump-diffusion and the perpetual put option, *Insurance: Mathematics and Economics* **22**, 263–276.
- [7] Gerber, H. & Shiu, E.S.W. (1998). On the time value of ruin, *North American Actuarial Journal* **2**(1), 48–72, Discussions, 72–78.
- [8] Gerber, H.U. & Shiu, E.S.W. (2005). The time value of ruin in a Sparre Andersen model, *North American Actuarial Journal* **9**(2), 49–84.
- [9] Huzak, M., Perman, M., Sikic, H. & Vondracek, Z. (2004). Ruin probabilities and decompositions for general perturbed risk processes, *Annals of Applied Probability* **14**, 1378–1397.
- [10] Kalashnikov, V. (1997). *Geometric Sums: Bounds for Rare Events with Applications*, Kluwer Academic Publishers, Dordrecht.
- [11] Klüppelberg, C., Kyprianou, A.E. & Maller, R.A. (2004). Ruin probabilities and overshoots for general Levy insurance risk processes, *Annals of Applied Probability* **14**, 1766–1801.
- [12] Li, S. & Garrido, J. (2004). On ruin for the Erlang(n) risk process, *Insurance: Mathematics and Economics* **34**, 391–408.
- [13] Lin, X.S. & Wang, T. (2009). Pricing perpetual American catastrophe put options: a penalty function approach, *Insurance: Mathematics and Economics*, **44**, 287–295.
- [14] Lundberg, F. (1909). Über die theorie der rückversicherung, *Transactions of VI International Congress of Actuaries* **1**, 877–948.
- [15] Tsai, C.C.-L. & Willmot, G.E. (2002). A generalized defective renewal equation for the surplus process perturbed by diffusion, *Insurance: Mathematics and Economics* **30**, 51–66.
- [16] Willmot, G.E. (2007). On the discounted penalty function in the renewal risk model with general interclaim times, *Insurance: Mathematics and Economics* **41**, 17–31.
- [17] Willmot, G.E. & Lin, X.S. (2001). *Lundberg Approximations for Compound Distributions with Insurance Applications*, Lecture Notes in Statistics 156, Springer-Verlag, New York.
- [18] Yang, H.L. & Zhang, L. (2001). Spectrally negative Lévy processes with applications in risk theory, *Advances in Applied Probability* **33**, 281–291.

Related Articles

American Options; Cramér's Theorem; Insurance Derivatives; Insurance Risk Models; Lévy Processes; Large Deviations; Ruin Theory.

X. SHELDON LIN

Actuarial Premium Principles

An actuarial premium principle is a method for assigning an appropriate price for an insurance policy. This price is the amount that has to be paid by the insured for exchanging “risk” with “certainty.” The agreed premium represents an equilibrium between the parties involved. In fact, individual agents look at the price they are willing to pay for hedging (partially or totally) their risks, whereas insurers take into account how much premium they should charge in order to avoid technical ruin and to guarantee the rentability of the underlying insurance portfolio. Premium principles deal with the analysis and the characterization of such equilibrium.

In the first section, we introduce basic notions, an axiomatic characterization, and a review of premium principles. Next, an economic method to interpret the relation between premium principles and market systems is described. Then, we present a unified approach based on a generalized Markov inequality for tail probabilities and show how several types of premium principles can be derived from it.

In the last section, we consider the analysis of insurance portfolios (i.e., collections of insurance contracts) traded in financial markets. In this case, portfolios are transferred either from an insurer to another insurer or from an insurer to the market (*see Securitization*). We devote the last paragraph to the analysis of pricing such portfolios.

In this article, we focus on the main aspects of the theory of premium principles. Topics that are not covered include insurer profitability and solvency requirements (*see Solvency*).

Notice that in the recent literature, premium principles are often considered as akin to risk measures (*see Convex Risk Measures; Value-at-Risk; Expected Shortfall; Spectral Measures of Risk*).

The Axiomatic Characterization of Premium Principles

Classes of premium principles may be derived from sets of axioms. The introduction of such axioms is not an arbitrary or dogmatic approach but rather an essential step for defining the properties a premium

principle has to satisfy. Indeed, in case a set of axioms proposed does not suit a certain intuition of insurance risk, we could modify it properly. In this way, it is possible to justify an axiomatic characterization by showing its strengths or to criticize it by demonstrating its lack of practical importance.

The analysis of premium principles using an axiomatic characterization was introduced in the actuarial context in [20].

Let us consider a probability space (Ω, F) , where Ω is the set of the scenarios equipped with a σ -algebra F . A risk is a random variable defined on (Ω, F) , that is, $X : \Omega \rightarrow \Re$ is a risk if $X^{-1}((-\infty, x]) \in F$ for all $x \in \Re$. When X takes only positive values, that is, $X \geq 0$, we call it a *risk*. The class of all random variables on (Ω, F) is denoted by $\mathbf{W}$. We equip the measurable space with a probability measure P , which we call the *base probability measure*.

A premium principle π is a functional that assigns a real number to any random variable X in $\mathbf{W}$.

The above definition may be extended to the case where risks are noninsurable, that is, when $\pi[X] = +\infty$.

In the following, we introduce some axioms and properties a premium principles may (or may not) satisfy. All random variables are assumed to be elements of the set $\mathbf{W}$ introduced above.

Monotonicity: π is monotonic if $\pi[Y] \leq \pi[X]$ when $Y(\omega) \leq X(\omega)$ for all $\omega \in \Omega$. π is P -monotonic if $\pi[Y] \leq \pi[X]$ when $Y \leq X$, P -almost sure.

Positive homogeneity (scale invariance, scale equivariance): π is positively homogeneous if $\pi[aX] = a\pi[X]$ for all $a \geq 0$.

Translation invariance (translation equivariance, translativity, consistency): π is translation invariant if $\pi[X + c] = \pi[X] + c$ for all real c .

Subadditivity: π is subadditive if $\pi[X + Y] \leq \pi[X] + \pi[Y]$. If the inequality is reversed, then the property is called *superadditivity*.

Additivity for independent risks: π is additive if $\pi[X + Y] = \pi[X] + \pi[Y]$ for independent $X, Y \in \mathbf{W}$.

Comonotonic additivity: π is comonotonic additive (*see Copulas in Insurance*) if $\pi[X + Y] = \pi[X] + \pi[Y]$ whenever X and Y are comonotonic.

See [17, 23] for a review of independent and comonotonic additive premium principles and [10, 11] for an introduction to comonotonicity.

For taking in account violations of the axioms of subadditivity and positive homogeneity due to liquidity risk (see **Liquidity Premium**), it is possible to replace these two with the requirement by convexity. The interested reader is referred to [1, 16, 22] and **Convex Risk Measures** and also to [12].

Convexity: π is convex if $\pi[aX + (1 - a)Y] \leq a\pi[X] + (1 - a)\pi[Y]$ for all $a \in (0, 1)$.

Law-invariance: π is law invariant if $\pi[X] = \pi[Y]$ in case $X \stackrel{d}{=} Y$.

Preserving stop-loss order: π preserves stop-loss order if $\pi[X] \leq \pi[Y]$ when $E[(X - d)_+] \leq E[(Y - d)_+]$ for all $d \in \mathbb{R}$.

Preserving first-order stochastic dominance (FOSD): π preserves first-order stochastic dominance if $\pi[X] \leq \pi[Y]$ when $P[X \leq x] \geq P[Y \leq x]$ for all $x \in \mathbb{R}$.

These two properties are closely linked to each other. Indeed, the latter is a direct consequence of the former. For a review on the applications of stop-loss and stochastic ordering in actuarial science, see [9].

Positive risk loading: π induces a risk loading if $\pi[X] \geq E[X]$.

Having a positive risk loading is a suitable property because without it an insurer will make a loss in the long run. We will see the relevance of this property comparing the net and the expected value principle.

No rip-off (maximal loss): $\pi[X] \leq \text{ess sup}[X]$.

Not unjustified: π is not unjustified if $\pi[X] = c$, when $X \in \mathbf{W}$ is equal to $c \geq 0$ almost everywhere.

Premium Principles: A Review

In this section, we introduce several premium principles as well as their relevant properties. We also give the definition of the Dutch premium calculation principle.

Note that all the following premium principles implicitly assume the existence of a given utility function (see **Utility Theory: Historical Perspectives; Utility Function**). For an analysis of expected utility theory and subjective expected utility theory, the interested reader is referred to [29, 35]. For more applications of the theory of choice under uncertainty and the utility framework within actuarial science, we refer to [7].

Expected value principle: The expected value principle is given as

$$\pi[X] = (1 + \alpha)E[X] \quad (1)$$

where $\alpha \in \mathbb{R}$. In case $\alpha = 0$ we find the net premium that does not load for risk. This principle is justified by the strong law of large numbers (SLLN). Given that the risk $X = \sum_{i=1}^n X_i$ is linked to a large homogeneous portfolio, let us suppose that X_i are independent and identically distributed (i.i.d.) and their first moment exists and is finite, $E[X_i] = \mu < \infty$, for $i = 1, 2, \dots$. Thus, from the SLLN, we find that

$$\lim_{n \rightarrow \infty} P[|n^{-1}X - \mu| < \varepsilon] = 1 \quad (2)$$

for all $\varepsilon > 0$.

The net premium principle is appropriate only in case insurers are risk-neutral. Indeed from the ruin theory (see **Ruin Theory**), it is known that ruin will inevitably occur whenever there is no loading.

Equivalent utility principle: The equivalent utility premium is obtained for the premium that makes the expected utility of the income $\pi[X] - X$ of the insurer equal to the utility of not accepting the risk. We could assume that the utility function is subjective and the probability measure objective, if we choose the framework introduced in [35]. Otherwise, following the results in [29], the probability measure is also considered to be subjective and not known.

The minimal premium, $\pi^-[X]$, and the maximal premium, $\pi^+[X]$, are derived by solving

$$E[u(w + \pi[X] - X)] = u(w) \quad (3)$$

and

$$E[\bar{u}(\bar{w} - X)] = \bar{u}(\bar{w} - \pi[X]) \quad (4)$$

respectively, where X is a positive random variable, w a real number, and $\bar{u}$ denotes the utility function of the insured and u the one of the insurer.

Thus, the insurer will sell the insurance contract if and only if $\pi[X] \geq \pi^-[X]$, whereas the insured will buy it when $\pi[X] \leq \pi^+[X]$. Hence, the equilibrium price lies in the interval

$$[\pi^-[X], \pi^+[X]]$$

In case both the insurer and the insured are risk-averse (see **Risk Aversion**), it follows that

$$E[X] \leq \pi^-[X] \leq \pi^+[X] \quad (5)$$

Different specifications of the utility function yield different expressions $\pi^-[X]$ and $\pi^+[X]$. It is not guaranteed that an explicit expression for $\pi^-[X]$ or $\pi^+[X]$ is obtained.

As a final remark, the difference $\pi^+[X] - \pi^-[X]$ denotes the mark-up price. It is maximum in the case of insurer's monopoly, whereas in a competitive insurance market it tends to zero.

Distortion principle: The distortion principle as a premium principle was introduced in [36, 37] and it can be seen as an attempt to overcome some of the shortcomings of the expected utility theory. Indeed as a consequence of the Allais paradox [3], some nonexpected utility theories were proposed. The distortion principle refers to Yaari's dual theory [38]. Yaari showed that the amount of money $\pi(v)$ (certainty equivalent) that makes it indifferent to hold the lottery V or to receive the amount $\pi(v)$ with certainty is given as

$$\pi(v) = -\int_{-\infty}^0 (1 - g(\bar{F}_V(v)))dv + \int_0^{+\infty} g(\bar{F}_V(v))dv \quad (6)$$

where $\bar{F}_V(v) = 1 - P[V \leq v]$. The function $g : [0, 1] \rightarrow [0, 1]$ is known as *distortion function* and it is nondecreasing and it fulfills $g(0) = 0$ and $g(1) = 1$.

Using the distorted distribution function $F_g^*(x) = 1 - g(\bar{F}_V(x))$ together with equation (6), we get that $\pi(v) = \int_{-\infty}^{+\infty} x dF_g^*(x)$.

For an overview on distortion risk measures, the reader is referred to [9, 13].

Proportional hazard (PH) principle: The PH principle arises as a special case of the distortion principle when $\bar{g}(x) = x^{1/\alpha}$, $\alpha \geq 1$, and it is given as

$$\begin{aligned} \pi[X] = & -\int_{-\infty}^0 (1 - (\bar{F}_X(x))^{1/\alpha}) dx \\ & + \int_0^{+\infty} ((\bar{F}_X(x))^{1/\alpha}) dx \end{aligned} \quad (7)$$

See [36] for a review of the properties of this premium principle.

Mean value principle: The mean value principle arises as the root π of the following equation:

$$f(\pi) = E[f(x)] \quad (8)$$

where f is a nondecreasing and nonnegative function defined on $\mathfrak{R}$.

Variance principle: For some $\alpha > 0$, the variance principle is given as

$$f(\pi) = E[X] + \alpha \text{Var}[X] \quad (9)$$

Standard deviation principle: For some $\alpha > 0$, the standard deviation principle is given as

$$f(\pi) = E[X] + \alpha \sqrt{\text{Var}[X]} \quad (10)$$

The last two principles are built on the expected value by including a risk loading that is proportional to the variance or the standard deviation of the risk. Denneberg [8] suggested to replace the standard deviation with the absolute deviation. See [27, 30] for a generalization of the variance and standard deviation principles to the dynamic case.

Exponential principle: The exponential principle is given as

$$\pi[X] = \frac{1}{\alpha} \log E[\exp(\alpha X)] \quad (11)$$

where $\alpha > 0$. The exponential principle represents a special case of the equivalent utility principle in the case of an exponential utility function; see [17, 23, 26].

Esscher principle: The Esscher premium is given as

$$\pi[X] = \frac{E[Xe^{\alpha X}]}{E[e^{\alpha X}]} \quad (12)$$

where $\alpha > 0$. This premium is based on the notion of Esscher transform (*see **Esscher Transform***), introduced by Esscher [15]. Let us consider a random variable X and a positive real number α such that $E[e^{\alpha X}] < \infty$; one considers a transformed distribution function:

$$F_{\tilde{X}}(x) = \int_0^x \frac{e^{\alpha X}}{E[e^{\alpha X}]} dF_X(x) \quad (13)$$

In this way, we obtain the Esscher measure that is mutually absolute continuous, that is, equivalent (*see **Equivalence of Probability Measures***), with respect to the original probability measure. For a review of the characterizations of Esscher principle, we refer to [32]. In the last paragraph, we demonstrate the importance of Esscher transform in an insurance pricing framework.

Swiss principle: The Swiss premium π is the root of the following equation in π :

$$E[w(W - p\pi)] = w((1 - p)\pi) \quad (14)$$

4 Actuarial Premium Principles

Table 1 All premiums mentioned above are law invariant. All principles are not unjustified except the expected value principle for $\alpha > 0$

Principle	M	PH	TI	SAd	Ad	IAd	CAd	C	SL	NR	RL
Expected Value	Y	Y	N	Y	Y	Y	Y	Y	Y	N	Y
Equivalent utility	Y	N	Y	N	N	N	N	N	Y	Y	Y
Distortion	Y	Y	Y	Y	N	N	Y	Y	Y	Y	Y
Proportional Hazard	Y	Y	Y	Y	N	N	Y	Y	Y	Y	Y
Mean value	Y	N	Y	N	N	N	N	N	Y	Y	Y
Variance	N	N	Y	N	N	Y	N	N	N	N	Y
Standard Deviation	N	Y	Y	Y	N	N	N	Y	N	N	Y
Exponential	Y	N	Y	N	N	Y	N	N	Y	Y	Y
Esscher	N	N	Y	N	N	Y	N	N	N	Y	Y
Swiss	Y	N	N	N	N	N	N	N	Y	Y	Y
Dutch	Y	Y	N	Y	N	N	N	Y	Y	Y	Y
Orlicz	Y	Y	N	Y	N	N	N	Y	Y	Y	Y

M, monotonicity; PH, positive homogeneity; TI, translation invariance; SAd, subadditivity; Ad, additivity; IAd, independent additivity; CAd, comonotonic additivity; C, convexity; SL, stop-loss order; NR, no rip-off; RL, risk loading

for a nonnegative and nondecreasing function w defined on $\mathfrak{R}$ and a given parameter $0 < p \leq 1$.

Both the equivalent utility principle and the mean value principle can be obtained as particular cases of the Swiss premium principle; see [20].

Dutch principle: The Dutch principle is given as

$$\pi[X] = E[X] + \theta E[(X - \alpha E[X])_+],$$

$$\alpha \geq 1, \quad 0 < \theta \leq 1 \quad (15)$$

The Dutch premium principle was introduced in [31].

Orlicz principle: The Orlicz principle is given by the root of

$$E[\psi(X/\pi)] = 1 \quad (16)$$

where $X \geq 0$ and ψ is a normalized Young function on $\mathfrak{R}_+ \cup \{0\}$. This premium reflects a change in risk in a proportional change in premium (Table 1).

The Economic Method

In 1980, Bühlmann derived the economic premium principle in contrast to classical premium principles that depend only on the risk to be covered; see [4]. He took into account general market conditions and proposed a premium principle based on economic arguments. In this way, the economic premium principle is a function that associates with a random variable X representing insurance risk and a random variable Z denoting market risk π , that is, $\pi : (X, Z) \rightarrow \mathfrak{R}$.

Let consider a market with n insurer agents. We suppose that the insurer i bears risk X_i , $i = 1, 2, \dots, n$, and that his/her risk aversion is characterized by the exponential utility function $u_i(x) = \frac{1}{\alpha_i}(1 - e^{(-\alpha_i x)})$. Bühlmann obtained the following expression for the equilibrium pricing density for the risk X :

$$\pi[X, Z] = \frac{E[Xe^{\alpha Z}]}{E[e^{\alpha Z}]} \quad (17)$$

that is, the Esscher premium; see [4]. Moreover, it can be considered as a Pareto optimal risk exchange. In the case where X and $Z - X$ are independent to each other, we derive the standard premium principle:

$$\begin{aligned} \pi[X, Z] &= \frac{E[Xe^{\alpha X}]E[Xe^{\alpha(Z-X)}]}{E[e^{\alpha X}]E[e^{\alpha(Z-X)}]} \\ &= \frac{E[Xe^{\alpha Z}]}{E[e^{\alpha Z}]} = \pi[X] \end{aligned} \quad (18)$$

Bühlmann derived the above result using the concept of a pricing density (pricing kernel and state price deflator). The interested reader is referred to [28].

Bühlmann generalized this result in [5], allowing for insurers with general utility functions. For an extension of the above model to the dynamic multi-period case, see [25].

Table 2 Premium principles derived from the generalized Markov inequality

Premium principle	$\nu(x)$	$\phi(x, \pi)$
Mean value	1	$\frac{f(x)}{f(\pi)}$
Equivalent utility	1	$\frac{u(\pi - x)}{u(0)}$
Swiss	1	$\frac{\omega(x - p\pi)}{\omega((1 - p)\pi)}$
Orlicz	1	$\psi(x/\pi)$

A Unified Approach by Markov Inequality for Tail Probabilities and VaR

A unified approach for generating premium principles based on the Markov inequality of tail probabilities has been introduced in [21]. Indeed, using two exogenous functions $\nu(\cdot)$ and $\phi(\cdot, \cdot)$ (Table 2), an exogenous parameter $\alpha \leq 1$, and assuming that $\nu(\cdot)$ is both nonnegative and nondecreasing and fulfills $\phi(x, \pi) \geq 1_{\{x > \pi\}}$, where $1_{\{x > \pi\}} = \begin{cases} 1 & \text{if } x > \pi \\ 0 & \text{if } x < \pi \end{cases}$, then by the generalized Markov inequality we find that

$$P[X > \pi] \leq \frac{E[\phi(X, \pi)\nu(X)]}{E[\nu(X)]} \quad (19)$$

Focusing on the upper bound of the above equation, if it is set to the confidence level $\alpha \leq 1$, we have that

$$P[X > \pi] \leq \frac{E[\phi(X, \pi)\nu(X)]}{E[\nu(X)]} = \alpha \leq 1 \quad (20)$$

When $\alpha = 1$, equation (21) can be used for generating several well-known premium principles:

$$\frac{E[\phi(X, \pi)\nu(X)]}{E[\nu(X)]} = 1 \quad (21)$$

It should be remarked that this unified approach is also directly linked to the quantile-based approach. Indeed, for all ϕ, ν , using the results shown above, we find that $\pi \geq \inf\{x \in \mathfrak{R} : P[X \leq x] \geq 1 - \alpha\}$; see [20].

Insurance Pricing

This section is devoted to the analysis of insurance portfolios (i.e., collections of insurance contracts) traded in financial markets.

Although traditional actuarial pricing of insurance relies on economic theory of decision under uncertainty, financial pricing of contingent claims mainly relies on risk-neutral valuation, which can be justified within a no-arbitrage setting (*see Arbitrage: Historical Perspectives; Arbitrage Pricing Theory*). Even if this assumption may seem unrealistic in the case of traditional insurance markets, it is suitable for insurance products traded on financial markets; see [2, 7, 33, 34]. Moreover, owing to the incompleteness of insurance markets, we will observe an infinite number of equivalent martingale measures (*see Equivalence of Probability Measures; Martingales*) and hence an uncountable set of possible prices.

Insurance prices can be derived by means of the Esscher transform, as suggested by Gerber and Shiue [18, 19]. These authors defined the Esscher transform for exponential Lévy processes (*see Lévy Processes*), that is, processes with independent and stationary increments. Under the Esscher transform probability measure, it is possible to establish a change of measure for a random variable. Choosing suitable Esscher parameters, discounted prices become martingales and thus no-arbitrage prices can be calculated. In the case where equivalent martingale measure is unique, it is derived by the Esscher transform. Otherwise, if it is not unique, as is usual in the case of insurance markets, the equivalent martingale measure can be derived by the Esscher transform assuming the existence of a representative agent that maximizes its expected utility to a power utility function.

The interested reader is referred to [6, 14, 24] for a review of recent developments in this topic.

Acknowledgments

The authors acknowledge the financial support of the Onderzoeksfond K.U. Leuven (GOA 2007–2011: Risk modeling and valuation of insurance and financial cash flows, with applications to pricing, provisioning, and solvency). Jan Dhaene and Marc Goovaerts also thank the financial support of Fortis (K.U. Leuven Fortis Chair in Financial and Actuarial Risk Management).

References

- [1] Acerbi, C. & Scandolo, G. (2008). Liquidity and coherent risk measures, *Quantitative Finance* forthcoming.
- [2] Albrecht, P. (1992). Premium calculation without arbitrage? A note on a contribution by G. Venter, *Astin Bulletin* **22**, 247–254.

- [3] Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine, *Econometrica* **21**, 503–546.
- [4] Bühlmann, H. (1980). An economic premium principle, *Astin Bulletin* **14**, 13–22.
- [5] Bühlmann, H. (1984). The general economic premium principle, *Astin Bulletin* **14**, 89–101.
- [6] Bühlmann, H., Delbaen, F., Embrechts, P. & Shirayev, A.N. (1996). No arbitrage, change of measure and conditional Esscher transforms, *CWI Quarterly* **9**, 291–317.
- [7] Delbaen, F. & Haezendonck, J. (1989). A martingale approach to premium calculation principles in an arbitrage free market, *Insurance: Mathematics and Economics* **8**, 269–277.
- [8] Dennenberg, D. (1990). Premium calculation: why standard deviation should be replaced by absolute deviation, *Astin Bulletin* **20**, 181–190.
- [9] Denuit, M., Dhaene, J., Goovaerts, M.J. & Kaas, R. (2005). *Actuarial Theory for Dependent Risks*, Wiley, New York.
- [10] Dhaene, J., Denuit, M., Goovaerts, M.J., Kaas, R. & Vyncke, D. (2002). The concept of comonotonicity in actuarial science and finance: theory, *Insurance: Mathematics and Economics* **31**, 3–33.
- [11] Dhaene, J., Denuit, M., Goovaerts, M.J., Kaas, R. & Vyncke, D. (2002). The concept of comonotonicity in actuarial science and finance: applications, *Insurance: Mathematics and Economics* **31**, 133–161.
- [12] Dhaene, J., Laeven, R.J.A., Vanduffel, S., Darkiewicz, G. & Goovaerts, M.J. (2008). Can a coherent risk measure be too subadditive? *Journal of Risk and Insurance* **75**, 365–386.
- [13] Dhaene, J., Vanduffel, S., Tang, Q., Goovaerts, M.J., Kaas, R. & Vyncke, D. (2006). Risk measures and comonotonicity: a review, *Stochastic Models* **22**, 573–606.
- [14] Embrechts, P. (2000). Actuarial versus financial pricing of insurance, *Risk Finance* **1**, 17–26.
- [15] Esscher, F. (1932). On the probability function in the collective theory of risk, *Scandinavian Actuarial Journal* **15**, 175–195.
- [16] Follmer, H. & Schied, A. (2002). Convex measures of risk and trading constraints, *Finance and Stochastics* **6**, 429–447.
- [17] Gerber, H.U. (1974). On additive premium calculations principles, *Astin Bulletin* **7**, 215–222.
- [18] Gerber, H.U. & Shiu, S.W.E. (1994). Option pricing by Esscher transform, *Transactions of the Societies of Actuaries* **46**, 99–191.
- [19] Gerber, H.U. & Shiu, S.W.E. (1996). Actuarial bridges to dynamic hedging and option pricing, *Insurance: Mathematics and Economics* **18**, 183–218.
- [20] Goovaerts, M.J., De Vylder, F.E.C. & Haezendonck, J. (1984). *Insurance Premiums*, North Holland Publishing, Amsterdam.
- [21] Goovaerts, M.J., Kaas, R., Dhaene, J. & Tang, Q. (2003). A unified approach to generate risk measures, *Astin Bulletin* **33**, 173–191.
- [22] Goovaerts, M.J., Kaas, R., Dhaene, J. & Tang, Q. (2004). Some new classes of consistent risk measures, *Insurance: Mathematics and Economics* **34**, 505–516.
- [23] Goovaerts, M.J., Kaas, R., Dhaene, J., Laeven, R.J.A. & Tang, Q. (2004). A comonotonic image of independence for additive risk measures, *Insurance: Mathematics and Economics* **35**, 581–594.
- [24] Goovaerts, M.J. & Laeven, R.J.A. (2008). Actuarial risk measures for financial derivative pricing, *Insurance: Mathematics and Economics* **42**, 540–547.
- [25] Iwaki, H., Kijima, M. & Morimoto, Y. (2001). An economic premium principle in a multiperiod economy, *Insurance: Mathematics and Economics* **28**, 325–339.
- [26] Kaas, R., Goovaerts, M.J., Dhaene, J. & Denuit, M. (2001). *Modern Actuarial Risk Theory*, Kluwer Academic Publishers, Amsterdam.
- [27] Møller, T. (2001). On transformations of actuarial valuation principles, *Insurance: Mathematics and Economics* **28**, 281–303.
- [28] Rubinstein, M. (1976). The valuation of uncertain income and the pricing of options, *The RAND Corporation* **7**, 407–425.
- [29] Savage, L.J. (1954). *The Foundations of Statistics*, Wiley, New York.
- [30] Schweizer, M. (2001). From actuarial to financial valuation principles, *Insurance: Mathematics and Economics* **28**, 31–47.
- [31] Van Heerwaarden, A.E. & Kaas, R. (1992). The Dutch premium principle, *Insurance: Mathematics and Economics* **11**, 129–133.
- [32] Van Heerwaarden, A.E., Kaas, R. & Goovaerts, M.J. (1989). Properties of the Esscher premium calculation principle, *Insurance: Mathematics and Economics* **8**, 261–267.
- [33] Venter, G.G. (1991). Premium calculation implications without arbitrage, *Astin Bulletin* **21**, 223–230.
- [34] Venter, G.G. (1992). Premium calculation implications without arbitrage Authors reply on the note by P. Albrecht, *Astin Bulletin* **22**, 255–256.
- [35] Von Neumann, J. & Morgenstern, O. (1947). *Theory of Games and Economic Behavior*. Princeton university Press, Princeton.
- [36] Wang, S.S. (1996). Premium calculation by transforming the layer premium density, *Astin Bulletin* **26**, 71–92.
- [37] Wang, S.S., Young, V.R. & Panjer, H.H. (1997). Axiomatic characterization of insurance prices, *Insurance: Mathematics and Economics* **21**, 173–183.
- [38] Yaari, M.E. (1987). The dual theory of choice under risk, *Econometrica* **55**, 95–115.

MARC J. GOOVAERTS, JAN DHAENE & OMAR
RACHEDI

Reinsurance

In a *reinsurance contract*, one party (the *reinsurer*) agrees to indemnify for a certain premium another party (called the *reinsured*, the *first-line insurer* or also the *ceding company*) for specified parts of its underwritten insurance risk. Hence, reinsurance is a form of *risk sharing* between insurance companies (cf. [3]). One among the main motivations for an insurance company to buy reinsurance is to reduce the probability of suffering losses that it cannot easily cope with, in particular, the appearance of excessively large or unusually many claims from the underwritten policies. Also, reinsurance is a way to increase the underwriting capacity (which is particularly but not exclusively relevant for smaller insurance companies) and can be interpreted as a (virtual) increase in the solvency (see **Solvency**) capital of the ceding company. Replacing a part of the random costs (claims) by fixed costs (premiums) is the core insurance activity and helps the policy holders to stabilize their business. But this also applies to the first-line insurer as a reinsurance customer and reinsurance can increase the effectiveness of the marketplace. For the same reason, a reinsurance company may itself, again for some suitable premium, pass on parts of the reinsured risks to another reinsurer (both domestic and international), which is then called *retrocession*. For the general concept of the *need for reinsurance* see, for example, Gerber [16] and Schnieper [25].

The type and amount of reinsurance coverage may be negotiated for individual risks in the portfolio (*facultative reinsurance*). But predominantly *reinsurance treaties* are signed, which are obligatory agreements that specify the type and amount of reinsurance coverage for an entire portfolio of risks within an insurance branch (usually for a time horizon of one year).

Among the crucial questions in the context of reinsurance are:

- Which type of reinsurance form is appropriate in a given situation?
- How much reinsurance should be purchased?
- What should be the premium?

Forms of Reinsurance

Let $X(t) := \sum_{i=1}^{N(t)} X_i$ denote the aggregate claim amount in an insurance portfolio up to time t , where

the counting process N_t specifies the number of claims up to time t and the random variables X_i denote the individual claim sizes. In every reinsurance treaty, this amount is split into $X(t) = D(t) + R(t)$, where $D(t)$ is the amount that the insurance company retains for itself (the *deductible*) and $R(t)$ is the *reinsured* amount to be covered by the reinsurance company.

There are two types of proportional reinsurance forms.

Quota-Share Reinsurance

A very common and simple reinsurance form is the quota-share (QS) treaty, where one has

$$R(t) := a X(t), \quad D(t) := (1 - a) X(t) \quad (1)$$

for some proportionality factor $0 < a < 1$. In this case, it is straightforward to calculate the distribution of $R(t)$ and $D(t)$, once the distribution of $X(t)$ is known.

Surplus Reinsurance

While in a QS treaty also small claims are reinsured (which the first-line insurer often wants to keep for himself), a surplus treaty only reinsures claims X_i whose corresponding insured sum Q_i exceeds some retention M . In the latter case, the first-line insurer deducts a proportion M/Q_i and shifts the remaining part to the reinsurer. Consequently,

$$R(t) = \sum_{i=1}^{N(t)} \left(1 - \frac{M}{Q_i}\right) X_i 1_{\{Q_i \geq M\}},$$

$$D(t) = \sum_{i=1}^{N(t)} \left(X_i I(Q_i \leq M) + M \frac{X_i}{Q_i} 1_{\{Q_i > M\}}\right) \quad (2)$$

Surplus treaties are quite popular in practice, in particular in fire, marine, and storm insurance.

Other types of reinsurance forms are of nonproportional nature.

Excess-of-Loss Reinsurance

In an excess-of-loss (XL) treaty, for each individual claim the excess over some retention M is paid by

the reinsurer:

$$R(t) := \sum_{i=1}^{N(t)} (X_i - M)^+, \quad D(t) := \sum_{i=1}^{N(t)} \min(X_i, M) \quad (3)$$

XL covers are, for instance, popular in casualty and windstorm insurance. In most situations, there is an upper limit L of the XL cover of each risk as well, resulting in the treaty

$$R(t) := \sum_{i=1}^{N(t)} \min\{(X_i - M)^+, L\} \quad (4)$$

where L is then the size of the so-called *XL reinsurance layer* (the usual notation is then L *xs* M). It is also common to have an upper limit $k \cdot L$ on the total reinsured amount $R(t)$ for some integer k . In the latter case, one is said to have k *reinstatements* and, depending on the contract, the premium for a higher reinstatement often only has to be paid once the lower reinstatement is used up.

A variant of XL reinsurance is the *per-event XL cover* (or sometimes called *catastrophe XL cover*), where all claims due to some external event such as a storm or an earthquake are first added and then treated as a single claim in an XL contract (avoiding the issue that for many small claims the XL treaty may not provide sufficient protection).

Stop-Loss Reinsurance

The stop-loss (SL) treaty acts on the aggregate claim amount $X(t)$ of the portfolio:

$$R(t) := \left\{ \sum_{i=1}^{N(t)} X_i - C \right\}^+, \quad D(t) := \min(X(t), C) \quad (5)$$

where C is some retention. In general, there is again an additional upper limit on $R(t)$.

In practice, often combinations of the above reinsurance forms are employed (together with upper limits on the total reinsurance cover).

Large-Claim Reinsurance

Let $\{X_1^*, X_2^*, \dots, X_{N(t)}^*\}$ denote the order statistics of the claims ordered according to their size. Then the

largest claims reinsurance is defined through

$$R(t) := \sum_{i=1}^r X_{N(t)-i+1}^*, \quad D(t) := \sum_{i=1}^{N(t)-r} X_i^* \quad (6)$$

that is, the reinsurer covers the r largest claims of the portfolio that have occurred up to time t (which again usually is 1 year). Another variant called *excedent du cout moyen relatif* (ECOMOR) was introduced by Thépaut [27] with a reinsured amount of the form

$$R(t) := \sum_{i=1}^{N(t)} \{X_i - X_{N(t)-r}^*\}^+ \quad (7)$$

so that in this case the reinsurer covers only that part of the r largest claims that overshoots the random retention $X_{N(t)-r}^*$. Such a contract gives the reinsurer protection against claim inflation, since if the claim amounts increase, the retention will also increase.

Although theoretically appealing, large-claim reinsurance treaties are rarely applied in practice.

Reinsurance Premiums

In the first place, reinsurance is a specific form of insurance. Consequently, the principles of premium calculation (*see Actuarial Premium Principles*) can, to some extent, also be applied in the reinsurance setting, once the moments or even the distribution of the number and the size of the reinsured claims are available (although costs and loading factors will in general be different). Extreme value statistics is of vital importance in this context, since typically the reinsured claims are heavy tailed and their distribution has to be estimated, often on the basis of only a few data points (*see Heavy Tails in Insurance* and [14]).

Whereas for proportional reinsurance the actuarial reinsurance premium will simply be the corresponding proportion of the premium that the first-line insurer has received for the underlying risks (minus some acquisition and administration costs), the situation for nonproportional reinsurance is more involved.

The claim number process for the reinsured claims in an XL treaty can be interpreted as an independent thinning of the point process (*see Point Processes*) that generates the claims for the first-line insurer. For many practically relevant examples, the resulting

claim number process for the reinsurer is then of the same form with just some parameters modified (see, e.g., [4]). For the premium calculation of an XL treaty with retention M , under the assumption of independence between the number and the sizes of the claims as well as identically distributed claim sizes, often the *reduction effect*

$$r(M) := \frac{E(D(t))}{E(X(t))} = \frac{1}{E(X)} \int_0^M (1 - F(w)) dw \quad (8)$$

is used, which turns out to be the *equilibrium distribution function* of the distribution function F of an individual claim X . In practice, one distinguishes between *exposure rating*, where the reinsurer uses the portfolio and premium information of the first-line insurer, and *experience rating*, which is mainly based on the previous claim experience of the XL layer itself. Especially, the latter method often faces the challenge of having only a few data points available, in which case credibility estimates (see **Credibility Theory**) may be used. Another important aspect in XL treaties is potential claims inflation, which may be substantial (as the time until settlement of a claim may be large) and would often just affect the reinsurer. Hence, typically some *index clause* is arranged, which splits the amount of claims inflation between the first-line insurer and the reinsurer.

For SL treaties, the aggregate claim distribution (see **Accumulated Claims**) of the first-line insurance portfolio will be the crucial quantity to determine suitable reinsurance premiums, using, for instance, generalized stop-loss transforms to determine moments of the reinsured claim amount (for numerical and algorithmic issues, see [20]).

Premium calculations for large-claim reinsurance contracts are more involved, but several asymptotic results using extreme value theory are available (see [4] for a survey). Note that for all these treaties it is also important to adequately assess the possible dependence among the risks, as this may drastically influence the appropriate premium (see, for instance, [12]).

In addition to the resulting actuarial premium, the actual reinsurance premium will also be influenced by other factors like competition, volume of the portfolio, and perhaps a loading that accounts for potential *moral hazard* and *adverse selection*. Here moral hazard refers to the tendency of the reinsured

not to be as strict in the claim settlement process as without reinsurance (see, e.g., [13]), whereas adverse selection accounts for the asymmetry of information of the cedant and the reinsurer on the underlying risk in general (see [19]). Also, for contracts like the XL treaty, the reinsurer typically has no insight into the number and sizes of the claims below the retention and needs to get information from other sources, which also increases the costs (see, e.g., [17, 18]). General considerations on planning in reinsurance can be found in [26].

Robustness questions of reinsurance premiums are discussed in Brazauskas [8]. In addition, the economic environment influences the pricing of reinsurance risks (see [1, 2]).

Choice of Reinsurance

Each reinsurance form has its particular advantages and disadvantages in terms of the type of protection it provides (frequency risk, large-claim risk), premium calculation, practical handling, administration and processing of loss estimation (including issues like moral hazard and adverse selection). An insurer may have a concrete objective such as to maximize expected profit after reinsurance (subject to a certain security level condition) or to minimize the probability of ruin after reinsurance (see **Ruin Theory**). There are a number of theoretical results available which under specific assumptions on the underlying premium principle identify the optimal reinsurance form with respect to given maximization criteria and constraints (see [4] for an overview). A natural and close connection between such optimality questions and utility theory (see **Utility Function**) is, for example, discussed by Borch [7] and Bühlmann and Jewell [9]. For instance, under an expected value premium principle of the reinsurer and a given reinsurance premium, an SL contract maximizes the expected utility of terminal wealth for any risk-averse utility function (a result attributed to K. Arrow, see **Arrow, Kenneth**). As a consequence, in this setting an SL contract also minimizes the retained variance $\text{Var}(D)$ of the first-line insurer. On the other hand, under a variance or standard deviation premium principle and a fixed value of $\text{Var}(R)$, a QS contract minimizes $\text{Var}(D)$ [5, 23]. Already these two examples illustrate that the selected reinsurance premium principle plays a decisive rule for the optimal decision. Also, the

optimality results usually heavily depend on the type and amount of transaction costs involved. In recent years many refinements of such optimality results have been developed (see [4, 21] for a general survey on this topic). For a dual approach in the framework of solvency in finance and insurance, see [29]. In the collective model, it is also quite common to look for reinsurance strategies that minimize the Lundberg adjustment coefficient (see **Cramér–Lundberg Estimates**; [15] and [10]). An overview on optimal dynamic reinsurance strategies to minimize the probability of ruin can be found in [24] (see **Ruin Theory and Stochastic Control in Insurance**). For other optimization criteria, see [30].

In practice, the concrete form and amount of reinsurance choice for a certain portfolio often also will be influenced by experience, availability of reinsurance offers, and current market prices.

Reinsurance and Finance

The financial component in any insurance and reinsurance activity has become more important over the last years, not the least due to higher interest rates, longer time horizons for claim settlement, and the general need to combine the analysis and management of assets and liabilities (see, e.g., [28]). Financial pricing of reinsurance contracts in various contexts is discussed in [11] (see **Insurance Derivatives**), and note that a (re)insurance market is incomplete, (see **Complete Markets**). Since reinsurance is a global business, foreign exchange risk (see **Foreign Exchange Markets**) also has to be considered (cf. [6]).

Nowadays, classical reinsurance is often complemented by other mechanisms to cope with risk (summarized under the name *alternative risk transfer*) that have a certain financial flavor such as *captives*, *finite risk reinsurance*, *multi-trigger products*, and *contingent capital* (cf. [22]). Further bridging the gap between insurance and finance, insurance companies also started to transfer some of their risk directly to the capital market (see **Securitization**). This includes *reinsurance sidecars*, where investors act like a QS reinsurer, as well as the issuance of catastrophe bonds (see **Catastrophe Bonds**). Recently, some structured insurance-linked products such as *catastrophe collateralized debt obligations* (CDOs) (see **Collateralized Debt Obligations (CDO)**) have also been issued in

the financial market, which sometimes receive quite competitive credit ratings for their tranches. At the same time, some of the traded catastrophe CDOs offer loss coverage that can be significantly cheaper than in classical reinsurance contracts.

References

- [1] Aase, K.K. (1993). Equilibrium in a reinsurance syndicate: existence, uniqueness and characterization, *ASTIN Bulletin* **23**, 185–211.
- [2] Aase, K.K. (1993). Premiums in a dynamic model of a reinsurance market, *Scandinavian Actuarial Journal* (2), 134–160.
- [3] Aase, K.K. (2002). Perspectives of risk sharing, *Scandinavian Actuarial Journal* (2), 73–128.
- [4] Albrecher, H. & Teugels, J. (2009). *Reinsurance: Actuarial and Financial Aspects*, John Wiley & Sons, Chichester. To appear.
- [5] Beard, R., Pentikäinen, T. & Pesonen, E. (1984). *Risk Theory*, Chapman & Hall, London.
- [6] Blum, P., Dacorogna, M., Embrechts, P., Neghawi, P. & Niggli, H. (2001). Using DFA for modelling the impact of foreign exchange risk on reinsurance decisions, *CAS Forum Summer*, 49–94.
- [7] Borch, K. (1961). The utility concept applied to the theory of insurance, *Astin Bulletin* **1**, 245–255.
- [8] Brazauskas, V. (2003). Influence functions of empirical nonparametric estimators of net reinsurance premiums, *Insurance, Mathematics and Economics* **32**(1), 115–133.
- [9] Bühlmann, H. & Jewell, W. (1979). Optimal risk exchanges, *Astin Bulletin* **10**, 243–262.
- [10] Centeno, M. (2002). Measuring the effects of reinsurance by the adjustment coefficient in the Sparre Anderson model, *Insurance, Mathematics and Economics* **30**(1), 37–49.
- [11] De Waegenaere, A. (1994). Equilibria in a mixed financial-reinsurance market with constrained trading possibilities, *Insurance, Mathematics and Economics* **14**(3), 205–218.
- [12] Denuit, M., Dhaene, J., Goovaerts, M. & Kaas, R. (2005). *Actuarial Theory for Dependent Risks*, Wiley, Chichester.
- [13] Doherty, N. & Smetters, K. (2006). Moral hazard in reinsurance markets, *Journal of Risk and Insurance* **72**(3), 375–391.
- [14] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events*, Springer, New York, Berlin, Heidelberg, Tokyo.
- [15] Gerber, H. (1979). *An Introduction to Mathematical Risk Theory*. Huebner Foundation Monograph 8, Huebner Foundation, Homewood, Ill.
- [16] Gerber, H. (1980). Principles of premium calculation and reinsurance, *Transactions of the 21th International Congress of Actuaries* **I**, 137–142.

-
- [17] Hürlimann, W. (1994). A note on experience rating, reinsurance and premium principles, *Insurance, Mathematics and Economics* **14**(3), 197–204.
 - [18] Hürlimann, W. (1995). Links between premium principles and reinsurance, *Transactions of the 25th International Congress of Actuaries*, Brussels, pp. 141–168.
 - [19] Jean-Baptiste, E. & Santomero, A. (2000). The design of private reinsurance contracts, *Journal of Financial Intermediation* **9**, 274–297.
 - [20] Kaas, R. (1993). How to (and how not to) compute stop-loss premiums in practice, *Insurance, Mathematics and Economics* **13**(3), 241–254.
 - [21] Mack, T. (1996). *Schadensversicherungsmathematik*. Verlag Versicherungswirtschaft E.V., Karlsruhe.
 - [22] Neuhaus, W. (2004). Alternative risk transfer, in *Encyclopedia of Actuarial Science*, J. Teugels & B. Sundt, eds John Wiley & Sons, Ltd., Chichester, Chapter.
 - [23] Pesonen, M.I. (1984). Optimal reinsurances, *Scandinavian Actuarial Journal* (2), 65–90.
 - [24] Schmidli, H. (2008). *Stochastic Control in Insurance*, Springer-Verlag London Ltd., London.
 - [25] Schnieper, R. (1990). Insurance premiums, the insurance market and the need for reinsurance, *Schweizerische Vereinigung der Versicherungsmathematiker Mitteilungen* **1990**(1), 129–147.
 - [26] Straub, E. (1984). Actuarial remarks on planning and controlling in reinsurance, *Astin Bulletin* **14**, 183–191.
 - [27] Thépaut, A. (1950). Une nouvelle forme de réassurance. le traité d'excédent du coût moyen relatif (ECOMOR), *Bulletin Trimestriel de l'Institut des Actuaire Français* **49**, 273.
 - [28] Wüthrich, M.V., Bühlmann, H., and Furrer, H. (2008). *Market-consistent Actuarial Valuation*, EAA Lecture Notes, Springer, Berlin.
 - [29] Yaari, M. (1987). The dual theory of choice under risk, *Econometrica* **55**, 95–115.
 - [30] Young, V.R. (1999). Optimal insurance under Wang's premium principle, *Insurance, Mathematics and Economics* **25**(2), 109–122.

HANSJÖRG ALBRECHER

Catastrophe Bonds

Catastrophe bonds (also called *cat bonds*) were developed to facilitate direct transfer of catastrophe insurance risk (see **Insurance Risk Models**) from (re)insurers (see **Reinsurance**) and corporations (referred to as *sponsors* for the cat bonds) to investors. As a market innovation in the form of insurance risk securitization, cat bonds were designed to protect sponsoring companies from financial losses caused by large natural catastrophes (such as hurricanes and earthquakes), as an alternative or supplement to traditional reinsurance (see **Reinsurance**). From the sponsor's perspective, the fully collateralized nature of cat bonds provides an important alternative source of capital capacity. From the investors' perspective, the past performances of cat bonds have been very attractive and cat bonds provide further diversification relative to the typical financial asset portfolios.

Cat losses covered under the cat bond are often characterized by a loss-exceedance curve, $S(x) = \Pr\{X > x\}$, that is, the probability that the cat loss X will exceed amount $\$x$. Investors of a cat bond are generally provided with a loss-exceedance curve $S(x)$ that is obtained either

- by running company exposure data through commercially available cat modeling software, or
- by designing payout functions along some parametric indicators (e.g., the Richter scale of an earthquake at a specified location, an aggregate industry loss index, etc.).

The following information is embedded in a loss-exceedance curve $S(x)$:

- the expected frequency of default (see **Credit Risk**); and
- the recovery rate, given default.

As compensation for credit risks (see **Credit Risk**), cat bonds normally offer investors yields that are higher than the risk-free interest rate (e.g., the London interbank offered rate (LIBOR); see **Bond**). The excess yields spread over the risk-free rate should more than compensate for the expected default rate, since it should also contain risk load for uncertainties associated with the default risk.

For investors, it is desirable to compare the relative attractiveness of the yields spreads between cat bonds and corporate bonds. In order to compare risk-adjusted performance of various asset classes, what is needed is a common yardstick that is applicable to all types of risks. For mutual funds (see **Mutual Funds**), a popular measure of risk-adjusted performance is the Sharpe ratio (see **Capital Asset Pricing Model**), namely, the excess return per unit of volatility. The Sharpe ratio works well for assets whose returns follow normal distributions. However, for a single cat bond issue, the traditional Sharpe ratio concept is not applicable since the asset return is skewed and with jumps: most of the probability mass is centered at zero loss, although there is a small probability of potentially large negative returns. Other pricing methods have been developed for cat bonds to capture the characteristics of cat bond losses.

Lane Pricing Model

Lane [1] proposed the following model for the cat bond yields:

$$\text{EER} = \gamma(\text{PFL})^\alpha(\text{CEL})^\beta \quad (1)$$

where probability of first loss (PFL) is the expected annual frequency of incurring a loss. Conditional expected loss (CEL) represents the average severity as percentage of principal given that a loss occurs. Expected excess return (EER) represents the adjusted spread over LIBOR, further net of expected loss (that is, $\text{PFL} \times \text{CEL}$). EER represents the risk loading.

Lane showed that his model (1) fits reasonably well to empirical prices for a dozen CAT bond transactions in 1999–2000, for which the best-fit parameters were $\gamma = 55\%$, $\alpha = 49\%$, and $\beta = 57\%$.

Probability Transform Method

Wang [2] establishes a pricing method by directly applying the following Wang transform to convert a loss curve S to a price curve S^* :

$$S^*(y) = t(\Phi^{-1}(S(y)) + \lambda) \quad (2)$$

where t represents Student- t distribution with degree of freedom k . The difference in the areas between S^* and S constitutes the risk load. Based on the same 1999–2000 cat bond transaction data as in [1], the estimated parameters of Wang transform in

equation (2) are $\lambda = 0.454$ and the degree of freedom $k = 5$. Here the student- t factor is essential to capture the low frequency and high-severity characteristic of credit losses associated with cat bonds.

References

- [1] Lane, M.N. (2000). Pricing risk transfer transactions, *ASTIN Bulletin* **30**(2), 259–293.

- [2] Wang, S.S. (2004). *Cat Bond Pricing Using Probability Transforms*. The Geneva Papers on Risk and Insurance—Issues and Practice.

Related Articles

Arbitrage Pricing Theory; Credit Risk; Bond; Insurance Risk Models; Mutual Funds; Reinsurance.

SHAUN WANG

Insurance Derivatives

Derivatives are financial instruments whose value depends on some underlying variables. Insurance derivatives are those derivatives that are related to insurance. Roughly speaking, there are two main types of insurance derivatives in the market. The first type is the so-called catastrophe derivatives (see **Catastrophe Bonds**). They are similar to ordinary financial derivatives like options (see **Options: Basic Definitions**) and futures (see **Forwards and Futures**) except that the underlying variables are not financial instruments but insured natural catastrophe risk. The second type is equity-linked products. These are policies sold by insurance companies in which payoffs to the investors are linked to the performance of other financial instruments like equities and stock index.

PCS Catastrophe Insurance Options

A traditional way that insurance companies can get capital injection (in the form of debt or equity) in the event of a natural catastrophe is through the so-called contingent capital programs. This process permits access to risk capital only if following a loss exceeds a certain threshold. However, insurance risk is not transferred in this arrangement. To enhance the claims-paying ability of insurers and reinsurers following severe natural catastrophes, and to take advantage of the global capital markets, securitization (see **Securitization**) of catastrophe insurance risk has been proposed in recent years so that the risk can be transferred from the insurance industry to the capital market.

Catastrophe insurance derivatives (CAT-derivatives) were first developed by the Chicago Board of Trade in 1992. The first generation of derivatives included futures and options of futures. For a detailed description of these contracts, we refer to [29] and [7]. Reference [37] contains a concise summary of catastrophe derivatives. In 1995, the first generation derivatives were replaced by property claim service (PCS) catastrophe insurance options. Underlying these options are some catastrophe loss indices compiled daily by PCS. These indices reflect estimated and published insured losses due to natural catastrophes for a defined region and a defined time period.

Let $X = (X_t)_{t \geq 0}$ be a stochastic process modeling a PCS loss index. The value of this process at time t can be interpreted as the accumulated insured losses (see **Accumulated Claims**) due to natural catastrophes up to time t . The payoff of a European type CAT-derivative with maturity T depends on the value of the index at time T , that is, $\text{payoff} = \phi(X_T)$ for some function ϕ .

Models of the PCS Loss Index

Since the index value represents the accumulated insured losses due to catastrophic events, and catastrophic events occur unpredictably, it is natural to model the loss index by a continuous time stochastic process with jumps (see **Jump Processes**) at random time points. For example, Cummins and Geman [16] modeled the instantaneous claim process as a geometric Brownian motion with jumps:

$$dS_t = \alpha S_{t-} dt + \sigma S_{t-} dW_t + k \cdot dN_t \quad (1)$$

where k is a positive constant representing the losses due to catastrophes and N is an independent Poisson process (see **Poisson Process; Point Processes**) with constant intensity. This Poisson process counts the number of catastrophic events that occurred in the past. The process S models the instantaneous loss amount, and hence accumulated losses up to time t (that is, the index value up to time t) is given by

$$X_t = \int_0^t S(u) du \quad (2)$$

Assuming the market is arbitrage-free (see **Arbitrage Strategy; Fundamental Theorem of Asset Pricing**), or, equivalently, the existence of an equivalent martingale probability measure (see **Equivalent Martingale Measures**), the valuation problem of futures contracts written on the loss index X is equivalent to that of Asian options (see **Asian Options**) written on S with zero exercising price. Therefore, the valuation problem of CAT-futures is transformed to a standard Asian option pricing problem. In [22], the authors have modeled accumulated losses directly by a jump diffusion:

$$dX_t = S_t dt + k dN_t \quad (3)$$

where the process S is a geometric Brownian motion (see **Black-Scholes Formula**) representing

the randomness in the reporting of the claims, k is a positive constant, and (N_t) is a Poisson process with constant intensity. This model preserves the quasi-completeness of insurance derivative markets obtained by applying the Delbaen and Haezendonck methodology [19] to the class of layers of reinsurance (see **Reinsurance**) replicating the call spreads (see **Call Spread**) [21]. To derive a unique price of insurance derivatives written on the index (X_t) , the authors assumed that the insurance market is both arbitrage free and complete. The justification of completeness is that one can choose an equivalent martingale measure that best fits the prices of various reinsurance layers. By using the techniques of stochastic time changes (see **Time Change**) and Laplace transforms, the authors obtained quasi-analytical solutions for CAT-option prices.

Aase [1] modeled the loss index using a compound Poisson process:

$$X_t = \sum_{k=1}^{N_t} Y_k \quad (4)$$

where N , as before, is a Poisson process counting the number of catastrophic events in the past, and Y_k is the jump size of the index at the random time point of occurrence of the k th catastrophe. It was assumed that $Y_1, Y_2, \dots$ form an independent and identically distributed sequence and are independent of the counting process N . The compound Poisson process is commonly used in actuarial risk theory (see **Insurance Risk Models**) to model aggregate losses. The loss index modeled in this way is time homogeneous with independent increments. These properties were justified empirically by the actuarial studies performed by Levi and Partrat [27]. Aase investigated the valuation problems of CAT-futures and derivatives on futures. In the framework of partial equilibrium theory with uncertainty, the author derived closed form formulas under some specific assumptions on the utility functions (see **Utility Function**). A similar model was used by Chang *et al.* [11] to evaluate CAT-futures and futures call spreads. They applied the randomized operational time approach to transform the compound Poisson process to a more trackable pure diffusion. Embrechts and Meister [20] generalized Aase's compound Poisson model to mixed compound Poisson process with stochastic intensity. Similarly, Dassios and Jang [18] used the Cox process to model the claim arrival process for catastrophic

events. As pointed out in [36], a market comprises a risk-free bond and a loss index alone is very often incomplete because of two reasons: (i) the underlying loss index (X_t) is not traded and hence it is impossible to construct a hedging portfolio (see **Hedging**) based on the bond and the index; and (ii) there is no instrument to hedge the stochastic jump sizes of the loss index.

Equity-linked Insurance Products

Traditionally, annuities are contracts sold by insurance companies to provide a stream of fixed payments over the life of the contract. In recent years, variable annuities in the United States are getting more popular with investors. In contrast to traditional annuity products, the payments of a variable annuity are linked to the performance of a portfolio of securities like stocks, bonds, and mutual funds. Through investing in variable annuities, investors can not only participate in the stock market but also partially hedge the downside risk of the market. Nowadays, variable annuities with guaranteed minimum withdrawal benefits (GMWBs) are quite common. Policyholders are allowed to withdraw funds on an annual or semiannual basis without a penalty as long as the withdrawal rate is less than a prespecified level. With the GMWB, the sum of cash flows received by the policyholder is guaranteed to be at least the original premium paid. Other common riders include guaranteed minimum death benefit (GMDB), guaranteed living benefit (GLB), guaranteed minimum income benefit (GMIB), and guaranteed minimum accumulation benefit (GMAB). For example, with the GMDB rider, a variable annuity will pay to the policyholder's beneficiary the greater of the account value or a guaranteed minimum amount (such as the total purchase payments). Products similar to variable annuities are also sold in other countries, such as unit-linked annuities in the United Kingdom and segregated funds in Canada. A concise description of the various guaranteed benefits in these contracts can be found in [29, 30].

A similar product is the equity-indexed annuity (EIA). EIA was first introduced by Keyport Life in 1995, and is one of the most successful innovations in the financial markets in the last decade. The notional amount of EIAs sold has increased dramatically in recent years. Essentially, EIA is an equity-linked

deferred annuity whose returns are based on the performance of an equity mutual fund or a stock index like S&P 500. Moreover, return on the initial premium is guaranteed. On each anniversary of the contract, the insurance company measures the growth of the underlying index and credits that calculated growth to the investors. If the growth is negative, the companies normally would still guarantee a certain minimum rate of return, say 3%. However, the guaranteed rate is typically much less than the rate of return offered on US Treasury securities with the same maturity. Since return on the initial premium is guaranteed, EIA can be regarded as a fixed interest deposit with an embedded call option (*see Call Options*) on an underlying index. Since options are implicitly embedded in EIA contracts, various option pricing and hedging techniques are applicable. For detailed discussions and descriptions on equity-linked insurance products, we refer to [23, 26, 32, 39, 40], and [25].

In the literature, the main emphases about variable annuities and EIA have been on pricing and hedging of the options embedded in the contracts. For example, Brennan and Schwartz [10], Boyle and Schwartz [9], and Bacinello and Ortu [3] used standard martingale techniques to price and hedge the embedded options in variable annuities under the assumption that the market is complete. Moller [32–34] investigates pricing and hedging of insurance contracts in the presence of mortality risk. Bacinello and Ortu [4, 5], Nielsen and Sandmann [37], Miltersen and Persson [31], Bacinello and Persson [6], and Zaglauer and Bauerb [41] studied the situation when interest rate risk presents, assuming that the financial market is complete. In an incomplete market, there is no unique fair option price. Perfect hedge by self-financing trading strategies does not exist. As an alternative to perfect hedge, Boyle and Hardy [8] and Hardy [24] studied delta hedging (*see Delta Hedging*) for the options embedded in contracts with GMDB. Another possible way is to construct hedging strategy that minimizes certain measure of the intrinsic risk. For instance, Moller [32–34] and Lin and Tan [27] studied the effect of mortality risk on the risk minimization hedging strategies in the presence of interest rate risk. Coleman *et al.* [14] studied hedging strategies for options embedded with ratchet features, under both equity (including jump) risk and interest rate risk. They assumed that the market is

complete under mortality risk, but the financial market is incomplete, due to a suitable equity model for fat tails or to discrete hedging. Coleman *et al.* [15] compared the relative performances of delta hedging and dynamic discrete risk minimization hedging strategies in the presence of both jump risk and volatility risk.

Recent research also studied the embedded withdrawal and surrender options of equity-linked insurance products. For example, Bacinello [2] used discrete-time binomial models (*see Binomial Tree*) to study the fair valuation problem of participating policies embedded with a surrender option. Chu and Kwok [13] analyzed the withdrawal right embedded in dynamically protected funds. Siu [38] studied the withdrawal and surrender options of a participating life insurance policy by modeling the underlying assets as Markov-modulated geometric Brownian motion. Cheung and Yang [12] studied the optimal surrender time for equity-linked products in a discrete-time regime-switching model (*see Regime-switching Models*). They examined the impact of the Markov transition matrix on the optimal surrender strategy. Milevsky and Salisbury [30] developed the pricing model of variable annuities with GMWB under both static and dynamic withdrawal policies. In [17], the authors have studied the optimal withdrawal strategy using techniques in singular stochastic control theory (*see Stochastic Control*).

References

- [1] Aase, K.K. (1999). An equilibrium model of catastrophe insurance futures and spreads, *Geneva Papers on Risk and Insurance* **24**, 69–96.
- [2] Bacinello, A.R. (2003). Fair valuation of a guaranteed life insurance participating contract embedding a surrender option, *Journal of Risk and Insurance* **70**, 461–487.
- [3] Bacinello, A.R. & Ortu, F. (1993). Pricing of equity-linked life insurance with endogenous minimum guarantees, *Insurance: Mathematics and Economics* **12**, 245–257.
- [4] Bacinello, A.R. & Ortu, F. (1993). Pricing of guaranteed security-linked life insurance under interest rate risk. Actuarial approach for financial risks. *Transactions of the Third AFIR International Colloquium*, Rome, 35–55.
- [5] Bacinello, A.R. & Ortu, F. (1994). Single and periodic premiums for guaranteed equity-linked life insurance under interest rate risk: the “lognormal \$+\$ \$ Vasicek” case, in *Financial Modeling*, L. Pecatti & M. Viren, eds, Physica-Verlag, 1–25.

- [6] Bacinello, A.R. & Persson, S.A. (2002). Design and pricing of equity-linked life insurance under stochastic interest rates, *Journal of Risk Finance* **3**, 6–21.
- [7] Banks, E. (2005). *Catastrophic Risk: Analysis and Management*. John Wiley & Sons, Chichester.
- [8] Boyle, P. & Hardy, M. (1997). Reserving for maturity guarantees: two approaches, *Insurance: Mathematics and Economics* **21**, 113–127.
- [9] Boyle, P. & Schwartz, E. (1977). Equilibrium prices of guarantees under equity-linked contracts, *Journal of Risk and Insurance* **44**, 639–680.
- [10] Brennan, M. & Schwartz, E. (1976). The pricing of equity-linked life insurance policies with an asset value guarantee, *Journal of Financial Economics* **3**, 195–213.
- [11] Chang, C.W., Chang, J.S.K. & Yu, M.T. (1996). Pricing catastrophe insurance futures call spreads: a randomized operational time approach, *Journal of Risk and Insurance* **63**, 599–616.
- [12] Cheung, K.C. & Yang, H. (2005). Optimal stopping behavior of equity-linked investment products with regime switching, *Insurance: Mathematics and Economics* **37**, 599–614.
- [13] Chu, C.C. & Kwok, Y.K. (2004). Reset and withdrawal rights in dynamic fund protection, *Insurance: Mathematics and Economics* **34**, 273–295.
- [14] Coleman, T.F., Kim, Y., Li, Y. & Patron, M. (2007). Robustly hedging variable annuities with guarantees under jump and volatility risks, *Journal of Risk and Insurance* **74**, 347–376.
- [15] Coleman, T.F., Li, Y. & Patron, M. (2006). Hedging guarantees in variable annuities under both equity and interest rate risks, *Insurance: Mathematics and Economics* **38**, 215–228.
- [16] Cummins, J.D. & Geman, H. (1994). An Asian option approach to the valuation of insurance futures contracts, *Review of Futures Markets* **13**, 517–557.
- [17] Dai, M., Kwok, Y.K. & Zong, J. (2008). Guaranteed minimum withdrawal benefit in variable annuities, *Mathematical Finance* **18**, 595–611.
- [18] Dassios, A. & Jang, J.K. (2003). Pricing of catastrophe reinsurance and derivatives using the Cox process with shot noise intensity, *Finance and Stochastic* **7**, 73–95.
- [19] Delbaen, F. & Haezendonck, J. (1989). A Martingale approach to premium calculation principles in an arbitrage-free market, *Insurance: Mathematics and Economics* **8**, 269–277.
- [20] Embrechts, P. & Meister, S. (1997). Pricing insurance derivatives, the case of CAT futures. *Proceedings of the 1995 Bowles Symposium on Securitization of Insurance Risk*, Georgia State University, Atlanta, Georgia. Society of Actuaries, Monograph M-FI97-1, 15–26.
- [21] Geman, H. (1994). CAT calls, *Risk* **7**, 86–89.
- [22] Geman, H. & Yor, M. (1997). Stochastic time changes in catastrophe option pricing, *Insurance: Mathematics and Economics* **21**, 185–193.
- [23] Grosen, A. & Jørgensen, P.L. (2000). Fair valuation of life insurance liabilities: the impact of interest rate guarantees, surrender options, and bonus policies, *Insurance: Mathematics and Economics* **26**, 37–57.
- [24] Hardy, M. (2000). Hedging and reserving for single-premium segregated fund contracts, *North American Actuarial Journal* **4**, 63–74.
- [25] Hardy, M. (2004). Options and guarantees in life insurance, in *Encyclopedia of Actuarial Science*, J.L. Teugels & B. Sundt, eds, John Wiley & Sons, 1216–1225.
- [26] Lee, H. (2001). Discussion of valuing equity-indexed annuities, *North American Actuarial Journal* **5**, 133–136.
- [27] Levi, C. & Partrat, C. (1991). Statistical analysis of natural events in the United States, *ASTIN Bulletin* **21**, 253–276.
- [28] Lin, X.S. & Tan, K.S. (2003). Valuation of equity-indexed annuities under stochastic interest rate, *North American Actuarial Journal* **7**, 72–91.
- [29] Meister, S. (1995). Contribution to the mathematics of catastrophe insurance futures. Technical report. Department of Mathematics, ETH Zürich.
- [30] Milevsky, M.A. & Posner, S.E. (2001). The titanic option: valuation of the guaranteed minimum death benefit in variable annuities and mutual funds, *The Journal of Risk and Insurance* **68**, 91–126.
- [31] Milevsky, M.A. & Salisbury, T.S. (2006). Financial valuation of guaranteed minimum withdrawal benefits, *Insurance: Mathematics and Economics* **38**, 21–38.
- [32] Miltersen, K.R. & Persson, S.A. (1999). Pricing rate of return guarantees in a HJM framework, *Insurance: Mathematics and Economics* **25**, 307–325.
- [33] Moller, T. (1998). Risk-minimizing hedging strategies for unit-linked life insurance contracts, *ASTIN Bulletin* **28**, 17–47.
- [34] Moller, T. (2001). Hedging equity-linked life insurance contracts, *North American Actuarial Journal* **5**, 79–95.
- [35] Moller, T. (2001). Risk-minimizing hedging strategies for insurance payment processes, *Finance and Stochastics* **5**, 419–446.
- [36] Muermann, A. (2001). *Pricing Catastrophe Insurance Derivatives*. LSE FMG Discussion Paper Series No. 400.
- [37] Muermann, A. (2004). Catastrophe derivatives, in *Encyclopedia of Actuarial Science*, J.L. Teugels & B. Sundt, eds, John Wiley & Sons, 231–236.
- [38] Nielsen, J.A. & Sandmann, K. (1995). Equity-linked life insurance: a model with stochastic interest rate, *Insurance: Mathematics and Economics* **16**, 225–253.
- [39] Siu, T.K. (2005). Financial fair valuation of participating policies with surrender options and regime switching, *Insurance: Mathematics and Economics* **37**, 533–552.

- [40] Streiff, T.F. & DiBiase, C.A. (1999). *Equity Indexed Annuities*, Dearborn, Chicago.
- [41] Tiong, S. (2000). *Equity indexed annuities in the black-scholes environment*. Doctoral thesis. The University of Iowa, Ames.
- [42] Zaglauer, K. & Bauer, D. (2007). Risk-neutral valuation of participating life insurance contracts in a stochastic interest rate environment, *Insurance: Mathematics and Economics* **43**, 29–40.

Related Articles

Catastrophe Bonds; Economic Capital Allocation; Insurance Risk Models; Option Pricing; General Principles; Reinsurance; Securitization.

KA CHUN CHEUNG

Heavy Tails in Insurance

Heavy-tailed Distributions

In the class of heavy-tailed distribution functions, *subexponential* distribution functions are a special class, which have just the right level of generality for risk measurement in insurance and finance models. The name arises from one of their properties: the fact that their right tail decreases more slowly than any exponential tail. This implies that large values can occur in a sample with nonnegligible probability, which proposes the subexponential distribution functions as natural candidates for situations where extremely large values occur in a sample compared to the mean size of the data. Such a pattern is often seen not only in insurance data, for instance, in fire, wind–storm, or flood insurance (collectively known as *catastrophe insurance*) but also in data from finance and communications engineering; see, for example [1, 6, 7]. Subexponential insurance claims can account for large fluctuations in the risk process of a company. Textbook accounts can be found in [2, 4, 6–8].

We present two defining properties of subexponential distribution functions. Let $(X_k)_{k \in \mathbb{N}}$ be independently and identically distributed positive random variables with distribution function F such that $F(x) < 1$ for all $x > 0$. Denote by $\overline{F}(x) = 1 - F(x)$ for $x \geq 0$, the tail of F , and for $n \in \mathbb{N}$,

$$\begin{aligned} \overline{F^{n*}}(x) &= 1 - F^{n*}(x) \\ &= \mathbb{P}(X_1 + \cdots + X_n > x), \quad x \geq 0 \end{aligned} \quad (1)$$

the tail of the n -fold convolution of F . We then say that F (or X) is subexponential (written $F \in \mathcal{S}$) if one of the following equivalent conditions holds:

- (a) $\lim_{x \rightarrow \infty} \frac{\overline{F^{n*}}(x)}{\overline{F}(x)} = n$ for some (all) $n \geq 2$ and
- (b) $\lim_{x \rightarrow \infty} \frac{\mathbb{P}(X_1 + \cdots + X_n > x)}{\mathbb{P}(\max(X_1, \dots, X_n) > x)} = 1$ for some (all) $n \geq 2$.

The heavy-tailedness of $F \in \mathcal{S}$ is demonstrated by the implications

$$F \in \mathcal{S} \implies \lim_{x \rightarrow \infty} \frac{\overline{F}(x-y)}{\overline{F}(x)} = 1 \quad \forall y \in \mathbb{R} \quad (2)$$

$$\implies \frac{\overline{F}(x)}{e^{-\varepsilon x}} \xrightarrow{x \rightarrow \infty} \infty \quad \forall \varepsilon > 0 \quad (3)$$

A famous subclass of $\mathcal{S}$ is the class of distribution functions with regularly varying tails; see [3]. For a positive measurable function f , we write $f \in \mathcal{R}(\alpha)$ for $\alpha \in \mathbb{R}$ (f is *regularly varying with index* α) if

$$\lim_{x \rightarrow \infty} \frac{f(tx)}{f(x)} = t^\alpha \quad \forall t > 0 \quad (4)$$

Let $\overline{F} \in \mathcal{R}(-\alpha)$ for $\alpha \geq 0$; then it has the representation

$$\overline{F}(x) = x^{-\alpha} \ell(x), \quad x > 0 \quad (5)$$

for some $\ell \in \mathcal{R}(0)$.

For regularly varying distribution tails, we can check (a) for $n = 2$ by splitting the convolution integral and use partial integration to obtain

$$\begin{aligned} \frac{\overline{F^{2*}}(x)}{\overline{F}(x)} &= 2 \int_0^{x/2} \frac{\overline{F}(x-y)}{\overline{F}(x)} dF(y) \\ &\quad + \frac{(\overline{F}(x/2))^2}{\overline{F}(x)}, \quad x > 0 \end{aligned} \quad (6)$$

Immediately, by equation (4), the last term tends to 0. The integrand satisfies $\overline{F}(x-y)/\overline{F}(x) \leq \overline{F}(x/2)/\overline{F}(x)$ for $0 \leq y \leq x/2$; hence, Lebesgue-dominated convergence applies and, since F satisfies equation (2), the integral on the right-hand side tends to 1 as $x \rightarrow \infty$. Examples of distribution functions with regularly varying tail include the Pareto, Burr, transformed beta (also called *generalized F*), log-gamma, and stable distribution functions (see Table 1.2.6 in [4]).

In addition, the lognormal, the two Benktander families, and the heavy-tailed Weibull (shape parameter less than 1) belong to $\mathcal{S}$; see again Table 1.2.6 in [4]. However, a direct proof of this is more difficult than that for the regularly varying case. Subexponentiality is typically established via an integral test for the hazard rate known as *Pitman's criterion* [2].

Application to Insurance

The two main examples in insurance risk theory that use the tail properties of subexponential distributions

are aggregated claims tails and ruin probabilities. The *aggregated claims* denoted by A (see **Accumulated Claims**) are defined as the sum of all claims to the insurance company in a given period. The usual model is

$$A = X_1 + \cdots + X_N \quad (7)$$

where N is the number of claims and $X_1, X_2, \dots$ are the *claim sizes* in the given time interval (N and $(X_k)_{k \in \mathbb{N}}$ are assumed to be independent and $(X_k)_{k \in \mathbb{N}}$ is an independently and identically distributed sequence). The tail of A and the associated quantiles are important for assessing the probability of big losses and for Value-at-Risk (see **Value-at-Risk**) calculations. The main result in the heavy-tailed case states that if the X_k are subexponential, then $(a(u) \sim b(u))$ as $u \rightarrow \infty$ means that $\lim_{u \rightarrow \infty} a(u)/b(u) = 1$

$$\mathbb{P}(A > x) \sim \mathbb{E}(N)\mathbb{P}(X_1 > x) \quad \text{as } x \rightarrow \infty \quad (8)$$

provided in addition $\mathbb{E}(z^N) < \infty$ for some $z > 1$ (cf. [2], Chapter IX, Lemma 2.2).

The classical insurance risk model is the *Cramér–Lundberg model* (cf. [2, 4, 6, 8] and see **Ruin Theory**), where the *claim times* occur at the jump times of a Poisson(λ) process $(N(t))_{t \geq 0}$ (see **Poisson Process**) and the claims $(X_k)_{k \in \mathbb{N}}$ are again an independently and identically distributed sequence with finite mean. The *risk process* is for *initial reserve* $u \geq 0$ and *premium rate* $c > 0$ defined as

$$R(t) = u + ct - \sum_{k=1}^{N(t)} X_k, \quad t \geq 0 \quad (9)$$

Then the *ruin probability* $\psi(u)$ in *infinite time* is the probability that R ever falls below 0, that is, $\psi(u) = \mathbb{P}(\inf_{t \geq 0} R(t) < 0)$. Define the *integrated tail distribution function* F_I by

$$F_I(x) = \frac{1}{\mathbb{E}(X_1)} \int_0^x \bar{F}(y) dy, \quad x \geq 0 \quad (10)$$

If F_I is subexponential (which is satisfied for all subexponential distributions F mentioned above under the condition that they have a finite mean) and $\rho = \lambda \mathbb{E}(X_1)/c < 1$, then

$$\psi(u) \sim \frac{\rho}{1 - \rho} \int_u^\infty \bar{F}_I(y) dy \quad \text{as } u \rightarrow \infty \quad (11)$$

a result that is often associated with the names of (in alphabetical order) Borovkov, Cohen, Embrechts, Pakes, Veraverbeke, and von Bahr.

That F_I plays a role may be understood from the fact that F_I is the distribution function of the first undershoot of $R(t) - u$ below 0 under the condition that $R(t)$ falls below u in finite time. The number M of times, where $R(t)$ achieves a new local minimum in finite time plus 1, is geometrically distributed with parameter $(1 - \rho)$, that is, $\mathbb{P}(M = n) = (1 - \rho)\rho^n, n \in \mathbb{N}_0$. This easily leads to the *Beekman–Bowers–Pollaczek–Khinchine formula*

$$\psi(u) = (1 - \rho) \sum_{n=1}^{\infty} \rho^n \bar{F}_I^{n*}(u), \quad u \geq 0 \quad (12)$$

from which equation (11) is an easy consequence.

For further aspects of ruin theory with heavy tails and a comprehensive set of recent references, see [5].

References

- [1] Adler, R. & Feldman, R.E. (1998). *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, Birkhäuser, Boston.
- [2] Asmussen, S. (2001). *Ruin Probabilities*, World Scientific, Singapore.
- [3] Bingham, N.H., Goldie, C.M. & Teugels, J.L. (1987). *Regular Variation*, Cambridge University Press, Cambridge.
- [4] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.
- [5] Fasen, V. & Klüppelberg, C. (2008). Large insurance losses distributions, in E. Melnick & B. Everitt, eds, *Encyclopedia of Quantitative Risk Analysis and Assessment*, John Wiley & Sons, 961–969.
- [6] Mikosch, T. (2004). *Non-Life Insurance Mathematics*, Springer, Berlin.
- [7] Resnick, S.I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer, New York.
- [8] Rolski, T., Schmidli, H., Schmidt, V. & Teugels, J. (1999). *Stochastic Processes for Insurance and Finance*, Wiley, Chichester.

Related Articles

Accumulated Claims; Heavy Tails; Ruin Theory.

SØREN ASMUSSEN, VICKY FASEN & CLAUDIA KLÜPPELBERG

Insurance Risk Models

The Classical Risk Model

We start by formulating the simplest risk model, which is based on the following independent objects:

1. a Poisson process (see **Poisson Process**) $N = \{N(t); t \geq 0\}$ with $N(0) = 0$ and intensity λ_0 , that is, $E[N(t)] = \lambda_0 t$;
2. a sequence $\{Z_k\}_1^\infty$ of independent and identically distributed (i.i.d.) random variables, having the common distribution function F , with $F(0) = 0$, mean value μ , and variance σ^2 .

The *risk process*, X , is defined by

$$X(t) = ct - \sum_{k=1}^{N(t)} Z_k, \quad \left(\sum_{k=1}^0 Z_k \stackrel{\text{def}}{=} 0 \right) \quad (1)$$

where c is a real constant.

This is the *classical model of the risk business* of an insurance company, where $N(t)$ is to be interpreted as the number of claims on the company during the interval $(0, t]$. At each point of N , the company has to pay out a stochastic amount of money, and the company receives (deterministically) c units of money per unit time. The constant c is called the *gross risk premium* rate.

Many studies based on this model treat *ruin theory* (see **Ruin Theory**). Basic results about ruin probabilities for the classical risk model go back to 1926 [17] and 1930 [6], while the general ideas underlying the risk model go even farther back to 1903 [16]. A stringent treatment, based on Wiener–Hopf methods, is due to Cramér [7].

In risk theory, the most interesting situation is when $c > 0$ and $F(0) = 0$. This case is generally called *positive risk sums*, and includes most nonlife branches and also the ordinary types of life insurance, where a certain amount of money is paid on the death of a policyholder.

There are, however, situations where the circumstances are reversed, that is, where $c < 0$ and $F(0) = 1$ (see **Life Insurance**). The typical example is life annuity—or pension—insurance, where a life annuity rate $-c$ is paid from the company to the policyholder and where the claim, that is, the death of the policyholder, will place an amount of money

corresponding to the “expected pension to be paid” at the company’s free disposal. Thus the claim means an income, or a negative cost, for the company. This situation is generally called *negative risk sums*.

Models for the Occurrence of Claims

A natural extension, at least from a mathematical point of view, of the Poisson process is to consider the case where the occurrence of the claims is described by a renewal process N . Let S_k denote the epoch of the k th claim. A point process is called a *renewal process* (with interoccurrence time distribution K^0) if the variables $S_1, S_2 - S_1, S_3 - S_2, \dots$ are independent and if $S_2 - S_1, S_3 - S_2, \dots$ have the same distribution K^0 . Further, N is called an *ordinary* renewal process if S_1 also has distribution K^0 . N is called a *stationary* renewal process if K^0 has finite mean $1/\lambda_0$ and if S_1 has distribution K given by

$$K(t) = \lambda_0 \int_0^t (1 - K^0(s)) \, ds \quad (2)$$

A Poisson process (see **Poisson Process**) is a renewal process with exponentially distributed interoccurrence times. In that case, the ordinary and stationary versions coincide. Probably, the Poisson process is the most realistic renewal process for describing the occurrence of the claims. This, however, does not mean that it is uninteresting to consider renewal models, since often better mathematical clarity is achieved by explicit use of an interoccurrence time distribution. In connection with ruin theory, the renewal model was first studied in [2] and the corresponding risk process is often called the *Sparre Andersen model*. A detailed treatment is given in [21].

Let us now go back to the Poisson case. It is natural to assume that λ_0 is proportional to the size of the insurance company. Naturally, the size of the company may vary in time. A simple way to take size variation into account is to let N be a nonhomogeneous Poisson process.

Definition 1 A point process N is called a nonhomogeneous Poisson process with intensity function $\lambda_0(t)$ if

- (i) $N(t)$ has independent increments;
- (ii) $N(t) - N(s)$ is Poisson distributed with mean $\int_s^t \lambda_0(\tau) \, d\tau$.

If $\lambda_0(t) \equiv \lambda_0$, we are back to the Poisson process. The *intensity measure* is defined by

$$\Lambda_0(t) \stackrel{\text{def}}{=} \int_0^t \lambda_0(s) \, ds, \quad t \geq 0 \quad (3)$$

It is natural to let $\Lambda_0(t)$ have the representation (3), but it is enough to assume that $\Lambda_0(t)$ is a nondecreasing function with $\Lambda_0(0) = 0$ and $\Lambda_0(t) < \infty$ for each $t < \infty$. If $\Lambda_0(t)$ is continuous, then N is *simple*, that is, $N(t)$ increases exactly one unit at its epochs of increase.

There may also be variation in the underlying risk. Typical examples are automobile insurance and fire insurance. It is natural to describe the underlying risk as a random process $\lambda = \{\lambda(t); t \geq 0\}$ and to let N be the corresponding *Cox process*.

We think of a Cox process N as generated in the following way. First, a realization λ_0 of a random process λ is generated. Conditioned upon that realization, N is a nonhomogeneous Poisson process with intensity function λ_0 . A Cox process is stationary if and only if the underlying random process is stationary.

Let equation (3) the *random intensity measure* be defined by

$$\Lambda(t) \stackrel{\text{def}}{=} \int_0^t \lambda(s) \, ds, \quad t \geq 0 \quad (4)$$

Assume that $E[\Lambda(t)^2] < \infty$ for all $t \geq 0$. Then

$$\begin{aligned} E[N(t)] &= E[\Lambda(t)] \quad \text{and} \quad \text{Var}[N(t)] \\ &= E[\Lambda(t)] + \text{Var}[\Lambda(t)] \quad \text{for all } t \geq 0 \end{aligned} \quad (5)$$

Thus a Cox process is overdispersed relative to a Poisson process.

For details of Cox processes, we refer to [12] or [13].

Example 1 The Mixed Poisson Process A Cox process with underlying random process $\lambda(t) \equiv \lambda$, where λ is a random variable, is called a *mixed Poisson process*. A mixed Poisson process where λ is Γ -distributed is called a *Pólya process* and for any fixed value of t the random variable $N(t)$ is negatively binomially distributed.

The mixed Poisson process is a natural model for the accident pattern of an individual policyholder. For instance, in automobile insurance, λ may describe the skillfulness of the individual as a driver.

The distribution of λ may be regarded as a prior distribution. When the company knows more about the driver, the posterior distribution may be used instead.

In 1940, Ove Lundberg had already presented his thesis [18] about Markov point processes and, more particularly, mixed Poisson processes. The mixed Poisson processes have ever since played a prominent role in actuarial mathematics. A survey of mixed Poisson processes is found in [15].

Example 2 The Ammeter Process Let $\Delta > 0$ be a fixed value and $\{L_k; k = 0, 1, \dots\}$ a sequence of nonnegative, i.i.d. random variables. A rather simple Cox process is obtained by letting

$$\lambda(t) = L_k \quad \text{for } k\Delta \leq t < (k+1)\Delta \quad (6)$$

The Cox process with this intensity is called an *Ammeter process*, since it is essentially the model considered by Ammeter [1]. A modern treatment of the Ammeter process is found in [14] and [15].

Example 3 The Markov-modulated Poisson Process Let $J = \{J(t); t \geq 0\}$ be an M -state Markov process with state space $\{1, 2, \dots, M\}$ and $\{\lambda_1, \lambda_2, \dots, \lambda_M\}$ a set of nonnegative numbers. The Cox process $N(t)$ with $\lambda(t) = \lambda_{J(t)}$ is called a *Markov-modulated Poisson process*.

An equivalent way to express N is to consider M independent Poisson processes $\{N_1, N_2, \dots, N_M\}$ with intensities $\{\lambda_1, \lambda_2, \dots, \lambda_M\}$ and to let the occurrence of the claims follow N_k when $J = k$. More formally, we let $dN(t) = dN_{J(t)}(t)$. This model was first studied in insurance by Reinhard [19], in the case $M = 2$ and exponentially distributed claims, and by Asmussen [3] in the general case.

Example 4 Independent Jump Intensity Intuitively, an independent jump intensity is a jump process where the jump times form a renewal process and where the value of the intensity between two successive jumps may depend only on the distance between these two jumps. More formally, let $\Sigma_k, k = 1, 2, \dots$ denote the epoch of the k th jump of the intensity process and let $\Sigma_0 \stackrel{\text{def}}{=} 0$. Put

$$\begin{aligned} \sigma_n &= \Sigma_n - \Sigma_{n-1} \\ L_n &= \lambda(\Sigma_{n-1}) \end{aligned} \quad n = 1, 2, 3, \dots \quad (7)$$

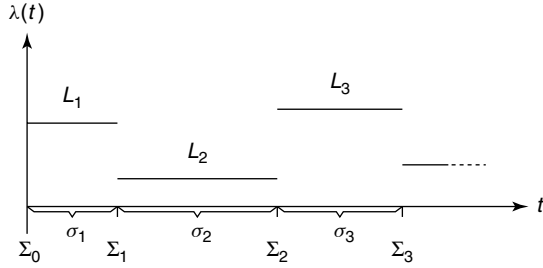


Figure 1 Illustration of notation

Here we understand that λ has right-continuous realizations so that $\lambda(t) = L_n$ for $\Sigma_{n-1} \leq t < \Sigma_n$. These notations are illustrated in Figure 1.

An intensity process λ is called an *independent jump intensity* if the random vectors

$$(L_1, \sigma_1), (L_2, \sigma_2), (L_3, \sigma_3), \dots$$

are independent and if $(L_2, \sigma_2), (L_3, \sigma_3), \dots$ have the same distribution.

This model was introduced in [4].

Models for Cost of Claims and Income

In the classical risk model, the random variables $\{Z_k\}_1^\infty$ represent the cost of the claims and the gross risk premium rate c , the income. We, when nothing else is said, assume that $\{Z_k\}_1^\infty$ is i.i.d. and independent of N .

Usually, the distribution F of Z_k is assumed to be either *exponentially bounded* (see **Cramér–Lundberg Estimates**) or to belong to the class $\mathcal{S}$ of *subexponential distributions* (see **Heavy Tails in Insurance**). These classes of distributions are also referred to as *light tailed* and *heavy tailed*, respectively.

Definition 2 A distribution F on $[0, \infty)$ belongs to $\mathcal{S}$ if

$$\lim_{z \rightarrow \infty} \frac{1 - F^{2*}(z)}{1 - F(z)} = 2 \quad (8)$$

The Γ -distribution is a typical example of an exponentially bounded distribution, while the Pareto distribution and the lognormal distribution are subexponential. A class of exponentially distributed distributions of great interest is that of the phase-type distributions (see **Phase-type Distribution**).

It may very well be the case that a light-tailed F is too “kind” while an $F \in \mathcal{S}$ is too “dangerous”. Then it may be tempting to try to find some intermediate case with nice properties. $\mathcal{S}(\beta)$ is such a class:

Definition 3 A distribution F on $[0, \infty)$ belongs to the class $\mathcal{S}(\beta)$, $\beta \geq 0$, if

- (i) $\hat{h}(\beta) < \infty$;
- (ii) $\lim_{z \rightarrow \infty} (1 - F^{2*}(z))/(1 - F(z)) = c$,
 $0 < c < \infty$;
- (iii) $\lim_{z \rightarrow \infty} (1 - F(z + y))/(1 - F(z)) = e^{-\beta y}$

for all real y .

Embrechts [11] has shown that certain generalized inverse Gaussian distributions belong to $\mathcal{S}(\beta)$.

The classes $\mathcal{S}$ and $\mathcal{S}(\beta)$ were introduced by Chistyakov [5].

Sometimes, it may be natural to allow for dependence between $\{Z_k\}_1^\infty$ and N . As an example, consider the Markov-modulated Poisson process motioned in Example 3. Let $\{X_1, X_2, \dots, X_M\}$ be M independent risk processes, as in equation (1), where X_k has Poisson intensity λ_k , claim cost distribution F_k , and gross risk premium rate c_k . In automobile insurance, it is natural to allow the weather situation to have influence on both the intensity and the cost. The risk business X now follows X_k when $J = k$, that is, $dX(t) = dX_{J(t)}(t)$.

The generalization of a constant premium rate c mentioned above, may not be the most natural one. Examples of more relevant extensions of the premium rate within the risk model are given below:

1. The premiums may depend on the result of the risk business. It is natural to let the premium decrease when risk business increases and *vice versa*.
2. Inflation and interest may be included in the model. If the interest is higher than the inflation, then the premium decreases when risk business decreases and *vice versa*.
3. Part of the risk business may be risky investments (see **Ruin Models with Investment Income**).

References [8, 10] are very readable studies focusing mainly on generalizations (1) and (2).

Early studies of generalizations (1) and (2) are found in [9, 20], respectively.

The question of “fair” premiums for an individual policyholder is of utmost importance for a company (see **Actuarial Premium Principles and Credibility Theory**). The discussion in Example 1 about prior and posterior distributions of λ is related to credibility theory.

References

- [1] Ammeter, H. (1948). A generalization of the collective theory of risk in regard to fluctuating basic probabilities, *Skandinavisk Aktuarietidskrift*, 171–198.
- [2] Sparre Andersen, E. (1957). On the collective theory of risk in the case of contagion between the claims, *Transactions XVth International Congress of Actuaries*, Vol. 2, New York, 219–229.
- [3] Asmussen, S. (1989). Risk theory in a Markovian environment, *Scandinavian Actuarial Journal*, 66–100.
- [4] Björk, T. & Grandell, J. (1988). Exponential inequalities for ruin probabilities in the Cox case, *Scandinavian Actuarial Journal*, 77–111.
- [5] Chistyakov, P. (1964). Теорема о суммах независимых положительных случайных величин и ее приложения к ветвящимся случайным процессам. Теория Вероятностей и ее Применения **9** 710–718 (English translation: A theorem on sums of independent random variables and its applications to branching processes. *Theory of Probability and its Applications* **9**, 640–648).
- [6] Cramér, H. (1930). *On the Mathematical Theory of Risk*, Skandia Jubilee Volume, Stockholm (Reprinted in 1994 *Harald Cramér Collected Works*, Vol. 1, Martin-Lof, A., ed, Skandia Insurance Company, Springer-Verlag, Berlin, p. 601–678).
- [7] Cramér, H. (1955). *Collective Risk Theory*, Skandia Jubilee Volume, Stockholm (Reprinted in *Harald Cramér Collected works Vol. II*. (1994), Ed. by Martin-Lof, A., Skandia Insurance Company, Springer-Verlag, Berlin, 1028–1115).
- [8] Dassios, A. & Embrechts, P. (1989). Martingales and insurance risk. *Communications in Statistics. Stochastic Models* **5**, 181–217.
- [9] Davidson, A.A. (1946). *Om ruinproblemet i den kollektiva riskeorin under antagande av variabel säkerhetsbelastning*, Försäkringsmatematiska studier tillägnade Filip Lundberg, Stockholm, 32–47 (English translation: (1969). On the ruin problem in the collective theory of risk under the assumption of variable safety loading. *Skand. AktuarTidskr.Suppl.* 70–83).
- [10] Delbaen, F. & Haezendonck, J. (1987). Classical risk theory in an economic environment. *Insurance: Mathematics and Economics* **6**, 85–116.
- [11] Embrechts, P. (1983). A property of the generalized inverse Gaussian distribution with some applications. *Journal of Applied Probability* **20**, 537–544.
- [12] Grandell, J. (1976). *Doubly Stochastic Poisson Processes*, Lecture Notes in Mathematics, 529, Springer-Verlag, Berlin.
- [13] Grandell, J. (1991). *Aspects of Risk Theory*, Springer-Verlag, New York.
- [14] Grandell, J. (1995). Some remarks on the Ammeter risk process. *Mitteilungen der Vereinigung schweizerischer Versicherungsmathematiker* **95**, 43–72.
- [15] Grandell, J. (1997). *Mixed Poisson processes*, Chapman & Hall, London.
- [16] Lundberg, F. (1903). *I. Approximerad Framställning av Sannolikhetsfunktionerna. II. Återförsäkring av Kollektivrisker*, Almqvist & Wiksell, Uppsala.
- [17] Lundberg, F. (1926). *Försäkringsteknisk Riskutjämning*, F. Englund's boktryckeri A.B., Stockholm.
- [18] Lundberg, O. (1964). *On Random Processes and their Application to Sickness and Accident Statistics*, 2nd Edition, Almqvist & Wiksell, Uppsala.
- [19] Reinhard, J.M. (1984). On a class of semi-Markov risk models obtained as classical risk models in a Markovian environment, *Astin Bulletin*, **16**, 23–43.
- [20] Segerdahl, C.O. (1942). Über einige risikothoretische Fragestellungen, *Skandinavisk Aktuarietidskrift*, 43–83.
- [21] Thorin, O. (1982). Probabilities of ruin, *Scandinavian Actuarial Journal* 65–102.

Related Articles

Actuarial Premium Principles; Cramér–Lundberg Estimates; Gerber–Shiu Function; Poisson Process; Point Processes; Ruin Theory; Ruin Models with Investment Income.

JAN GRANDELL

Solvency

Capital Requirements

Capital, most often, is the sum of contributed capital and retained earnings. In modern times, financial economics has developed a variety of other instruments to supply capital to companies. Capital is often referred to in the insurance industry as *surplus*. Whatever term is used to describe the concept (we shall here use the term *capital*), a primary concern is the amount of capital that must be held.

The amount of required capital will depend upon the methods used to determine policy liabilities within financial reporting since different methods provide varying degrees of conservatism that can be applied to offset risk and required capital. Therefore, a change in financial reporting may necessitate alteration of capital requirements.

Determination of required capital involves consideration of a number of factors including the time over which solvency is to be protected (the “time horizon”), the degree of probability to which future solvency is to be maintained, and the scope of threats to future solvency (the “risks”) to be considered in the calculation. The usual approach is to consider a variety of important risks; hence modern capital requirements are usually termed *risk based*.

The initial and historical approach to solvency and required capital is through ruin theory (*see Ruin Theory* and [8]). Here, the only source of risk considered is the insurance claims process. From a solvency perspective, the aim of the theory is to determine the amount of capital required to ensure that the probability of continued solvency (i.e., nonruin) up to some chosen time horizon is as high as desired. There are discrete and continuous time versions of the theory and the time horizon can be either finite or infinite. Ruin theory can definitely provide insight into the insurance process. However, its concentration on a single source of risk leads it to be most applicable to nonlife or general insurance and of much less use in life insurance (*see Life Insurance*) where financial risks play a much larger role. The theory is structured to calculate the probability of ruin given the initial available capital. The calculations are difficult and the theory has not often been applied in practice although it was

the basis for the first dynamic capital requirements created in Finland in the early 1940s.

An important advance came in the early 1980s with the availability of high-speed computers and the development of models of the insurance process that could be used to simulate an insurer’s financial progress. The leading work in this area was carried out in Finland by Pentikäinen and his colleagues [11]. To construct their model, this group first extended classical risk theory to include a number of significant variables, in addition to the claims process. This work is described in [12]. An extended, more modern, and readily available description of this approach can be found in [2].

Regulatory Approach to Solvency

Insurance regulators and supervisors are primarily charged with protecting the interests of policyholders and others holding contracts issued by insurance companies. They are clearly interested in insurers’ solvency. Historically, the emphasis in statutory financial reporting was on a conservative valuation of policy liabilities as a means of maintaining solvency. Although many jurisdictions required insurers to have a minimum amount of capital before being licensed to carry on the business of insurance, there were often no requirements that insurers had to maintain specific capital levels. In those jurisdictions where insurance companies were required to maintain capital of a uniform monetary amount, that amount usually did not vary by the size of the company, the nature of its portfolio of policies issued, its financial experience and condition, or its method of operation.

The earliest development of more modern risk-based capital requirements began in Finland in the 1940s; the determination of an insurer’s required capital was based upon ruin theory. The focus was upon excessively high claim frequency and severity. A second important step was the introduction, in the First Insurance Directive of the European Union, of a capital requirement for life insurance companies that was calculated in two parts: a fixed percentage of mathematical reserves and a smaller percentage of the net amount at risk. The portion related to mathematical reserves was meant to provide for credit and market risks related to the assets supporting these liabilities while the second portion was meant to provide for insurance (claims) risk.

Both the Finnish and the European innovations represent an effort to relate a company's capital requirement to the risks to which it is exposed. Further developments in this direction required a systematic study and listing of these sources of risk. The Committee on Valuation and Related Problems of the Society of Actuaries carried out one of the earliest such studies; its report was published in 1979 [4]. The committee suggested that risk could be categorized in three parts and that capital, or contingency reserves, could be provided for each part. The parts are C1, asset depreciation; C2, pricing inadequacy; and C3, interest-rate fluctuation (*see Bond and Term Structure Models*). In more modern terms, these would now be called C1, *credit risk*; C2, *insurance risk*; C3, *asset/liability mismatch risk*.

Work on the first modern risk-based capital requirement began in Canada in 1984, based upon the committee's categorization. By 1989, a test calculation had been established. This was subject to extensive field tests in the life insurance industry and became a regulatory requirement in 1992; it is known as the *Minimum Continuing Capital and Surplus Requirement* (MCCSR) [9]. A similar requirement was not established for Canadian property and casualty (general) insurers until 2002 [10]. In the United States, following an increase in the number of failures of insurance companies due to economic experience in the late 1980s and early 1990s, the National Association of Insurance Commissioners (NAIC) began work on a similar capital requirement for life insurers. The structure of the NAIC formula, known widely as *risk-based capital* (RBC), follows the committee's categorization and was quite similar to the Canadian formulation. Both national requirements have evolved independently; although they remain similar, there are significant differences between them. Shortly after it established the life insurance company formula, the NAIC introduced a capital requirement for property and casualty insurers. It has since introduced a separate requirement for health insurers.

A number of other jurisdictions have adopted regulatory capital requirements for insurance companies. These tend to follow the North American formulations.

The development of the more comprehensive insurance regulatory capital requirements was parallel to but independent of the similar development of a requirement for banks under the Basel Accord (*see*

Regulatory Capital). Although these developments occurred independently, future prospects are that the banking and insurance capital requirements will eventually have a strong influence on each other. This is due to two factors: the recent emergence of large financial conglomerates or groups that engage in both banking and insurance, and the trend in many countries to form integrated regulators/supervisors that oversee both industries. The formerly distinct cultures of banking and insurance are thereby influencing each other both in business and government.

Developments in Regulatory Capital since 2000

The first decade of the twenty-first century has seen a great deal of activity and advances in capital requirements and solvency. This is due to an increased appreciation of the risks faced by financial institutions, a heightened awareness of their need for capital, the emergence of the discipline of enterprise risk management, the drive to establish uniform international financial reporting standards (IFRS) by the International Accounting Standards Board (IASB), and regulatory moves, on a worldwide basis, by the International Association of Insurance Supervisors (IAIS) and in Europe by the European Union, to establish international supervisory standards.

Since its formation in the middle 1980s, the IAIS has been developing a series of papers on best practices for insurance supervisors. Solvency regulation and supervision has naturally been a major focus of IAIS activities [6]. In preparation for the development of papers on capital requirements, the IAIS asked the International Actuarial Association (IAA) for technical assistance. In 2002, the IAA formed its Insurer Solvency Assessment Working Party, which produced its report [5] in 2004. The IAIS continues to develop its series of papers on this topic.

The IAA report has become the conceptual basis for much of the recent work on insurance capital requirements. This is particularly the case with respect to work being done in Europe under the Solvency II project. The results of this project, a Europe-wide unified approach to insurance financial reporting and solvency supervision, is expected to come into effect in 2011. At the time of writing, the exact form of the Solvency II capital requirements has not been

fixed. The Committee of European Insurance and Occupational Pension Supervisors (CEIOPS), which has been charged with development of these requirements, has been conducting a series of qualitative impact studies that seek industry tests of proposed formulations of these requirements.

Advanced Approaches and Internal Models

An important feature of the Solvency II proposed capital requirement is the availability of an advanced approach for technically advanced insurers. Under this approach, insurers will determine their required capital using internal models that have been approved by supervisors. To be permitted to adopt this approach, insurers must satisfy a number of conditions involving not only the model itself but also the quality of implementation and execution of the insurer's risk management program. In particular, insurers will be required to satisfy a "use test" and show that the internal model is used within the firm's operations. This test is intended to assure that insurers have sufficient confidence in the quality and accuracy of their models.

Internal models have been used as part of the MCCR and RBC requirements in Canada and the United States, respectively. In particular, they have been applied to the liability valuation and capital requirements for certain financial guarantees that insurers offer in connection with certain (variable annuity) investment contracts. They have also been used for banking capital requirements under the Basel Accord, initially with respect to market risk in the trading book and more lately, under Basel II [1], with respect to credit and operational risks (*see Regulatory Capital and Operational Risk*).

Use of an advanced approach is likely to often produce lower capital requirements than would be produced by a standard and generally conservative regulatory formula. Supervisors justify accepting reduced capital because they require the financial institution to have a strong risk management program before being allowed to adopt an advanced approach.

The use of internal models was anticipated in the IAA report [5]. More recently the IAIS has published a standard and a paper on internal models, primarily directed to supervisors [7]. The IAA is preparing a companion paper directed to practitioners who construct and operate these models.

Beyond Regulatory Capital Requirements

The supervision of solvency extends beyond regulatory capital requirements. Many supervisors have adopted a structure that resembles one laid out in Basel II. The structure is described in terms of three Pillars: Pillar 1—Minimum Capital Requirements, Pillar 2—Supervisory Review Process, and Pillar 3—Market Discipline. The discussion heretofore has centered on Pillar 1. These are described in detail in [1].

An interesting feature of Pillar 2 involves the assessment of an insurer's capital adequacy, extending beyond minimum Pillar 1 requirements. An early and trail-blazing approach has been taken by the Financial Services Authority (FSA) of the United Kingdom with the introduction of individual capital adequacy assessment [3]. Each insurer is expected to have the ability to assess its own view of its capital needs, its economic capital. This assessment will usually be made with an internal model. The FSA may ask insurers to submit their models for examination. This is not a formal model approval. Rather, the FSA examines the model to ensure it covers all significant sources of risk for the particular firm and provides its views to the firm as individual capital guidance.

The third Pillar is based upon the notion that market discipline will act to force an insurer to operate its business in a sound and prudential manner, thus stimulating a convergence between the interests of the firm, its policyholders, and its supervisors. Supervisors' primary activity in this area centers on encouraging and/or requiring firms to publicly disclose information that is required by policyholders and investors to assess the firm's condition.

Solvency, Risk Management, and Economic Capital

Much of the thrust for improved capital requirements and risk management has come from the regulatory/supervisory sector. However, many insurers, particularly the larger and multinational firms, have undertaken their own initiatives to develop economic capital requirements and to apply them as an integral part of their business operations. In addition, these firms have adopted many ideas and practices from the growing field of enterprise risk management. The role of the chief risk officer has gained considerable

importance within these insurers. Many of the responsibilities that were formally assigned to the company actuary have been assumed by the chief risk officer. Economic capital and risk management have become the firm's operational tools to measure and control the risk of its own insolvency.

References

- [1] Basel Committee on Banking Supervision (2006). *Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework*, Bank for International Settlements, Basel.
- [2] Daykin, C.D., Pentikäinen, T. & Pesonen, M. (1994). *Practical Risk Theory for Actuaries*. Chapman & Hall, London.
- [3] Financial Services Authority *Individual Capital Assessment*, Prudential sourcebook for insurers (INSPRU 7), London, available at www.fsa.gov.uk.
- [4] Hickman, J.C., Cody, D.C., Maynard, J.C., Trowbridge, C.L. & Turner, S.H. (1979). Discussion of the preliminary report of the Committee on Valuation and Related Problems, *Record of the Society of Actuaries* 5, 241–284.
- [5] International Actuarial Association (2004). *A Global Framework for Insurer Solvency Assessment*, report of the Insurer Solvency Assessment Working Party, IAA, Ottawa, available at www.actuaries.org.
- [6] International Association of Insurance Supervisors (2007). *Guidance Paper no. 2.1.1 on the Structure of Regulatory Capital Requirements*, available at www.iaisweb.org.
- [7] International Association of Insurance Supervisors (2007). *Guidance Paper no. 2.2.7 on the Use of Internal Models for Risk and Capital Management Purposes by Insurers*, available at www.iaisweb.org.
- [8] Klugman, S.A., Panjer, H.H. & Willmot, G.E. (2004). *Loss Models: From Data to Decisions, Second Edition*, Wiley, New York.
- [9] Office of the Superintendent of Financial Institutions Canada (2007). *Minimum Continuing Capital and Surplus Requirements (MCCSR) for Life Insurance Companies*, Life Guideline A, Ottawa, available at www.osfi-bsif.gc.ca.
- [10] Office of the Superintendent of Financial Institutions Canada (2008). *Minimum Capital Test (MCT)*, P&C Guideline A, Ottawa, available at www.osfi-bsif.gc.ca.
- [11] Pentikäinen, T. (ed) (1982). *Solvency of Insurers and Equalization Reserves; vol. 1 General Aspects*, Insurance Publishing Co., Helsinki.
- [12] Rantala, J. (ed) (1982). *Solvency of Insurers and Equalization Reserves vol. 2 Risk Theoretical Model*, Insurance Publishing Co., Helsinki.

Related Articles

Economic Capital; Internal-ratings-based Approach; Regulatory Capital; Ruin Theory.

ALLAN BRENDER

Cramér's Theorem

Let $X, X_1, X_2, \dots$ be independent and identically distributed (i.i.d.) random variables with partial sums $S_n = \sum_{k=1}^n X_k$ and empirical means $\hat{S}_n = n^{-1} S_n$ for $n = 1, 2, \dots$. Assuming a finite mean $\mu = E[X]$, by the law of large numbers, it holds for every $x > \mu$ that

$$\lim_{n \rightarrow \infty} P(\hat{S}_n > x) = 0 \quad (1)$$

Cramér's theorem refines this result by identifying an exact exponential decay rate of the small probability in equation (1). Note that the central limit theorem cannot, in general, capture this subtle decay rate because the central limit theorem focuses on average-sized fluctuations while the rare event $(\hat{S}_n > x)$ is mainly determined by large values of the random variables $X_1, \dots, X_n$.

Study of the probability of such a rare event has important applications in insurance and finance. For example, in insurance context, we may understand the random variables $X_1, \dots, X_n$ as respective claim amounts of n policy holders during a time period or of a single policy holder during n time periods. The event $(\hat{S}_n > x)$ with $x \geq \mu + \delta$, where $\delta > 0$ is interpreted as the safety loading, describes a critical situation that the insurance company loses money. To make sure that the probability of this rare event is sufficiently small, we need to determine what value of δ is appropriate. A good approximation for the probability $P(\hat{S}_n > x)$ enables us to get a simple yet powerful solution to this problem.

Let us take another example from financial risk management. Denote by $Q_\alpha[S_n]$ and $\text{CTE}_\alpha[S_n]$, respectively, the quantile (value at risk) and conditional tail expectation of the sum S_n of n risk variables, where $\alpha \in (0, 1)$ represents the confidence level, usually chosen to be close to 1 (see **Value-at-Risk** and **Expected Shortfall**). By definition,

$$Q_\alpha[S_n] = \inf\{y : P(S_n > y) \leq 1 - \alpha\} \quad (2)$$

and

$$\begin{aligned} \text{CTE}_\alpha[S_n] &= E[S_n | S_n > Q_\alpha[S_n]] \\ &= Q_\alpha[S_n] + \frac{\int_{Q_\alpha[S_n]}^{\infty} P(S_n > y) dy}{P(S_n > Q_\alpha[S_n])} \end{aligned} \quad (3)$$

Hence, an efficient approximation for the probability in equation (1) is desired for calculation of both $Q_\alpha[S_n]$ and $\text{CTE}_\alpha[S_n]$.

Define the cumulant generating function of X as

$$\Lambda(\lambda) = \ln E[e^{\lambda X}], \quad \lambda \in \mathbb{R} \quad (4)$$

Note that $\Lambda(0) = 0$ and $-\infty < \Lambda(\lambda) \leq \infty$ for all $\lambda \in \mathbb{R}$. Further define the Fenchel–Legendre transform of $\Lambda(\cdot)$ as

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda(\lambda)\}, \quad x \in \mathbb{R} \quad (5)$$

It can be shown that both $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$ are convex (but not necessarily strictly convex) functions on $\mathbb{R}$ and that $\Lambda^*(\cdot)$ is lower semicontinuous (i.e., $\Lambda^*(\cdot)$ takes values in $[0, \infty]$ and is such that, for every $y \in [0, \infty)$, the level set $\{x : \Lambda^*(x) \leq y\}$ is closed).

The following are some concrete and easily verifiable examples for $\Lambda^*(\cdot)$ (see, e.g., page 35 of [3]):

1. If X follows a Poisson distribution with mean $\theta > 0$ then $\Lambda^*(x) = \theta - x + x \ln \frac{x}{\theta}$ for $x \geq 0$ and $\Lambda^*(x) = \infty$ otherwise.
2. If X follows a Bernoulli distribution with success probability $p \in (0, 1)$ then $\Lambda^*(x) = x \ln \frac{x}{p} + (1-x) \ln \frac{1-x}{1-p}$ for $x \in [0, 1]$ and $\Lambda^*(x) = \infty$ otherwise.
3. If X follows an exponential distribution with mean $1/\theta$ then $\Lambda^*(x) = \theta x - 1 - \ln(\theta x)$ for $x > 0$ and $\Lambda^*(x) = \infty$ otherwise.
4. If X follows a normal distribution with mean μ and variance σ^2 then $\Lambda^*(x) = \frac{(x - \mu)^2}{2\sigma^2}$.

For a Borel set Γ , denote by $\bar{\Gamma}$ and $\bar{\Gamma}$ its interior and closure, respectively. Throughout, $\inf \emptyset = \infty$, by convention.

Cramér's Theorem For every Borel set Γ with nonempty interior,

$$\begin{aligned} -\inf_{x \in \bar{\Gamma}} \Lambda^*(x) &\leq \liminf_{n \rightarrow \infty} \frac{1}{n} \ln P(\hat{S}_n \in \Gamma) \\ &\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \ln P(\hat{S}_n \in \Gamma) \\ &\leq -\inf_{x \in \bar{\Gamma}} \Lambda^*(x) \end{aligned} \quad (6)$$

A sequence of probability measures, $\{P_n\}$, is said to satisfy the large deviation principle (see **Large Deviations**) with a rate function $I(\cdot)$ if, for every Borel set Γ ,

$$\begin{aligned} -\inf_{x \in \bar{\Gamma}} I(x) &\leq \liminf_{n \rightarrow \infty} \frac{1}{n} \ln P_n(\Gamma) \\ &\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \ln P_n(\Gamma) \leq -\inf_{x \in \bar{\Gamma}} I(x) \end{aligned} \quad (7)$$

Thus, Cramér's theorem shows that the probability laws of the empirical means $\hat{S}_n$ satisfy the large deviation principle with rate function $\Lambda^*(\cdot)$ (see **Large Deviations**).

Note that Cramér's theorem does not require $\Lambda(\lambda) < \infty$ for all $\lambda \in \mathbb{R}$. It is applicable even when the mean of X does not exist. For $\mu \in \bar{\Gamma}$, Cramér's theorem does not contribute because both sides of equation (6) turn out to be zero. If $\inf_{x \in \bar{\Gamma}} \Lambda^*(x) = \inf_{x \in \bar{\Gamma}} \Lambda^*(x)$, as is often the case in applications, then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln P(\hat{S}_n \in \Gamma) = -\inf_{x \in \bar{\Gamma}} \Lambda^*(x) \quad (8)$$

or, equivalently,

$$P(\hat{S}_n \in \Gamma) = \exp \left\{ -n \left(\inf_{x \in \bar{\Gamma}} \Lambda^*(x) + o(1) \right) \right\} \quad (9)$$

where $o(1)$ tends to 0 as $n \rightarrow \infty$. It can be shown that, for every real number y ,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln P(\hat{S}_n \geq y) = -\inf_{x \geq y} \Lambda^*(x) \quad (10)$$

see, for example, Corollary 2.2.19 of [3].

The first statement of Cramér's theorem for distributions possessing densities was given by Cramér [2], who applied it to model the insurance business (see **Cramér–Lundberg Estimates**). This is the first rigorous result concerning large deviations and is viewed as the historical starting point of large deviation theory. An extension to general distributions was done by Chernoff [1]. Cramér's theorem is also often cited as the Cramér–Chernoff theorem in the literature of large deviation theory.

Various versions of Cramér's theorem have been included in many monographs in large deviation theory. This most general version of Cramér's theorem is due to [3]. Textbook treatments of Cramér's theorem can be found in [6, 16], and [10], among others.

A similar result for the empirical means of i.i.d. random vectors in $\mathbb{R}^d$ has been established. A further extension to dependent, not necessarily identically distributed, random vectors in $\mathbb{R}^d$ leads to the Gärtner–Ellis theorem.

Large Deviations Without Cramér's Condition

Note that if $\Lambda(\lambda) = \infty$ for all $\lambda \neq 0$, then $\Lambda^*(x) = 0$ for all $x \in \mathbb{R}$ and consequently equation (7) gives a trivial approximation

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln P(\hat{S}_n \in \Gamma) = 0 \quad (11)$$

Similarly, if $\Lambda(\lambda) = \infty$ for all $\lambda > 0$, then $\Lambda^*(x) = 0$ for all $x \geq \mu$ and, in this case, Cramér's theorem leads to equation (11) with $\Gamma = [\mu, \infty)$, while if $\Lambda(\lambda) = \infty$ for all $\lambda < 0$ then $\Lambda^*(x) = 0$ for all $x \leq \mu$ and in this case Cramér's theorem leads to equation (11) with $\Gamma = (-\infty, \mu]$. To avoid such trivialities, it is necessary to assume $\Lambda(\lambda) < \infty$ in a neighborhood of $\lambda = 0$, which is usually referred to as *Cramér's condition*.

Study of large deviations for heavy-tailed random variables for which Cramér's condition is violated forms another important part of large deviation theory. The concept of subexponentiality plays a central role during the study in this direction. By definition, the common distribution of i.i.d. random variables $X, X_1, X_2, \dots$ is said to be subexponential (on the right)

if the relation

$$\lim_{x \rightarrow \infty} \frac{P\left(\sum_{k=1}^n X_k^+ > x\right)}{P(X^+ > x)} = n \quad (12)$$

holds for some (or, equivalently, for all) $n = 2, 3, \dots$, where $X^+ = \max\{X, 0\}$ (see **Heavy Tails in Insurance**). We remark that many popular distributions such as Pareto, lognormal, and heavy-tailed Weibull distributions are subexponential; see, for example, [5]. The very definition of subexponentiality opens a natural way to establish

$$\lim_{n \rightarrow \infty} \sup_{x \geq x_n} \left| \frac{P(S_n > x)}{nP(X > x)} - 1 \right| = 0 \quad (13)$$

for appropriate constants x_n serving as a threshold. The uniform asymptotic relation (13) is at the core of the mainstream study of large deviations in the presence of heavy tails. We usually say that results like equation (13) give precise asymptotics for large deviations while results like equations (6) and (10) give rough asymptotics for large deviations.

Pioneering works in the study of precise large deviations include [12, 13] and [7–9]. Overviews are available in [14] and [11], the latter of which contains a table summarizing choices of the threshold sequence $\{x_n\}$ in equation (13) for most subexponential distributions. Some recent developments of precise large deviations can be found in [15] and [4], among others.

References

- [1] Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Annals of Mathematical Statistics* **23**, 493–507.
- [2] Cramér, H. (1938). Sur un nouveau théorème-limite de la théorie des probabilités, *Actualités Scientifiques et Industrielles* **736**, 5–23.
- [3] Dembo, A. & Zeitouni, O. (1998). *Large Deviations Techniques and Applications*, 2nd Edition, Springer-Verlag, New York.
- [4] Denisov, D., Dieker, A.B. & Shneer, V. (2008). Large deviations for random walks under subexponentiality: the big-jump domain, *The Annals of Probability* **36**, 1946–1991.
- [5] Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, Berlin.
- [6] Gulinsky, O.V. & Veretennikov, A.Y. (1993). *Large Deviations for Discrete-time Processes with Averaging*, VSP, Utrecht.
- [7] Heyde, C.C. (1967). A contribution to the theory of large deviations for sums of independent random variables, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **7**, 303–308.
- [8] Heyde, C.C. (1967). On large deviation problems for sums of random variables which are not attracted to the normal law, *Annals of Mathematical Statistics* **38**, 1575–1578.
- [9] Heyde, C.C. (1968). On large deviation probabilities in the case of attraction to a non-normal stable law, *Sankhyā Series A* **30**, 253–258.
- [10] Klenke, A. (2008). *Probability Theory—A Comprehensive Course*, Springer-Verlag London, Ltd., London.
- [11] Mikosch, T. & Nagaev, A.V. (1998). Large deviations of heavy-tailed sums with applications in insurance, *Extremes* **1**, 81–110.
- [12] Nagaev, A.V. (1969). Integral limit theorems with regard to large deviations when Cramér's condition is not satisfied, *Theory of Probability and its Applications* **14**, 51–64.
- [13] Nagaev, A.V. (1969). Integral limit theorems with regard to large deviations when Cramér's condition is not satisfied, *Theory of Probability and its Applications* **14**, 193–208.
- [14] Nagaev, S.V. (1979). Large deviations of sums of independent random variables, *Annals of Probability* **7**, 745–789.
- [15] Tang, Q. (2006). Insensitivity to negative dependence of the asymptotic behavior of precise large deviations, *Electronic Journal of Probability* **11**, 107–120.
- [16] Varadhan, S.R.S. (1984). *Large Deviations and Applications*, SIAM, Philadelphia, PA.

QIHE TANG

Accumulated Claims

For measuring *operational risk* (see **Operational Risk**) or the risk of an insurance contract or an insurance portfolio, one has to find “good” models for the losses of the contract. These models need to be simple in order to be able to get desired answers but complicated enough to reflect the “real” situation. To achieve this goal, there are two different approaches: one approach is to model the number of claims and the individual claim sizes and the other is to model the distribution of the accumulated claim sizes directly. We consider the first approach in the second section and the second approach in the third section. In the fourth section, we discuss numerical methods to calculate the distribution function.

Compound Models

The most natural way to model the accumulated claims distribution is the distribution of a random sum

$$S = \sum_{k=1}^N Y_k \quad (1)$$

where $N \in \{0, 1, \dots\}$ is the (random) number of claims and Y_k is the (random) size of the k th claim. Here, we make the convention $\sum_{k=1}^0 Y_k = 0$. In a portfolio where claims do not occur simultaneously, one can assume that N and $\{Y_k\}$ are independent and that $\{Y_k\}$ are independent and identically distributed (i.i.d.). If claims occur simultaneously—as, for instance, claims caused by natural catastrophes—one has to consider Y_k as the accumulated claim from an event. The distribution of S is then called a *compound distribution* (see **Ruin Theory; Insurance Risk Models**).

One often needs some characteristics of the distribution of S . By conditioning on N , one finds

$$\mathbb{E}[S] = \mathbb{E}[N]\mathbb{E}[Y] \quad (2)$$

$$\text{Var}[S] = \text{Var}[N]\mathbb{E}[Y]^2 + \text{Var}[Y]\mathbb{E}[N] \quad (3)$$

$$\mathbb{E}[e^{rS}] = \mathbb{E}[\exp\{N \log \mathbb{E}[e^{rY}]\}] \quad (4)$$

We see that the characteristics can be easily calculated from the corresponding quantities for N and Y .

Let $p_k = \mathbb{P}[N = k]$. Popular models for N are the binomial distribution $p_k = \binom{n}{k} q^k (1-q)^{n-k}$, $n \in \mathbb{N}$, $q \in (0, 1)$; the Poisson distribution $p_k = e^{-\lambda} \lambda^k / k!$, $\lambda > 0$; the negative binomial distribution $p_k = \binom{\alpha + k - 1}{k} q^\alpha (1-q)^k$, $\alpha > 0$, $q \in (0, 1)$; and the logistic distribution $p_k = -q^{n+1} / [(n+1) \log(1-q)]$, $q \in (0, 1)$.

The Poisson distribution appears as a limit. Suppose that there are m independent individual contracts where a claim occurs with probability $q^{(j)}$. The average number of claims is approximately $\lambda = \sum_{j=1}^m q^{(j)}$. If now $m \rightarrow \infty$ such that λ remains constant and $\sum_{j=1}^m (q^{(j)})^2 \rightarrow 0$, then the number of claims converges to a Poisson distribution with parameter λ .

The compound Poisson model has a few more nice properties. Suppose that there are m portfolios $S^{(j)}$ modeled as compound Poisson with parameter $\lambda^{(j)}$ and the distribution $G^{(j)}(y)$ of the increments. Then $S = \sum_{j=1}^m S^{(j)}$ has a compound Poisson distribution with parameter $\lambda = \sum_{j=1}^m \lambda^{(j)}$ and distribution of the increments

$$G(y) = \sum_{j=1}^m \lambda^{-1} \lambda^{(j)} G^{(j)}(y) \quad (5)$$

Conversely, if S has a compound Poisson distribution and $B^{(j)}$ are sets such that $B^{(i)} \cap B^{(j)} = \emptyset$ for $i \neq j$, then $S^{(j)} = \sum_{k=1}^N Y_k \mathbb{I}_{Y_k \in B^{(j)}}$ are independent and compound Poisson distributed with parameter $\lambda^{(j)} = \lambda \mathbb{P}[Y \in B^{(j)}]$ and $Y_k^{(j)}$ has the distribution $\mathbb{P}[Y \in \cdot \mid Y \in B^{(j)}]$.

Easy treatable models are also mixed Poisson models, where $\Lambda \geq 0$ is a random variable, and N conditioned on Λ has a Poisson distribution with parameter Λ . One finds that

$$\mathbb{E}[N] = \mathbb{E}[\Lambda] \quad (6)$$

$$\text{Var}[N] = \text{Var}[\Lambda] + \mathbb{E}[\Lambda] \quad (7)$$

$$\mathbb{E}[e^{rN}] = \mathbb{E}[\exp\{\Lambda(e^r - 1)\}] \quad (8)$$

For example, if Λ follows a gamma distribution, which means that Λ has the density $\beta^\gamma x^{\gamma-1} e^{-\beta x} \mathbb{I}_{x>0} / \Gamma(\gamma)$, then N has a negative binomial distribution with $\alpha = \gamma$ and $q = \beta / (\beta + 1)$.

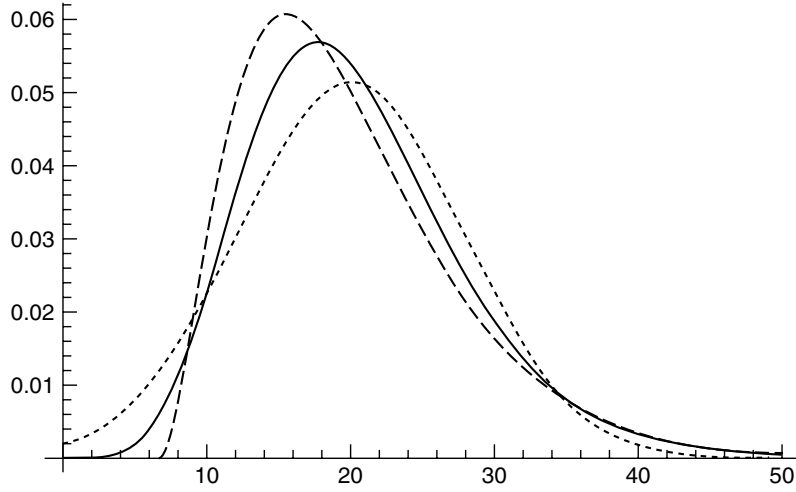


Figure 1 Densities of S (solid line), its normal approximation (dotted line), and its translated gamma approximation (dashed line)

Further risk models and details can be found in [1, 3, 4, 7, 9]. Applications to operational risk are described in [2, 5, 6].

Approximations

The calculation of the distribution of a compound sum is in general not possible. It involves n -fold convolutions, that is, n -fold integrals, which numerically is a problem. One, therefore, has to use approximations to the aggregate claims distribution. Let us denote by $\mu_k = \mathbb{E}[S^k]$ the moments of S , by $\sigma^2 = \mu_2 - \mu_1^2$ the variance, and by $\beta = \mathbb{E}[(S - \mu)^3]\sigma^{-3/2}$ the coefficient of skewness.

The *normal approximation* is the normal distribution with the same mean and variance as S . This approximation usually does not fit well because it does not take care of the skewness.

An approximation that usually works well is the *translated gamma approximation*. Let $\gamma = 4\beta^{-2}$, $\alpha = 2/(\beta\sigma)$, and $k = \mu - \gamma/\alpha$. Let Z be a gamma-distributed random variable with parameters γ and α . Then $Z + k$ has the same first three moments as S . That three moments are matched is also a reason for the good fit.

As an example we consider a compound Poisson distribution with $\lambda = 20$ and Pareto-distributed claim sizes with mean 1 and variance 2, that is,

$\mathbb{P}[Y > y] = (3/(3 + y))^4$. Figure 1 shows the densities of the exact distribution and its normal and translated gamma approximations. Both approximations do not fit well close to the mean value 20, but for risk management this region is not very interesting. We can see that in the right tail the normal approximation underestimates the true density considerably, whereas the fit in the tail of the translated gamma approximation is quite good.

Further possibilities for approximations are Edgeworth expansions, saddle point approximations, or the normal power approximation, [3, 4, 7, 9].

Panjer Recursion

Suppose that $Y_i > 0$ takes values in $\mathbb{N}$. Then also S takes values in $\mathbb{N}$. If we denote by $f_k = \mathbb{P}[Y = k]$ and $f_k^{*n} = \mathbb{P}[Y_1 + \dots + Y_n = k]$, then $g_k = \mathbb{P}[S = k]$ can be calculated by $g_0 = p_0$ and $g_n = \sum_{k=1}^n p_k f_n^{*k}$. The convolutions are obtained recursively: $f_k^{*(n+1)} = \sum_{\ell=1}^{k-1} f_\ell^{*n} f_{k-\ell}$, where $f_\ell^{*1} = f_\ell$. The problem is that numerous calculations have to be executed, which makes the procedure slow.

A faster procedure can be found if for $r \geq 1$

$$p_r = \left(a + \frac{b}{r}\right) p_{r-1} \quad (9)$$

for some constants a, b . It turns out that the distribution of N must be binomial ($a < 0$, $-b/a \in \mathbb{N}$), Poisson ($a = 0$, $b > 0$), or negative binomial ($a \in (0, 1)$, $a + b > 0$). In these cases, we have $g_0 = p_0$ and recursively

$$g_r = \sum_{k=1}^r \left(a + \frac{bk}{r} \right) f_k g_{r-k} \quad (10)$$

The method is fast and is also stable for the compound Poisson and the compound negative binomial distributions. This method goes back to Panjer [8].

Generalizations of formula (10) are treated for the case where $f_0 > 0$ or the case where the recursion (9) holds for $r \geq \ell > 1$ [4, 10–12].

References

- [1] Asmussen, S. (2000). *Ruin Probabilities*, World Scientific, Singapore.
- [2] Cruz, M.G. (2002). *Modeling, Measuring and Hedging Operational Risk*, Wiley, New York.
- [3] Daykin, C.D., Pentikäinen, T. & Pesonen, M. (1994). *Practical Risk Theory for Actuaries*, Chapman & Hall, London.
- [4] Dickson, D.C.M. (2005). *Insurance Risk and Ruin*, Cambridge University Press, Cambridge.
- [5] Embrechts, P., Kaufmann, R. & Samorodnitsky, G. (2004). Ruin theory revisited, stochastic models for operational risk, in *Risk Management for Central Bank Foreign Reserves*, C. Bernadell, P. Cardon, J. Coche, F.X. Diebold, S. Manganelli, eds, European Central Bank, Frankfurt A.M., pp. 243–261.
- [6] Frachot, A., Moudoulaud, O. & Roncalli, T. (2004). Loss distribution approach in practice, in M. Ong, ed., *The Basel Handbook: A Guide for Financial Practitioners*, Risk Books, London.
- [7] Gerber, H.U. (1979). *An Introduction to Mathematical Risk Theory*, Huebner Foundation Monographs, Philadelphia.
- [8] Panjer, H.H. (1981). Recursive evaluation of a family of compound distributions, *ASTIN Bulletin* **12**, 22–26.
- [9] Rolski, T., Schmidli, H., Schmidt, V. & Teugels, J.L. (1999). *Stochastic Processes for Insurance and Finance*, Wiley, Chichester.
- [10] Schröter, K.J. (1991). On a family of counting distributions and recursions for related compound distributions, *Scandinavian Actuarial Journal* 161–175.
- [11] Sundt, B. (1992). On some extensions of Panjer's class of counting distributions, *ASTIN Bulletin* **22**, 61–80.
- [12] Sundt, B. & Jewell, W.S. (1981). Further results on recursive evaluation of compound distributions, *ASTIN Bulletin* **12**, 27–39.

Related Articles

Cramér–Lundberg Estimates; Heavy Tails in Insurance; Insurance Risk Models; Operational Risk; Poisson Process; Ruin Theory.

HANSPETER SCHMIDLI

Gerber–Shiu Function

Risk theory studies **insurance risk models** that describe the uncertainty associated with the claims recorded by an insurance company for the losses incurred by its policy holders. From premium and investment income, insurers set aside funds (surplus) to cover such losses. Ruin theory studies the fluctuations of these surplus processes. Classical problems include ruin (low) and dividend (high) barrier hitting times (see [5]).

In the last decade, the expected discounted penalty function, proposed by Gerber and Shiu, has unified the treatment of the joint distribution of the time to ruin, the surplus just prior to ruin, and the deficit at ruin. This article centers on this expected discounted penalty function, commonly called the *Gerber–Shiu (G–S) function* in the actuarial literature.

The G–S function is somewhat akin to an expected discounted payoff function for financial instruments. A brief description of its general features is given here, together with references that discuss details, generalizations, and applications to insurance and finance.

Introduction

Risk theory models describe the uncertainty associated with the claims recorded by an insurance company. The aggregate claims up to time t ,

$$S(t) = X_1 + \cdots + X_{N(t)} = \sum_{i=1}^{N(t)} X_i, \quad t \geq 0 \quad (1)$$

(with $S(t) = 0$ if $N(t) = 0$) forms a random process of interest in **Insurance Risk Model**. It combines the two sources of uncertainty about the portfolio—the claim frequency and severity. The collection $\{S(t)\}_{t \geq 0}$ is called the *aggregate claims process*, see [9, Chapters 6, 9].

Another important insurance process is the accumulated surplus at t :

$$U(t) = u + ct - S(t) = u + ct - \sum_{i=1}^{N(t)} X_i, \quad t > 0 \quad (2)$$

where $U(0) = u \geq 0$ is the initial surplus at time 0. Here c denotes the aggregate premium amount

charged per unit time by the company to its portfolio policy holders, and $S(t)$ is given by equation (1).

Ruin Theory

The central focus of classical **Ruin Theory** is the time of ruin

$$T = \inf\{t > 0; U(t) < 0\} \quad (3)$$

when the surplus falls below zero for the first time. Depending on the aggregate claims model $S(t)$, if $c > \mathbb{E}[S(1)]$, the time of ruin T may be a defective random variable, with $\Psi(u) = \mathbb{P}\{T < \infty\} < 1$. This $\Psi(u)$ is called the *probability of ultimate ruin*, whereas the (possibly defective) distribution function of T , $\Psi(u; t) = \mathbb{P}\{T < t\}$, is called the *finite-time ruin probability* [1].

The expected discounted penalty function, introduced by Gerber and Shiu in [6], captures the joint distribution of three random variables related to ruin; the time of ruin T , the surplus immediately prior to ruin $U(T^-)$, a **Solvency** early warning, and the severity of the deficit at ruin $|U(T)|$, a recovery test measure. Also called the *Gerber–Shiu (G–S) function*, the expected discounted penalty function is defined as

$$\begin{aligned} \Phi_\delta(u) &= \mathbb{E} \left[e^{-\delta T} w \left(U(T^-), |U(T)| \right) I(T < \infty) \mid U(0) = u \right] \end{aligned} \quad (4)$$

where δ is the discount rate and w is a known penalty function, representing the cost to the insurer, at the time of ruin, of a surplus $U(T^-)$ prior to ruin and of a deficit $|U(T)|$ at ruin. The indicator function $I(T < \infty)$ applies the penalty only if ruin occurs.

Table 1 lists some interesting special choices of the penalty function w .

The G–S function was first derived in [6] for the classical Poisson risk model, in which $S(t)$ in equation (1) is a compound Poisson $(\lambda; F_X)$ process. Gerber and Shiu show that Φ_δ satisfies a defective renewal equation

$$\begin{aligned} \Phi_\delta(u) &= \frac{\lambda}{c} \int_0^u \Phi_\delta(u-x) \int_x^\infty e^{-\rho(y-x)} dF_X(y) dx \\ &\quad + \frac{\lambda}{c} e^{\rho u} \int_u^\infty e^{-\rho x} \int_x^\infty w(x, y-x) dF_X(y) dx \end{aligned} \quad (5)$$

Table 1 G–S function $\Phi_\delta(u)$ for different penalty functions w

δ	$w(x, y)$	Special case of the G–S function $\Phi_\delta(u)$
$\delta > 0$	e^{-tx-sy}	Trivariate Laplace transform of T , $U(T^-)$ and $ U(T) $
$\delta > 0$	1	the Laplace transform of T
$\delta = 0$	$x^m y^n$	Joint moments of $U(T^-)$ and $ U(T) $
$\delta = 0$	$I(x + y \leq z)$	(defective) distribution of ruin-causing claim at z
$\delta = 0$	$I(x \leq z, y \leq z')$	joint distribution of $U(T^-)$ and $ U(T) $ at (z, z')
$\delta = 0$	1	The probability of ultimate ruin $\Psi(u)$

where ρ is the unique nonnegative solution of Lundberg’s equation

$$\lambda + \delta - c\rho = \lambda \hat{F}_X(\rho) \quad (6)$$

and $\hat{F}_X(s) = \int_0^\infty e^{-sx} dF_X(x)$ is the Laplace–Stieltjes transform of F_X .

The solution of equation (5) can be expressed as an infinite series of convolutions $\Phi_\delta(u) = h * \sum_k g^{*k}(u)$, for given auxiliary functions g and h . This result extends to more general risk processes, including the Sparre Andersen (compound renewal) model [10], discounting of claims [2, 3], Lévy risk processes [4], or the Markov-modulated risk model [11].

Applications to Finance

The G–S function has been used also to analyze some financial instruments, such as several types of perpetual American options [7, 8]. As pointed out in these references, the pricing of perpetual options is related to the valuation of equity indexed annuities (EIA), hence the G–S function can help in analyzing the financial guarantee embedded in an EIA.

Current research work on G–S functions and extensions to financial instruments includes the valuation of contingent claims and the study of the optimal stopping time to exercise an option.

References

- [1] Asmussen, S. (2000). *Ruin Probabilities*, World Scientific, Singapore.
- [2] Cai, J. & Dickson, D.C.M. (2002). On the expected discounted penalty function at ruin of a surplus process with interest, *Insurance Mathematics and Economics* **30**, 389–404.
- [3] Cai, J., Feng, R. & Willmot, G.E. (2007). The compound Poisson surplus model with interest and liquid reserves: analysis of the Gerber–Shiu discounted penalty function, *Methodology and Computing in Applied Probability* in press.
- [4] Garrido, J. & Morales, M. (2006). On the expected discounted penalty function for Levy risk processes, *North American Actuarial Journal* **10**(4), 196–218.
- [5] Gerber, H.U. (1979). *An Introduction to Mathematical Risk Theory*, S.S. Huebner Foundation for Insurance, University of Pennsylvania.
- [6] Gerber, H.U. & Shiu, E.S.W. (1998). On the time value of ruin, *North American Actuarial Journal* **2**(1), 48–78.
- [7] Gerber, H.U. & Shiu, E.S.W. (1998). Pricing perpetual options for jump processes, *North American Actuarial Journal* **2**(3), 101–112.
- [8] Gerber, H.U. & Shiu, E.S.W. (2003). Pricing perpetual fund protection with withdrawal option, *North American Actuarial Journal* **7**(2), 1–33.
- [9] Klugman, S.A., Panjer, H.H. & Willmot, G.E. (2008). *Loss Models: From Data to Decisions*, 3rd Edition, John Wiley & Sons, New York.
- [10] Li, S. & Garrido, J. (2005). On a general class of renewal risk process: analysis of the Gerber–Shiu function, *Advances in Applied Probability* **37**, 836–8856.
- [11] Lu, Y. & Tsai, C.C.L. (2007). The expected discounted penalty at ruin for a Markov–modulated risk process perturbed by diffusion, *North American Actuarial Journal* **11**(2), 136–152.

ALEJANDRO BALBAS & JOSE GARRIDO

Ruin Models with Investment Income

The problem of ruin has a long history in risk theory, going back to Lundberg [13]. Initially the assumption was that companies do not earn any investment return on their capital. The first attempt to incorporate investment incomes was undertaken by Segerdahl [19]. Segerdahl's assumption was that capital earns interest at a fixed rate r . This model was further elaborated in [8, 11], and it remains very popular even today. Then inspired by ideas from mathematical finance, in [14] a model was suggested where capital is allowed to be invested in risky assets. Before presenting this model, we refer the reader to [16] for a more extensive survey, including several topics not covered here.

To make the ideas transparent, we introduce the risk process by means of two basic processes, that is,

- a basic risk process P with $P_0 = 0$ and
- a return on investment-generating process R with $R_0 = 0$.

If P and R belong to the rather general class of semimartingales, we can define the risk process as

$$Y_t = y + P_t + \int_0^t Y_{s-} dR_s \quad (1)$$

so that $Y_0 = y$. The solution of this equation is

$$Y_t = e^{\tilde{R}_t} \left(y + \int_0^t e^{-\tilde{R}_s} dP_s \right) \quad (2)$$

where $\tilde{R} = \log \mathcal{E}(R)$ is the logarithm of the Doléans–Dade exponential of R , that is, $V_t = \mathcal{E}(R)_t$ satisfies the equations $dV_t = V_{t-} dR_t$ and $V_0 = 1$.

In this article, it is assumed that P and R are of the following forms:

$$P_t = pt + \sigma_P W_{P,t} - \sum_{i=1}^{N_t} S_i \quad (3)$$

$$R_t = rt + \sigma_R W_{R,t} \quad (4)$$

where W_P and W_R are Brownian motions, N is a Poisson process with rate λ , and $\{S_i\}$ are nonnegative independent and identically distributed (i.i.d.) random

variables with distribution function F . Furthermore, W_P , W_R , N , and $\{S_i\}$ are all independent. The idea is that p is the premium rate, $\{S_i\}$ are claims, while W_P represents fluctuations in premium income and maybe also small claims. The return on investment-generating process R is the standard Black–Scholes return process. With these assumptions, Y becomes a homogeneous, strong Markov process, a fact that allows us to draw on the vast literature on Markov processes. In addition, $\tilde{R}_t = R_t - \frac{1}{2}\sigma_R^2 t = (r - \frac{1}{2}\sigma_R^2)t + \sigma_R W_{R,t}$.

The time of ruin is defined as $T = \inf\{t : Y_t < 0\}$, with $T = \infty$ if Y stays positive. The probability of ruin in finite *versus* infinite time is then defined as

$$\begin{aligned} \psi(t, y) &= P(T \leq t | Y_0 = y) \quad \text{and} \\ \psi(y) &= P(T < \infty | Y_0 = y) \end{aligned} \quad (5)$$

Mathematically, $\psi(y)$ is the easiest, and probably as a consequence, it has by far been the most popular in the literature.

A more general concept that also allows for problems relating to the time of ruin, the size of the deficit at ruin, or the surplus immediately before ruin is the Gerber–Shiu penalty function [9]. It is given as

$$\Phi_\alpha(y) = E[g(Y_{T-}, |Y_T|)e^{-\alpha T} 1_{\{T < \infty\}} | Y_0 = y] \quad (6)$$

where g is a nonnegative function and $\alpha \geq 0$. Various choices of g and α allow for the computation of many interesting quantities related to ruin [4].

Some General Results

It is shown in [15] that under weak assumptions, $\psi(y) < 1$ if $r > \frac{1}{2}\sigma_R^2$ (equivalence when $\sigma_R^2 > 0$), and in this case it is proved in [10] that under some weak additional assumptions, ψ is twice continuously differentiable and is a solution of the equation,

$$G\psi(y) = -\lambda \bar{F}(y) \quad (7)$$

with boundary conditions

$$\lim_{y \rightarrow \infty} \psi(y) = 0 \quad \text{and} \quad \psi(0) = 1 \quad \text{if} \quad \sigma_P > 0 \quad (8)$$

Here G is the integro-differential operator

$$Gh(y) = \frac{1}{2}(\sigma_P^2 + \sigma_R^2 y^2)h''(y) + (p + ry)h'(y) + \lambda \int_0^y h(y-x)dF(x) - \lambda h(y) \quad (9)$$

and $\bar{F}(y) = 1 - F(y)$. Sometimes it is more convenient to work with the survival probability $\phi(y) = 1 - \psi(y)$, in which case equation (2) becomes $G\phi(y) = 0$. It is shown in [4] that the Gerber–Shiu penalty function satisfies an equation similar to equation (2).

Following ideas from [11], it was shown in [14] that the ruin probability can be written as

$$\psi(y) = \frac{H(-y)}{E[H(-Y_T)|T < \infty]} \quad (10)$$

where H is the distribution function of the perpetuity

$$X = \int_0^\infty e^{-\left(r - \frac{1}{2}\sigma_R^2\right)t + \sigma_R W_{R,t}} dP_t \quad (11)$$

Corresponding to equation (2), the finite time ruin probability $\psi(t, y)$ should be the solution to the partial integro-differential equation

$$-\frac{\partial}{\partial t}\psi(t, y) + G\psi(t, y) = -\lambda\bar{F}(y) \quad (12)$$

with the additional boundary condition $\psi(0, y) = 1_{\{y < 0\}}$. Here, the operator G acts on the y variable.

Analytical and Numerical Solutions

Since most analytical results are for the infinite time horizon, we start with this case. When $\lambda = 0$, the process is a pure diffusion process and the problem becomes rather easy. A solution can be found in [14]. When $\lambda > 0$, things become markedly more complex, and analytical solutions can be obtained only in a few rather simple cases. In addition, several of these solutions are very complicated. In all known solutions, it is assumed that $\sigma_R = 0$, so in the sequel we shall therefore tacitly let $\sigma_R = 0$.

We have already mentioned Segerdahl's classical work when $\sigma_P = 0$ and claim sizes are exponentially distributed. In [17], these solutions were extended to the case where claims are mixtures of two exponential

distributions, as well as to the case where they are Erlang (2) distributed. Extensions beyond that seem very difficult though. In [17], the case with $\sigma_P > 0$ and claims exponentially distributed was also solved. In [5], rather explicit expressions for the Gerber–Shiu penalty function are given for the case where $\sigma_P = 0$ and claims are exponentially distributed.

In the finite time horizon case, analytical solutions are hard to come by. In [1], a recursion for the survival probability $\phi(t, y)$ when $\lambda = kr$ for some positive integer k is provided. This recursion is solved and exact solutions are given for $k = 1$ and $k = 2$.

The issue of computing numerical values has received comparatively little attention. In [18], using integration by parts on the equation $G\phi(y) = 0$, this equation was turned into a Volterra integral equation and methods from numerical analysis were used to solve this numerically. In the finite time case, several methods have been proposed when $\sigma_P = \sigma_R = 0$; see, for example [3, 6]. These methods are rather intuitive in nature and are not based on any particular known procedure from numerical analysis. Their efficiency is, therefore, slightly low, but as a bonus they provide upper and lower bounds.

An alternative to traditional numerical methods is the Monte Carlo simulation, which is particularly well suited for finite time ruin problems. For infinite time ruin problems, some care has to be taken as to when the simulation should stop. An alternative is to simulate under an equivalent measure $\tilde{P}$ so that $\tilde{P}(T < \infty | Y_0 = y) = 1$. Then with M_t equal to $\frac{d\tilde{P}}{dP}$ restricted to $\sigma\{Y_s : s \leq t\}$,

$$\psi(y) = \tilde{E}[M_T^{-1} | Y_0 = y] \quad (13)$$

Hence what is required is repetitive simulation of M_T under $\tilde{P}$. The catch is that this kind of importance sampling is very dangerous (due to the possible explosion of the variance of the Monte Carlo estimator) and can lead to completely wrong results.

Asymptotic Results

If there are only a few papers dealing with numerical solutions, there is certainly a fair number dealing with asymptotic results, that is, the behavior of $\psi(y)$ or $\psi(t, y)$ when y gets large. To present a few of the most prominent or recent results, we shall need some definitions. As before, let $\bar{F}(x) = 1 - F(x)$ and let $F^{*n}(x)$ be the n -fold convolution of F .

- $F \in \mathcal{R}_{-\alpha}$ if $\bar{F}(x) = x^{-\alpha}l(x)$ where l is a slowly varying function, that is, $\lim_{x \rightarrow \infty} \frac{l(tx)}{l(x)} = 1$ for all $t > 0$.
- $F \in \mathcal{S}$ if $\lim_{x \rightarrow \infty} \frac{\bar{F}^{*2}(x)}{\bar{F}(x)} = 2$.

We have $\mathcal{R}_{-\alpha} \subset \mathcal{S}$. The class $\mathcal{S}$ is called the *class of subexponential distributions*, and among others it contains the lognormal, loggamma as well as the Weibull distribution with $\bar{F}(x) = e^{-(x/\beta)^\gamma}$ for $\gamma < 1$. It can be shown that if $F \in \mathcal{S}$, then $E[e^{tS}] = \infty$ for all $t > 0$, and hence the name *subexponential distribution*.

Culminating through a series of papers, it was proved in [12] that when $\sigma_R = 0$ and $F \in \mathcal{S}$,

$$\psi(t, y) \sim \frac{\lambda}{r} \int_y^{ye^{rt}} \frac{\bar{F}(x)}{x} dx, \quad 0 < t \leq \infty \quad (14)$$

For the light-tailed case with $\sigma_R = 0$, results are a little less explicit. Let $\kappa = \sup\{a : E[e^{aS}] < \infty\}$. Then it was proved in [7] that for any $\varepsilon > 0$,

$$\begin{aligned} \lim_{y \rightarrow \infty} y^{(\kappa-\varepsilon)y} \psi(y) &= 0 \quad \text{and} \\ \lim_{y \rightarrow \infty} y^{(\kappa+\varepsilon)y} \psi(y) &= \infty \end{aligned} \quad (15)$$

It was also shown by examples that anything can happen to $\lim_{y \rightarrow \infty} y^\kappa \psi(y)$.

When $\sigma_R > 0$ the picture is slightly less complex. The reason for this is that the financial risk caused by variations in the return process R corresponds to claims $F \in \mathcal{R}_{-\kappa}$ with $\kappa = \frac{2\kappa}{\sigma_R^2} - 1$. Hence, when claims have lighter tails than this, the asymptotics is dominated by R . Various results in this direction appear in the literature and the most precise results for the model studied here are those in [10]. There it is assumed that $\sigma_P = 0$, but due to the light-tailed effect of W_P , the result is valid for $\sigma_P > 0$ as well. To present the results, remember that $\psi(y) = 1$ when $\kappa \leq 0$, so assume that $\kappa > 0$ and also that $E[S] < \infty$.

1. If for some $\varepsilon > 0$, $E[S^{\kappa+\varepsilon}] < \infty$, then $\lim_{y \rightarrow \infty} y^\kappa \psi(y) = c$ for some (for all practical purposes) unknown c .
2. If $\bar{F}(x) = x^{-\rho}l(x)$ where $\rho < \kappa$, then

$$\psi(y) \sim \frac{2\lambda}{\sigma_R^2 \rho (\kappa - \rho)} y^{-\rho} l(y) \quad (16)$$

The limiting case $F \in \mathcal{R}_{-\kappa}$ needs to be treated separately [10].

Inequalities

Several inequalities for the ruin probability exist, but the problem is that either they are not very sharp or they are difficult to compute. An example that follows directly from equation (3) is

$$H(-y) \leq \psi(y) \leq \frac{H(-y)}{H(0)} \quad (17)$$

and when F has decreasing failure rate, the right-hand side can be strengthened to $\frac{H(-y)}{E[H(S)]}$. The problem here is that the distribution function H is not known.

An upper bound that is easier to compute, but restricted to the case with $\sigma_P = \sigma_R = 0$ and light-tailed claims, is given in [2]. To explain, assume that the equation

$$\lambda \left(\int_0^\infty e^{\gamma(s)x} dF(x) - 1 \right) - \gamma(s)(p + rs) = 0 \quad (18)$$

has a positive solution $\gamma^*(s)$ for all $0 \leq s \leq y$. Then

$$\psi(y) \leq e^{-\int_0^y \gamma^*(s) ds} \quad (19)$$

References

- [1] Albrecher, H., Teugels, J.L. & Tichy, R.F. (2001). On gamma series expansion for the time-dependent probability of collective ruin, *Insurance, Mathematics and Economics* **29**, 345–355.
- [2] Asmussen, S. & Nielsen, H.M. (1995). Ruin probabilities via local adjustment coefficients, *Journal of Applied Probability* **32**, 736–755.
- [3] Brekelmans, R. & De Waegenaere, A. (2001). Approximating the finite-time ruin probability under interest force, *Insurance, Mathematics and Economics* **29**, 217–229.
- [4] Cai, J. (2004). Ruin probabilities and penalty functions with stochastic rates of interest, *Stochastic Processes and their Applications* **112**, 53–78.
- [5] Cai, J. (2007). On the time value of absolute ruin with debit interest, *Advances in Applied Probability* **39**, 343–359.
- [6] Cardoso, R.M.R. & Waters, H.R. (2003). Recursive calculation of finite time ruin probabilities under interest force, *Insurance, Mathematics and Economics* **33**, 659–676.

- [7] Embrechts, P. & Schmidli, H. (1994). Ruin estimation for a general insurance risk model, *Advances in Applied Probability* **26**, 404–422.
- [8] Gerber, H.U. (1971). Der Einfluss von Zins auf die Ruinwahrscheinlichkeit, *Mitteilungen der Schweizerischen Vereinigung der Versicherungsmathematiker* **71**, 63–70.
- [9] Gerber, H.U. & Shiu, E.S.W. (1998). On the time value of ruin, *North American Actuarial Journal* **2**, 48–78.
- [10] Grandits, P. (2004). A Karamata-type theorem and ruin probabilities for an insurer investing proportionally in the stock market, *Insurance, Mathematics and Economics* **34**, 297–305.
- [11] Harrison, J.M. (1977). Ruin problems with compounding assets, *Stochastic Processes and their Applications* **5**, 67–79.
- [12] Jiang, T. & Yan, H.-F. (2006). The finite-time ruin probability for the jump-diffusion model with constant interest force, *Acta Mathematicae Applicatae Sinica* **22**, 171–176.
- [13] Lundberg, F. (1903). *Approximerad Framställing av Sannolikhetsfunktioner II. Återforsäkring av Kollektivrisker*, Almqvist & Wiksell, Uppsala.
- [14] Paulsen, J. (1993). Risk theory in a stochastic economic environment, *Stochastic Processes and their Applications* **46**, 327–361.
- [15] Paulsen, J. (1998). Sharp conditions for certain ruin in a risk process with stochastic return on investments, *Stochastic Processes and their Applications* **75**, 135–148.
- [16] Paulsen, J. (2008). Ruin models with investment incomes, *Probability Surveys (Electronic)* **5**, 416–434.
- [17] Paulsen, J. & Gjessing, H.K. (1997). Ruin theory with stochastic return on investments, *Advances in Applied Probability* **29**, 965–985.
- [18] Paulsen, J., Kasozi, J. & Steigen, A. (2005). A numerical method to find the probability of ultimate ruin in the classical risk model with stochastic return on investments, *Insurance, Mathematics and Economics* **36**, 399–420.
- [19] Segerdahl, C.O. (1942). Über einige risikothoretische Fragestellungen, *Skandinavisk Aktuarietidskrift* **25**, 43–83.

Related Articles

Cramér–Lundberg Estimates; Merton Problem; Ruin Theory.

JOSTEIN PAULSEN

Phase-type Distribution

A phase-type distribution is a probability distribution that results from a system of one or more interrelated Poisson processes occurring in sequence or phases. Such distributions represent one of the most general classes of distributions permitting a probabilistic (i.e., Markovian) interpretation, as they are essentially based on the method of exponential stages technique that was introduced by Erlang in the early twentieth century and later generalized by Neuts in the mid-to-late 1970s. Phase-type distributions also possess a variety of desirable analytical properties that often facilitate their use in terms of yielding exact, algorithmically tractable, solution procedures. Phase-type distributions have been used extensively in many areas of applied probability, particularly in queueing theory (e.g., see [3]), actuarial risk theory (e.g., see [9]), and finance (e.g., see [4]). In addition, thorough treatments of phase-type distributions and their applications can be found in several reference texts including [1, 5, 7, 8, 10]. In what follows, we present only a brief overview of phase-type distributions and some of their key distributional properties.

In essence, phase-type distributions consist of a “general” mixture of exponentials and are characterized by an absorbing finite-state Markov chain. Specifically, let $\{X(t), t \geq 0\}$ represent a homogeneous, continuous-time Markov chain with transient states $\{1, 2, \dots, n\}$ and single absorbing state labeled $n + 1$. Define the row vector $\underline{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)$, wherein α_j represents the probability of beginning (at time 0) in a transient state j , $j = 1, 2, \dots, n$. The infinitesimal generator Q associated with this Markov process can be represented as

$$Q = \begin{bmatrix} S & \underline{s}_0 \\ \underline{0} & 0 \end{bmatrix} \quad (1)$$

In equation (1), S is an $n \times n$ nonsingular matrix containing the transition rates among the n transient states. The diagonal entries of S are necessarily negative, while the remaining entries of S are non-negative. Furthermore, $\underline{s}_0$ is the $n \times 1$ column vector of absorption rates into state $n + 1$ from the individual transient states. Necessarily, $\underline{s}_0 = -S\underline{e}$, where $\underline{e}$ denotes an $n \times 1$ column vector of ones. In addition, $\underline{0}$ represents a $1 \times n$ row vector of zeros.

Let Y represent the time to absorption into state $n + 1$, namely $Y = \inf\{t \geq 0 \mid X(t) = n + 1\}$. The random variable Y is said to have a continuous phase-type distribution with (parametric) representation $(\underline{\alpha}, S)$ of dimension n , often denoted as $Y \sim PH_n(\underline{\alpha}, S)$. The cumulative distribution function and density function of Y are given by

$$\begin{aligned} F(y) &= 1 - \underline{\alpha} \exp(Sy) \underline{e} \quad \text{and} \\ f(y) &= \underline{\alpha} \exp(Sy) \underline{s}_0 \quad \text{for } y \geq 0 \end{aligned} \quad (2)$$

where the matrix exponential in equation (2) is defined by $\exp(Sy) = \sum_{m=0}^{\infty} y^m S^m / m!$. Moreover, Y has Laplace–Stieltjes transform and integer moments given by

$$\begin{aligned} E\{e^{-tY}\} &= 1 - \underline{\alpha} \underline{e} + \underline{\alpha}(tI - S)^{-1} \underline{s}_0 \quad \text{and} \\ E\{Y^k\} &= (-1)^k k! \underline{\alpha} S^{-k} \underline{e} \end{aligned} \quad (3)$$

where I denotes an $n \times n$ identity matrix. Computation of the above matrix-based quantities is a routine task with the aid of most mathematical software packages in use today. A good reference for the calculation of matrix exponentials is [6].

One of the most important and useful features of the phase-type family of distributions is that the phase-type distribution is dense in the set of all positive-valued distributions, implying that it is (theoretically) possible to represent any positive-valued distribution with a “suitable” phase-type approximation. However, we emphasize the fact that the phase-type distribution is a light-tailed distribution (including combinations/mixtures of exponential and Erlang distributions as special cases), and thus it cannot be expected to fit a heavy-tailed distribution in the extreme tail. For an excellent treatment regarding phase-type fitting of arbitrary distributions, see [2] and references therein.

Numerous operations on phase-type distributions lead again to distributions that are phase-type with readily calculable parameters. This is a very appealing quality of phase-type distributions. In particular, convolutions, finite mixtures, and geometric compounds of phase-type distributions are again phase-type distributed. See [1, 7] for proofs of these results and further closure properties.

In conclusion, we state that a discrete analog to the continuous phase-type distribution exists by

considering the time to absorption in an equivalent discrete-time Markov chain. We direct interested readers to [5, 7] for further details on discrete phase-type distributions.

References

- [1] Asmussen, S. (2000). *Ruin Probabilities*, World Scientific, Singapore.
- [2] Asmussen, S. (2000). Matrix-analytic models and their analysis, *Scandinavian Journal of Statistics* **27**, 193–226.
- [3] Asmussen, S. (2003). *Applied Probability and Queues*, 2nd Edition, Springer-Verlag, New York.
- [4] Asmussen, S., Avram, F. & Pistorius, M. (2004). Russian and American put options under exponential phase-type Lévy models, *Stochastic Processes and their Applications* **109**, 79–111.
- [5] Latouche, G. & Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*, ASA SIAM, Philadelphia.
- [6] Moler, C. & Van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later, *SIAM Review* **45**, 3–49.
- [7] Neuts, M. (1981). *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*, Johns Hopkins University Press, Baltimore.
- [8] Neuts, M. (1989). *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York.
- [9] Rolski, T., Schmidli, H., Schmidt, V. & Teugels, J. (1999). *Stochastic Processes for Insurance and Finance*, John Wiley & Sons, Chichester.
- [10] Wolff, R. (1989). *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, NJ.

STEVE DREKIC

Credibility Theory

Credibility originates from the insurance industry and its basic idea goes back as far as the year 1918, when Whitney [20] was confronted with the problem of estimating risk premiums in workmen compensation. For the premium of an individual contract i , he suggested taking a weighted mean

$$\hat{\mu}_i = \alpha \bar{X}_i + (1 - \alpha) \bar{X} \quad (1)$$

where $\bar{X}_i$ is the observed average claim amount per volume unit for the individual contract i and $\bar{X}$ is the corresponding overall mean of the insurance portfolio. The expression *credibility* was originally used for the weight α in equation (1), which is a measure of reliability attached to the individual claims experience.

Since then, *credibility theory* has become an established discipline of insurance mathematics. Though mainly encountered in insurance applications, it also has the potential to be a powerful tool in the wider field of finance, where applications are found in credit risk (see **Credit Risk**) and in operational risk (see **Operational Risk**) assessments.

Credibility theory deals with the fundamental concepts of *individual risk* and *collective* or *individual claims experience* and *a priori knowledge*. The aim is to assess the individual risks. There are two sources of information on an individual risk: observations of the individual risk on the one hand, and *a priori* knowledge, gained from other similar risks in the portfolio (the collective) or from an expert opinion (personal belief), on the other.

Credibility theory can be described as follows:

- It answers the question of how one should combine the information out of the two sources to get a “best” estimate of the individual risk.
- It is the mathematical tool to describe heterogeneous collectives.
- It belongs mathematically to the area of Bayesian statistics.

A concise description of credibility theory and its development seen from the perspective of its essential mathematical structure can be found in [17]. References [5, 8] are some recent textbooks on credibility.

Credibility Estimators

In a simple set-up, the underlying paradigm is the “two-urn” model. The individual risk is characterized by its risk characteristic ϑ , which is drawn from the first urn describing the risk structure within the collective. Hence ϑ is a realization of a random variable Θ . Given $\Theta = \vartheta$, the claims are drawn from the second urn with distribution F_ϑ . The risk premium becomes a random variable $\mu(\Theta)$, which has to be estimated from the observed data $\mathbf{X}$. The objective is to find an estimator $\widehat{\mu(\Theta)}$, which minimizes the quadratic loss or mean squared error (mse)

$$\text{mse}(\widehat{\mu(\Theta)}) = E[(\widehat{\mu(\Theta)} - \mu(\Theta))^2] \quad (2)$$

The “two-urn” model formulates the problem in the language of Bayesian statistics. Essentially, this had already been done in the 1940s by Bailey [1], but it became well known only in the late 1960s, when it was set out clearly by Bühlmann in [2, 3].

The estimator minimizing equation (2) is the Bayes estimator

$$\mu(\Theta)^{\text{Bayes}} = E[\mu(\Theta) | \mathbf{X}] \quad (3)$$

In many models, it is a linear function of the observations. This is, in particular, the case for the one-parametric exponential family with its natural conjugate prior [13]. A famous example is the Poisson–Gamma model, where Θ is Gamma distributed and where the observations, the claim numbers, are Poisson distributed with parameter Θ . This model was the actuarial basis for the construction of bonus–malus systems in third-party motor liability. The case where the Bayes estimator is a linear function of the observations is referred to as *exact credibility*. As an example, we consider the situation where we want to design a bonus–malus scheme for fleets of cars. We assume that the *a priori* expected number of claims within a given period (e.g., 1 year) for a specific fleet i is equal to v_i , which is known and can be calculated on the basis of objective tariff variables. On the other hand, we know that things like driver education might have considerable influence and that the true expected number of claims for that particular fleet is $\mu(\Theta_i) = \Theta_i v_i$, where Θ_i is an unknown risk quality factor for that particular fleet with $E[\Theta_i] = 1$. We assume that we know from

portfolio experience of different fleets that the standard deviation of Θ_i is equal to τ . Now we have observed a number of N_i claims in the given period for fleet i . On the basis of the Poisson–Gamma model we then obtain the following bonus–malus factor by which the base premium is multiplied to obtain the experience-rated premium:

$$\widehat{\Theta}_i = 1 + \alpha_i \left(\frac{N_i}{v_i} - 1 \right), \text{ where } \alpha_i = \frac{v_i}{v_i + \tau^{-2}} \quad (4)$$

For instance, if $v_i = 20$, $N_i = 15$, and $\tau = 20\%$, then there results a bonus–malus factor of 89%.

In general, the Bayes estimator cannot be expressed in closed analytical form. Moreover, one has to specify the family $\{F_\vartheta\}$ of distributions as well as the structure distribution U , which are both mostly not known in practice.

The basic idea behind credibility is to *restrict* the class of allowable estimator functions to those which are *linear* in the components of the observation vector $\mathbf{X}$. *Credibility estimators* are therefore *linear Bayes estimators*, minimizing the mse within the class of linear estimators.

The so-called normal equations play a central role, that is, $\mu(\Theta)^{\text{cred}}$ is the credibility estimator of $\mu(\Theta)$ based on $\mathbf{X}^T = (X_1, \dots, X_n)$ if and only if it satisfies the normal equations

$$E[\mu(\Theta)^{\text{cred}} - \mu(\Theta)] = 0, \quad (5)$$

$$\begin{aligned} \text{Cov}[\mu(\Theta)^{\text{cred}}, X_j] &= \text{Cov}[\mu(\Theta), X_j] \\ \text{for } j &= 1, 2, \dots, n \end{aligned} \quad (6)$$

Hence the credibility estimator only depends on first and second moments and not on the whole underlying distributions.

The only mathematical structure in the most general credibility set-up is the fact that one wants to estimate a random variable $\mu(\Theta)$ based on a random vector $\mathbf{X}^T$. Neither $\mu(\Theta)$ nor the components of $\mathbf{X}^T$ are further specified, that is, the observations X_j can also be functions of some other original observations.

Credibility is most elegantly understood by looking at the space $\mathcal{L}^2$ of square integrable random variables. $\mathcal{L}^2$ together with the inner product $\langle X, Y \rangle := E[XY]$ is a Hilbert space. The corresponding norm of an X in $\mathcal{L}^2$ is $\|X\| = \sqrt{\langle X, X \rangle}$.

The advantage of looking at credibility in this way is that we can apply our intuitive understanding of the properties of linear vector spaces to the credibility problem. The mse of an estimator $\widehat{\mu}(\Theta)$ becomes $\|\widehat{\mu}(\Theta) - \mu(\Theta)\|^2$ and the credibility estimator is the point in the subspace $L(\mathbf{X}, 1) := \{\widehat{\mu}(\Theta) : \widehat{\mu}(\Theta) = a_0 + \sum_{j=1}^n a_j X_j, \quad a_0, a_1, \dots \in \mathbb{R}\}$ closest to $\mu(\Theta)$. Hence $\mu(\Theta)^{\text{cred}}$ is the orthogonal projection of $\mu(\Theta)$ on $L(\mathbf{X}, 1)$, and the normal equations (5) and (6) follow from the orthogonality property. For a historical perspective of the Hilbert space approach in the context of credibility, we refer to [9].

The Bühlmann–Straub Model

The Bühlmann–Straub (BS) [7] model is the best known credibility model that is widely used in practice. It extends the simple Bühlmann model [2] by allowing for observations with associated weights, which are very often encountered in practice such as when considering claim averages, claim frequencies, loss ratios, and so on.

The risks are embedded in a collective of similar risks numbered $i = 1, 2, \dots, I$, each of which is characterized by its own Θ_i and has observations $\{X_{ij} : j = 1, 2, \dots, n_i\}$ with associated weights w_{ij} . The BS model assumes that, conditionally on Θ_i , the random variables $\{X_{ij} : j = 1, 2, \dots, n_i\}$ are independent with

$$E[X_{ij} | \Theta_i] = \mu(\Theta_i) \quad (7)$$

$$\text{Var}[X_{ij} | \Theta_i] = \frac{\sigma^2(\Theta_i)}{w_{ij}} \quad (8)$$

that the pairs $(\Theta_1, \mathbf{X}_1), (\Theta_2, \mathbf{X}_2), \dots$ are independent and that $\Theta_1, \Theta_2, \dots$ are independent and identically distributed (i.i.d.).

A heterogeneous portfolio is modeled in this manner. The risks in the portfolio are different: they have different risk profiles Θ_i . But the risks in the portfolio also have something in common: *a priori*, they cannot be recognized as being different, which results from the fact that the risk profiles Θ_i are i.i.d.

The credibility estimator turns out to be a weighted mean

$$\mu(\Theta_i)^{\text{cred}} = \alpha_i \bar{X}_i + (1 - \alpha_i) \mu \quad (9)$$

where

$$\bar{X}_i = \sum_j \frac{w_{ij}}{w_{i\bullet}} X_{ij}, \quad w_{i\bullet} = \sum_j w_{ij}, \quad \alpha_i = \frac{w_{i\bullet}}{w_{i\bullet} + \frac{\sigma^2}{\tau^2}} \quad (10)$$

and where μ , σ^2 , and τ^2 are the so-called structural parameters

$$\mu = E[\mu(\Theta_i)], \quad \sigma^2 = E[\sigma^2(\Theta_i)], \quad \tau^2 = \text{Var}[\mu(\Theta_i)] \quad (11)$$

The mse is

$$\text{mse}(\mu(\Theta_i)^{\text{cred}}) = (1 - \alpha_i) \tau^2 = \alpha_i \frac{\sigma^2}{w_{i\bullet}} \quad (12)$$

The credibility estimator (9) has a natural interpretation. The expected value μ over the whole collective is the best estimator based only on the *a priori* knowledge and ignoring the individual observations, whereas $\bar{X}_i$ is the best linear estimator when looking only at the individual observations and ignoring the *a priori* knowledge (optimally compressed data, linear sufficient statistics). The credibility estimator takes both sources of information into account and is a weighted mean of the two components with weights inversely proportional to the mse. This *general intuitive principle* is also valid in the more general models.

The assumption that the risks in the portfolio are *a priori* equal is rather strong and often not fulfilled in practice. Often, there are known *a priori* differences between risks, that is, instead of equation (7) we have $E[X_{ij} | \Theta_i] = a_i \mu(\Theta_i)$ with known constants a_i . However, by a linear transformation one can transform the problem to the BS model and still apply it.

The credibility estimator (9) does not depend on the observations of the other risks in the portfolio. However, the structural parameters are usually not known, in practice, and have to be estimated from the data of the collective. For the overall mean μ , one can restrict the class of estimators further by allowing only linear estimators without a constant term. The best estimator out of this class is named the *homogeneous credibility estimator*, $\mu(\Theta_i)_{\text{hom}}^{\text{cred}}$, and contains a built-in estimator for μ . In the BS model it is obtained by replacing μ in equation (9) by

$$\hat{\mu} = \sum_i \frac{\alpha_i}{\alpha_{\bullet}} \bar{X}_i, \quad \text{where } \alpha_{\bullet} = \sum_i \alpha_i \quad (13)$$

When replacing the structural parameters in equation (9) by their estimated values, one arrives at the *empirical credibility estimator*. It can be shown [16] that under rather general conditions, the empirical credibility estimator converges in probability to the credibility estimator if the size of collateral data increases.

Hierarchical Credibility

Hierarchical credibility is an extension of the BS model in the sense that it generalizes the “two-urn” model to a “many-urn model”. In such models, the risk characteristics of the individual risks are drawn from a collective urn whose risk characteristics are drawn from a further parent urn and so on. Such hierarchies appear naturally in many situations. For instance, in insurance pricing, one often encounters the situation that individual risks are classified according to their “tariff positions”, tariff positions are grouped together into “subgroups”, subgroups into “groups”, groups into “main groups”, which together make the total of a line of business. A hierarchical model has the advantage of leading to a well-founded, properly balanced distribution of the burden of claims, in particular, of large claims, among the individual risks.

The concept of hierarchical credibility was introduced by Jewell [14] and Taylor [19]. It is best understood when looking at credibility estimators as orthogonal projections on linear subspaces, which, as shown in [6], leads to a recursive calculation of the credibility estimators. First, a linear sufficient statistics is calculated bottom up, and then the credibility estimators are calculated top down, where the estimators on the different levels have the same structure as the ones in the BS model.

Multidimensional Credibility

The essence of *multidimensional credibility* is that one wants to estimate a real-valued vector $\boldsymbol{\mu}(\Theta) = (\mu_1(\Theta), \dots, \mu_p(\Theta))^T$ as, for instance, the claims load for “normal claims” and “big claims”, “claim frequency” and “claim size”, the claims load in different layers, and so on. The point is that one estimates the components of $\boldsymbol{\mu}(\Theta)$ *simultaneously* on the basis of the same observation vector $\mathbf{X}$. For instance, when estimating the claims load for big

claims, one also looks at the observed claims load for normal claims, and *vice versa*.

Multidimensional credibility was introduced by Jewell [12]. A convenient way is to consider an abstract setup, where $\mathbf{X}$ has the same dimension as $\boldsymbol{\mu}(\Theta)$ with $E[\mathbf{X}|\Theta] = \boldsymbol{\mu}(\Theta)$ and $\text{Cov}(\mathbf{X}, \mathbf{X}^T|\Theta) = \Sigma(\Theta)$. This vector $\mathbf{X}$ may be obtained by data compression from the raw data. The structure of the credibility estimator inherits its form from the simple model and becomes

$$\boldsymbol{\mu}(\Theta)^{\text{cred}} = A\mathbf{X} + (I - A)\boldsymbol{\mu} \quad (14)$$

where

$$A = T(T + S)^{-1} \quad (15)$$

is the credibility matrix, I is the identity matrix, and

$$\begin{aligned} \boldsymbol{\mu} &= E[\boldsymbol{\mu}(\Theta)], \quad S = E[\Sigma(\Theta)], \\ T &= \text{Cov}[\boldsymbol{\mu}(\Theta), \boldsymbol{\mu}(\Theta)^T] \end{aligned} \quad (16)$$

are the structural parameters.

Mostly, one is not primarily interested in the individual components of the vector $\boldsymbol{\mu}(\Theta)$, but rather in a linear combination

$$\mu(\Theta) := \sum_{k=1}^p a_k \mu_k(\Theta) = \mathbf{a}^T \boldsymbol{\mu}(\Theta) \quad (17)$$

For the mse of $\mu(\Theta)$ one obtains

$$\text{mse}(\mu(\Theta)) = \mathbf{a}^T V_2 \mathbf{a} \quad (18)$$

where

$$V_2 = (I - A)T = AS \quad (19)$$

is the *mse matrix*.

The optimal data compression to reduce the raw data to a vector $\mathbf{X}$ of the same dimension as $\boldsymbol{\mu}(\Theta)$ is discussed in [5] and can be expressed as an orthogonal projection on a translated subspace

$$\mathbf{B} = \text{Pro}(\boldsymbol{\mu}(\Theta) | L_e^{\text{ind}}(\mathbf{X})) \quad (20)$$

where

$$\begin{aligned} L_e^{\text{ind}}(\mathbf{X}) &= \left\{ \widehat{\boldsymbol{\mu}(\Theta)} : \widehat{\mu_k(\Theta)} = \sum_j a_j^{(k)} X_j \right. \\ &\quad \left. \text{for } k = 1, \dots, p; E[\widehat{\boldsymbol{\mu}(\Theta)} | \Theta] = \boldsymbol{\mu}(\Theta) \right\} \end{aligned} \quad (21)$$

Credibility in the Regression Case

The credibility regression model was introduced by Hachemeister [11] and motivated by the problem of forecasting average claim amounts for bodily injury claims in third-party auto liability for various states in the United States. In contrast to classical statistics, the regression parameters are assumed to be a random vector $\boldsymbol{\beta}(\Theta)$, that is, it is assumed that

$$E[\mathbf{X}|\Theta] = Y\boldsymbol{\beta}(\Theta) \quad (22)$$

$$\text{Cov}(\mathbf{X}, \mathbf{X}^T|\Theta) = \Sigma(\Theta) \quad (23)$$

where $\mathbf{X}$ is the observation vector of a particular risk, Y is the corresponding design matrix and $\boldsymbol{\beta}(\Theta)$ the regression vector. The credibility estimator of $\boldsymbol{\beta}(\Theta)$ is most elegantly found by making use of the multidimensional credibility result. The optimal data compression is

$$\mathbf{B} = (Y^T S^{-1} Y)^{-1} Y^T S^{-1} \mathbf{X} \quad (24)$$

and by applying equation (14) one obtains

$$\boldsymbol{\beta}(\Theta)^{\text{cred}} = A\mathbf{B} + (I - A)\boldsymbol{\beta} \quad (25)$$

where

$$\begin{aligned} A &= T \left(T + (Y^T S^{-1} Y)^{-1} \right)^{-1} \\ \boldsymbol{\beta} &= E[\boldsymbol{\beta}(\Theta)], \quad S = E[\Sigma(\Theta)], \\ T &= \text{Cov}[\boldsymbol{\beta}(\Theta), \boldsymbol{\beta}(\Theta)^T] \end{aligned} \quad (26)$$

In many practical situations,

$$\Sigma(\Theta) = \sigma^2(\Theta)W^{-1} \quad (27)$$

where W is a diagonal matrix with known weights w_j down the diagonal. Then

$$\mathbf{B} = (Y^T W Y)^{-1} Y^T W \mathbf{X} \quad (28)$$

$$A = T \left(T + \sigma^2 (Y^T W Y)^{-1} \right)^{-1},$$

$$\text{where } \sigma^2 = E[\sigma^2(\Theta)] \quad (29)$$

When Hachemeister applied the theory to his data on bodily injury (a case of simple regression), he

obtained some “rather strange” results for some of the states. The solution is found in [4]: one should work with an orthogonal design matrix. In the simple linear regression case, this means that one should choose the barycenter not at the origin but rather at the center of gravity of time. Then the credibility estimators for the two regression components “barycenter” and “slope” split into two one-dimensional credibility formulas of the same type as in the BS-model.

Evolutionary Credibility Models

Evolutionary credibility models are characterized by the fact that the risk parameters vary stochastically in time. Such models are closely tied to recursive credibility formulae often referred to as *recursive credibility*.

References [10, 18] are some of the early actuarial papers devoted to *evolutionary or recursive credibility*. A well-known recursive procedure developed in an engineering context is the Kalman Filter [15] (see also **Filtering**).

In the *one-dimensional evolutionary credibility model*, a specific risk with observation vector $\mathbf{X}$ is again characterized by its individual risk profile. However, this risk profile is itself a stochastic process, that is, a sequence of random variables

$$\Theta = \{\Theta_1, \Theta_2, \dots, \Theta_j, \dots\} \quad (30)$$

The components Θ_j are to be understood as the risk profile in year j and it is assumed that

$$E[X_j | \Theta] = E[X_j | \Theta_j] = \mu(\Theta_j) \quad (31)$$

$$\text{Var}[X_j | \Theta] = \text{Var}[X_j | \Theta_j] = \frac{\sigma^2(\Theta_j)}{w_j} \quad (32)$$

where X_j is the observation in year j , w_j an appropriate weight and $E[\sigma^2(\Theta_j)] = \sigma^2$. The aim is to find for each year n the credibility estimator of $\mu(\Theta_n)$ based on the observations $\{X_1, \dots, X_{n-1}\}$.

It is useful to split up the step from $\mu(\Theta_n)^{\text{cred}}$ to $\mu(\Theta_{n+1})^{\text{cred}}$ into two steps:

1. Updating

Improve the estimation of $\mu(\Theta_n)$ on the basis of the newest information X_n .

2. Parameter movement

Change the estimation owing to the switch of the parameter from $\mu(\Theta_n)$ to $\mu(\Theta_{n+1})$.

In this connection, it is suitable to introduce the terminology and notation from state space models:

$$\begin{aligned} \mu_{n|n-1} &:= \text{Pro}(\mu(\Theta_n) | L(1, X_1, \dots, X_{n-1})), \\ \mu_{n|n} &:= \text{Pro}(\mu(\Theta_n) | L(1, X_1, \dots, X_{n-1}, X_n)), \\ q_{n|n} &:= E[(\mu_{n|n} - \mu(\Theta_n))^2], \\ q_{n|n-1} &:= E[(\mu_{n|n-1} - \mu(\Theta_n))^2] \end{aligned} \quad (33)$$

Here $\mu_{n|n-1}$ is just another notation for $\mu(\Theta_n)^{\text{cred}}$, which is the credibility estimator of $\mu(\Theta_n)$ based on the observations $X_1, \dots, X_{n-1}$, whereas $\mu_{n|n}$ is the updated estimator of $\mu(\Theta_n)$ based on the observations $X_1, \dots, X_{n-1}, X_n$.

If the observations in different years are, on the average, conditionally uncorrelated, and if the process $\mu(\Theta)$ has orthogonal increments, that is, if

$$E[\text{Cov}(X_k, X_j | \Theta)] = 0 \quad (34)$$

$$\begin{aligned} \text{Pro}(\mu(\Theta_{n+1}) | L(1, \mu(\Theta_1), \dots, \mu(\Theta_n))) \\ = \mu(\Theta_n) \quad \text{for } n \geq 1 \end{aligned} \quad (35)$$

then the credibility estimators can be calculated by the following recursive procedure (“Kalman filter notation”):

1. Anchoring (Period 1)

$$\begin{aligned} \mu_{1|0} &= E[\mu(\Theta_1)] \quad \text{and} \\ q_{1|0} &= E[(\mu_{1|0} - \mu(\Theta_1))^2] \end{aligned} \quad (36)$$

2. Recursion ($n \geq 1$)

(a) Updating

$$\mu_{n|n} = \alpha_n X_n + (1 - \alpha_n) \mu_{n|n-1} \quad (37)$$

$$\text{where } \alpha_n = \frac{q_{n|n-1}}{q_{n|n-1} + \frac{\sigma^2}{w_n}},$$

$$q_{n|n} = (1 - \alpha_n) q_{n|n-1} \quad (38)$$

(b) Change from $\mu(\Theta_n)$ to $\mu(\Theta_{n+1})$ (parameter movement)

$$\mu_{n+1|n} = \mu_{n|n} \quad (39)$$

$$q_{n+1|n} = q_{n|n} + \delta_{n+1}^2 \quad (40)$$

where $\delta_{n+1}^2 = \text{Var}[\mu(\Theta_{n+1}) - \mu(\Theta_n)]$.

The collection of models can be enlarged by allowing linear transformation on the right-hand side of equation (35) and adjusting equations (39) and (40) accordingly.

The evolutionary model sketched here can easily be extended to the *evolutionary regression credibility model*. By an appropriate choice of the components of $\beta(\Theta_n)$, the evolutionary regression model allows the modeling of a large number of evolutionary phenomena. In particular, one can model in this way all situations where the process $\{\mu(\Theta_j) : j = 1, 2, \dots\}$ is an autoregressive moving average (ARMA) process (see **Autoregressive Moving Average (ARMA) Processes**). Analogously, one can extend the idea to the multidimensional case and consider the *evolutionary multidimensional credibility model*. It allows to model situations where the individual risk over time depends both on specific individual factors as well as on global factors affecting all risks in the collective simultaneously (see [5] for further details). Such models might, besides insurance, also be suitable for applications in finance.

References

- [1] Bailey, A.L. (1950). Credibility procedures, Laplace's generalization of Bayes' Rule, and the combination of collateral knowledge with observed data, *Proceedings of the Casualty Actuarial Society* **37**, 7–23.
- [2] Bühlmann, H. (1967). Experience rating and credibility, *ASTIN Bulletin* **4**, 199–207.
- [3] Bühlmann, H. (1969). Experience rating and credibility, *ASTIN Bulletin* **5**, 157–165.
- [4] Bühlmann, H. & Gisler, A. (1997). Credibility in the regression case revisited, *ASTIN Bulletin* **27**, 83–98.
- [5] Bühlmann, H. & Gisler, A. (2005). *A Course in Credibility Theory and its Applications*, Universitext, Springer, Berlin.
- [6] Bühlmann, H. & Jewell, W.S. (1987). Hierarchical credibility revisited, *Bulletin of the Swiss Association of Actuaries* **87**, 35–54.
- [7] Bühlmann, H. & Straub, E. (1970). Glaubwürdigkeit für Schadensätze, *Bulletin of the Swiss Association of Actuaries* **76**, 111–133.
- [8] Dannenburg, D.R., Kaas, R. & Goovaerts, M.J. (1996). *Practical Actuarial Credibility Models*, Institute of Actuarial Science and Econometrics, Amsterdam.
- [9] De Vylder, F. (1976). Geometrical credibility, *Scandinavian Actuarial Journal* **76**, 121–149.
- [10] Gerber, H.U. & Jones, D.A. (1975). Credibility formulas of the updating type, *Transaction of the Society of Actuaries* **27**, 39–52.
- [11] Hachemeister, C.A. (1975). Credibility for regression models with application to trend, in *Credibility: Theory and Applications*, P.M. Kahn, ed., Academic Press, New York, pp. 129–163.
- [12] Jewell, W.S. (1973). *Multidimensional Credibility*, Report ORC Berkeley, Operations Research Center, Berkeley.
- [13] Jewell, W.S. (1974). Credibility means are exact Bayesian for simple exponential families, *ASTIN Bulletin* **8**, 77–90.
- [14] Jewell, W.S. (1975). The use of collateral data in credibility theory: a hierarchical model, *Giornale dell'Istituto Italiano degli Attuari* **38**, 1–16.
- [15] Kalmann, R.E. (1960). A new approach to linear filtering and prediction problems, *Transactions of ASME-Journal of Basic Engineering* **82**, 35–45.
- [16] Norberg, R. (1980). Empirical Bayes credibility, *Scandinavian Actuarial Journal* **38**, 177–194.
- [17] Norberg, R. (2004). Credibility theory, in *Encyclopedia of Actuarial Science*, J.L. Teugels & B. Sundt, eds, Wiley, Chichester, UK, pp. 398–406.
- [18] Sundt, B. (1981). Recursive credibility estimation, *Scandinavian Actuarial Journal* **3**–22.
- [19] Taylor, G.C. (1979). Credibility analysis of a general hierarchical model, *Scandinavian Actuarial Journal* **1**–12.
- [20] Whitney, A.W. (1918). The theory of experience rating, *Proceedings of the Casualty Actuarial Society* **4**, 274–292.

Related Articles

Credit Rating; Credit Scoring.

ALOIS GISLER

Life Insurance

Finance and life insurance have a lot in common. Life insurance contracts specify an exchange of streams of payments between the insurance company and the contract holder. These payment streams may cover the lifetime of the contract holder. Therefore, time valuation of money is crucial for any measurement of payments due in the past as well as in the future. Life insurance companies never keep their money idle, and accumulation and distribution of capital gains are always a part of the insurance business. With respect to the future, appropriate discounting of contractual obligations improves the estimates of liabilities.

The subject matter of life insurance draws on a wide set of mathematical disciplines including probability theory, stochastic processes, statistics, economics, and mathematical finance. However, in addition, less positivistic areas such as accounting and law, in general, and insurance accounting, contractual law, and regulatory law, in particular, often play important roles when theory meets practice. The broad title covers a survey of the mathematics of life insurance rather than life insurance in general.

At the end of the day, life insurance is a betting game on life history events such as death, survival, and disability. Betting in this context is very different from betting on horses or football since the benefit is paid conditional on a life history event that has already caused financial distress of the betting party: one buys a coverage against death to protect one's dependants—typically spouse and children—financially against the loss of future income upon death; one buys life annuities to protect oneself financially against living long by stretching out capital holdings until an uncertain time of death; and one buys a coverage against disability to protect oneself and one's dependants financially against the loss of income upon disability. In this respect, insurance and pension saving are betting games that hedge the personal financial position against financially adverse events. However, this explains why the society encourages its members to play this betting game (partly in contrast to other betting games) under slightly regulated conditions: it has a stabilizing effect on the financial position of individuals and families and thereby on the society as a whole.

The insurance companies and pension funds are the counterparts of this betting game. This exposition

deals, first of all, with the technical toolbox they use in order to provide relevant, fair, and reliable bets. However, it also deals with the enlargement of the toolbox caused by the integration of mathematical finance as a part of life insurance mathematics. New paradigms lead to new answers to what the terms relevant, fair, and reliable mean at all. However, the enlightenment is not one way: some of the classical tools of life insurance are relevant to modern finance and we provide an example.

The section Modeling of Payment Streams presents the classical modeling framework for life history risk and the payment streams linked to it, first in the so-called survival model and second in a multistate model. The section Valuation of Payment Streams deals with the valuation of and fairness criteria for these payment streams. First, we evaluate on the basis of classical actuarial assumptions and then discuss how the financial market helps to relax some of these assumptions. Finally, we exemplify by credit risk modeling and valuation, how methods of classical life insurance can also contribute to a structured view on modern financial topics. The section Design of Payment Streams formalizes the so-called participation feature in typical practical arrangements that add a dimension to the risk sharing taking place in an insurance contract. We distinguish between the classical with-profit construction, where design and decision is primarily a matter for the insurance company, and the more modern unit-link, where design and decision, to a larger extent, is a matter for the policy holder. Finally, in the section Management of Payment Streams, we focus on these decision processes. Optimal decision making on both sides of the balance scheme is the essence of asset-liability management (ALM). We conclude by relating these decision processes to those in the classical investment–consumption problem of personal finance.

Modeling of Payment Streams

The Survival Model Payment Streams

The payments of an insurance contract can be formalized by a payment process B , where $B(t)$ represents the accumulated payments from the insurance company to the policy holder until time t . The payments are then the increments $(dB(t))_{t \geq 0}$ and we have that

$$B(t) = \int_{0-}^t dB(s) \quad (1)$$

where, by definition, $\int_0^t = \int_{(0,t]}$ and $\int_{0-}^t = \int_{[0,t]}$. Thus, payments going from the insurance company to the policy holder (benefits) appear in B as positive increments, whereas payments going from the policy holder to the insurance company (premiums/contributions) appear in B as negative increments. In the formation of B , we have assumed that there are no payments before inception of the contract at time 0, that is, $B(0-) = 0$. We specify the payments in a continuous-time framework. We denote by T the time of death of a policy holder. To formalize the connection between payments and the life history of the insured, we introduce the indicator process $I(t)$. This process indicates whether the insured is alive in the sense that $I(t) = 1$ if the insured is alive at time t and $I(t) = 0$ if the insured is dead at time t . The process $Z(t) = 1 - I(t)$ indicates whether the insured is dead, which is illustrated in Figure 1.

We also introduce a counting process $(N(t))_{t \geq 0}$, counting the number of deaths of the insured (equals 0 or 1) over $[0, t]$. Note that $N = Z = 1 - I$. Finally, we introduce the notation $(\varepsilon^u(t))_{t \geq 0}$ for the deterministic indicator process, indicating that t exceeds u , that is, $\varepsilon^u(t) = 1 [t \geq u]$. Fixing a time horizon n for the insurance contract and introducing a retirement time $m \leq n$, most insurance payment processes in a survival model are given by a payment process B in the following form:

$$dB(t) = \Delta B^0(t)d\varepsilon^0(t) + b^0(t)I(t)dt + b^{01}(t) \times dN(t) + \Delta B^0(t)I(t)d\varepsilon^m(t), \quad t \leq n \quad (2)$$

Here, $\Delta B^0(0)$ is a possible single premium (negative) triggered at time 0, $b^0(t)$ is an annuity benefit rate (positive) or a level premium rate (negative) at time t , both conditional on survival by multiplication of I , $b^{01}(t)$ is a death benefit triggered upon death before time n , and $\Delta B^0(m)$ is a pension sum benefit triggered at time m upon survival to time m . In Table 1, we specify the elements in some standard forms of benefit and premium payment processes.

To have an idea about the future prospects of B , we need to specify the distribution of T . We form the

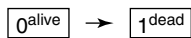


Figure 1 Survival model

survival function $\bar{F}$, the lifetime distribution function F , and the death density, respectively,

$$\begin{aligned} \bar{F}(t) &= P(T > t) = P(I(t) = 1) \\ F(t) &= P(T \leq t) = P(I(t) = 0) = 1 - \bar{F}(t) \\ f(t) &= \frac{d}{dt}F(t) \end{aligned} \quad (3)$$

The distribution of T is typically parameterized by the hazard rate conventionally denoted by μ (Greek m for mortality rate) and defined as the death rate conditional on survival,

$$\begin{aligned} \mu(t)dt &= P(t \leq T < t + dt | T > t) \\ &= \frac{f(t)dt}{\bar{F}(t)} = -\frac{\bar{F}'(t)dt}{\bar{F}(t)} \end{aligned} \quad (4)$$

This forms a differential equation for $\bar{F}$ with initial condition $\bar{F}(0) = 1$. Its solution $\bar{F}$ and the resulting F and f are given as

$$\begin{aligned} \bar{F}(t) &= e^{-\int_0^t \mu(s)ds}, \quad F(t) = 1 - e^{-\int_0^t \mu(s)ds}, \\ f(t) &= e^{-\int_0^t \mu(s)ds} \mu(t) \end{aligned} \quad (5)$$

thereby forming the elementary distribution characteristics in terms of the mortality rate. One of the advantages from representing these characteristics in terms of the mortality rate is that the corresponding quantities conditional on survival until a given age, say x , have the same structure with $\mu(t)$ replaced by $\mu(x+t)$, for example,

$$P(T > x + t | T > x) = e^{-\int_0^t \mu(x+s)ds} \quad (6)$$

This is useful when the insurance company sells a contract to an x -year old and takes into account the fact that he/she survived so far. There are many candidates for the mortality rate that can take any form as long as $\mu(t) \geq 0$ and it integrates to infinity over $[0, \infty)$. The Gompertz–Makeham mortality rate

$$\mu(t) = \alpha + \beta e^{\gamma t} \quad (7)$$

allows for a nonage-dependent effect α (death from accident) and an exponential age-dependent effect (death from deterioration of health over time). This is classic due to its good fit to data relative to its low number of parameters. In the Danish mortality law

Table 1 Elementary life insurances

	$\Delta B^0(0)$	$b^0(t)$	$b^{01}(t)$	$\Delta B^0(m)$
Benefits				
Pure endowment	0	0	0	1
Term insurance	0	0	$1[t \leq m]$	0
Endowment insurance	0	0	$1[t \leq m]$	1
Temporary life annuity	0	$1[t \leq n]$		
Deferred temporary life annuity	0	$1[m < t \leq n]$		
Premiums				
Single premium	-1			
Level premium		$-1[t \leq m]$		

G82 for Danish males, the parameters are $\alpha = 5 \times 10^{-4}$, $\beta = 7.5858 \times 10^{-5}$, and $\gamma = \log(1.09144)$.

The Multistate Model Payment Streams

Although the payment process in equation (2) formalizes a number of standard forms of benefits and premium payment processes, there is a number of contracts that are not covered. An example is a contract where the level premium is not paid during periods of disability, that is, the level premium is contingent on the state of health. Such so-called premium waiver conditions and different types of disability annuities can be covered by extending the survival model with a third state, “disabled” (Figure 2).

Another example is a widow’s/widower’s pension where the two spouses in a married couple are insured such that, for example, an annuity rate is payable in the time period where one spouse survives another. Insurances on two lives can be covered by combining two survival models (Figure 3). It is now easy to

imagine a combination of Figures 2 and 3 into a dependent lives model with disability.

In general, we let Z be a process moving around in a finite number of states numbered from 0 to J . For such a process, we introduce two vectors of processes: the indicator process I^j , $j \in \{0, \dots, J\}$ indicates sojourn in state j and the counting process N^j , $j \in \{0, \dots, J\}$ counts the number of jumps into state j . Note that, in contrast to the survival model, the counting processes are in general not necessarily equal to 0 or 1: for example, in a disability model, the policy holder can move back and forth between “disabled” and “active”. Formally,

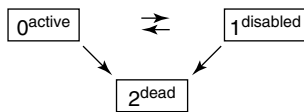
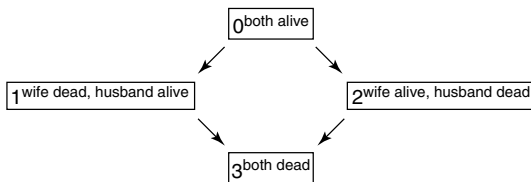
$$I^j(t) = 1[Z(t) = j]$$

$$N^j(t) = \#\{s \leq t : Z(s-) \neq j, Z(s) = j\} \quad (8)$$

We can now formalize a set of general payment processes driven by a multistate model (by convention, $Z(0) = 0$),

$$\begin{aligned} dB(t) = & \Delta B^0(t) d\epsilon^0(t) + \sum_j I^j(t) b^j(t) dt \\ & + \sum_k \sum_{j:j \neq k} I^j(t-) b^{jk}(t) dN^k(t) \\ & + \sum_j I^j(t) \Delta B^j(t) d\epsilon^m(t), t \leq n \quad (9) \end{aligned}$$

Here, $\Delta B^0(0)$ is a possible single premium (negative) triggered at time 0, $b^j(t)$ is a payment rate (positive or negative) at time t conditional on sojourn in state j , $\Delta B^j(t)$ is a lump sum payment at time t conditional on sojourn in state j , and $b^{jk}(t)$ is an insurance sum triggered and paid upon a transition from state j to state k . Note that, for example, in

**Figure 2** Disability model**Figure 3** Dependent lives model

the disability model, N^2 counts the number of deaths no matter from which state the death occurs; but by the specification of j in b^{j2} and multiplication by $I^j(t-)$ the death sum is allowed to depend on this information.

To have an idea about the future prospects of B , we need to specify the distribution of Z . The traditional approach is to assume the existence of deterministic piecewise continuous functions $\mu^{jk}(t)$ such that the distribution of N^k is characterized by its intensity $\mu^{Z(t-),k}(t)$. This means that the probability of experiencing a transition into state k over $[t, t + dt)$ is proportional to the time interval dt with proportionality factor $\mu^{Z(t-),k}(t)$. We allow this to depend on $Z(t-)$ such that, for example, in the disability model we can have a mortality intensity that depends on whether the policy holder is active or disabled. Given that these intensities exist, Z is a time-inhomogeneous multistate Markov process.

In more general models, one loosens up the Markov assumption (see **Markov Processes**) and works with, for example, semi-Markov models where the transition intensity into a given state does not only depend on where the policy holder is but also depends on how much time he/she has spent there. This is relevant in the disability model, where the mortality rate from the state “disabled” at a given age may depend on the duration of disability. Even if one sticks to a Markov model for Z , the duration of sojourn may still be needed as a state process to formalize, for example, the annuity payments depending on the duration. The main example is the so-called qualification period that has to be spent as disabled before the disability annuity starts. Note carefully the difference between taking duration as a part of the probability model and taking it as a part of the state process formalizing the payment process. In the following section, we give more examples of relaxation of the Markov property of Z .

Valuation of Payment Streams

Classical Valuation Based on Three Assumptions

We assume that the payments are currently deposited on (or withdrawn from) an account that bears interest at rate $r(t)$ at time t . We can now arrive at the present value at time t of a payment process by adding up the values of all elements in the payment process. We get the present value at time t of the payment process B :

$$\int_{0-}^n e^{-\int_t^s r(\tau) d\tau} dB(s) \quad (10)$$

The value at time 0 of a payment process B is the net loss at time 0 that the insurance company faces by initiation of the contract. If the processes describing the policy holder’s life history and the interest rate were known at time 0, this net loss could be calculated at time 0. To avoid losses (and gains), one should ideally balance the elements in the payment process such that the net loss is 0:

$$\int_{0-}^n e^{-\int_0^t r(s) ds} dB(t) = 0 \quad (11)$$

However, the processes for the policy holder’s life history and the interest rate are not known at time 0. We consider these processes as stochastic processes defined on a filtered (see **Filtrations**) probability space $(\Omega, \mathcal{F}, \{\mathcal{F}(t)\}_{t \geq 0}, P)$ such that the present value at time 0 becomes a stochastic variable and its outcome is revealed over time. The question is how one should balance the elements of the insurance contract in that situation.

A particular situation arises if we work under the following condition.

Condition 1

1. The insurance company issues (or can issue) contracts to a “large” number v of individuals with identically distributed payment processes $(B^i)_{i=1, \dots, v}$.
2. For all $i, j, i \neq j$, B^i is independent of B^j .
3. The interest rate is deterministic.

Then the law of large numbers applies and provides that the average gain from a contract converges toward the expected gain from a single contract (skipping the i then) as the number of contracts tends to infinity, that is,

$$\begin{aligned} & \frac{1}{v} \sum_{i=1}^v \int_{0-}^n e^{-\int_0^t r(s) ds} dB^i(t) \\ & \xrightarrow{P} E \left[\int_{0-}^n e^{-\int_0^t r(s) ds} dB(t) \right] \end{aligned} \quad (12)$$

To avoid *systematic* gains (and losses), one should, therefore, balance the elements in the payment

process such that the expected net gain is equal to 0,

$$E \left[\int_{0-}^n e^{-\int_0^t r(\tau) d\tau} dB(t) \right] = 0 \quad (13)$$

This balance equation formalizes the *principle of equivalence* that is fundamental in classical life insurance mathematics.

If one of the three assumptions fails, the classical principle of equivalence fails as a balancing tool for the elements in the payment process: if the insurance company cannot issue a large number of contracts, it makes no sense to draw conclusions from the law of large numbers; if B^i and B^j are dependent for $i \neq j$, the law of large numbers does not apply; and if the interest rate is not deterministic, we cannot deduce independence between present values from independence between payment processes. It should be mentioned, however, that Condition 1 can be weakened such that it holds only in a certain asymptotic sense.

Actually, the insurance company is interested in the valuation of the future payments not only at time 0 but also at any point in time t during the term of contract. This is required to inform the policy holder about the financial status of his/her contract and to inform the financial market and the regulators about the financial status of the insurance company. To estimate the present value of future payments at time t , $\int_t^n e^{-\int_t^s r(\tau) d\tau} dB(s)$, the insurance company naturally takes into account all information it possesses at that point in time. If we work under Condition 1 and in the multistate Markov model, we can argue, by the law of large numbers, for the valuation formula:

$$E \left[\int_t^n e^{-\int_t^s r(\tau) d\tau} dB(s) \middle| Z(t) \right] = \sum_j I^j(t) V^j(t) \quad (14)$$

where

$$V^j(t) = E \left[\int_t^n e^{-\int_t^s r(\tau) d\tau} dB(s) \middle| Z(t) = j \right] \quad (15)$$

These so-called statewise reserves are fundamental quantities in classical life insurance mathematics. They are characterized by a system of ordinary differential equations and presented in [6] as an inevitable tool for life insurance mathematicians. Substituting the payment process (9) into the reserve for state j , we get the differential equation (*see Markov Processes*):

$$\begin{aligned} \frac{d}{dt} V^j(t) &= r(t) V^j(t) - b^j(t) \\ &\quad - \sum_{k:k \neq j} \mu^{jk}(t) (b^{jk}(t) + V^k(t) - V^j(t)) \end{aligned} \quad (16)$$

which holds outside $\{0, m, n\}$. Inside $\{0, m, n\}$, the conditions are

$$\begin{aligned} V^j(n) &= 0 \\ V^j(m-) &= \Delta B^j(m) + V^j(m) \\ V^j(0-) &= \Delta B^j(0) + V^j(0) \end{aligned} \quad (17)$$

Note that, by taking the process Z to start in state 0 by convention, we can formulate the equivalence principle in terms of the statewise reserve:

$$V^0(0-) = \Delta B^0(0) + V^0(0) = 0 \quad (18)$$

The special case of the differential equation (16) that holds for the survival model was discovered by Thiele in 1875. In addition, generalized versions like, for example, equation (16) are spoken of as Thiele's differential equation. Interestingly, Thiele was actually also the first to model the Brownian motion mathematically in connection with his study of time series [5, 19]. Classical valuation formulas of classical life insurance products can be found in the textbook [4]. See also [13] for an account of different notions of reserves.

Example 1 Let us calculate the equivalence level premium for an endowment insurance sold to an x -year old based on the assumptions above. Plugging in the elementary endowment insurance benefits and a level premium rate π (Greek p for premium), the equivalence principle reads

$$\begin{aligned} &E \left[\int_{0-}^m e^{-\int_0^t r(s) ds} (-\pi I(t) dt + dN(t) \right. \\ &\quad \left. + I(t) d\mathcal{E}^m(t) \right) \Big] \\ &= E \left[\int_{0-}^m e^{-\int_0^t r(s) ds} (-\pi I(t) dt + dN(t) \right. \\ &\quad \left. + e^{-\int_0^m r(s) ds} I(m) \right) \Big] = 0 \end{aligned} \quad (19)$$

where expectation is actually conditional expectation given that the policy holder is x years old, that is, $T > x$. This forms one equation (the principle of equivalence) with one unknown (π) and its solution is called the equivalence premium. By plugging in the characteristics of the lifetime distribution in equation (5) and collecting terms, we get

$$\begin{aligned} & \int_{0-}^m e^{-\int_0^t (r(s) + \mu(x+s)) ds} (-\pi + \mu(x+t)) dt + e^{-\int_0^m (r(s) + \mu(x+s)) ds} = 0 \\ \Rightarrow & \frac{\int_{0-}^m e^{-\int_0^t (r(s) + \mu(x+s)) ds} \mu(x+t) dt + e^{-\int_0^m (r(s) + \mu(x+s)) ds}}{\int_{0-}^m e^{-\int_0^t (r(s) + \mu(x+s)) ds} dt} = \pi \end{aligned} \quad (20)$$

Actuaries have developed a shorthand notation for expected present values of standard payment streams. An actuary could write the premium formula above in code:

$$\pi = \frac{A_{x:m}}{a_{x:m}} = \frac{A_{x:m}^1 + {}_mE_x}{a_{x:m}} \quad (21)$$

The reserve to set up at time t given that the policy holder is alive is then given by

$$\begin{aligned} V^0(t) &= E \left[\int_t^m e^{-\int_t^s r(\tau) d\tau} (-\pi I(s) ds + dN(s)) \right. \\ &\quad \left. + e^{-\int_t^m r(s) ds} I(m) \middle| I(t) = 1 \right] \\ &= \int_t^m e^{-\int_t^s (r(\tau) + \mu(x+\tau)) d\tau} (-\pi + \mu(x+s)) ds \\ &\quad + e^{-\int_t^m (r(s) + \mu(x+s)) ds} \end{aligned} \quad (22)$$

It is easy to verify that this reserve actually fulfills the special case of the Thiele differential equation,

$$\frac{d}{dt} V^0(t) = r(t) V^0(t) + \pi - \mu(t) (1 - V^0(t)) \quad (23)$$

with the terminal condition given by definition,

$$V^0(m-) = 1 \quad (24)$$

and the initial condition as a result of inserting the equivalence premium calculated above,

$$V^0(0) = 0 \quad (25)$$

Can the Financial Market Relax the Three Assumptions?

In this subsection, we discuss how mathematical finance can help us relax parts of Condition 1 and the valuation principles and formulas that follow from such a relaxation.

The “Large Number” Assumption

This is inevitable for the law of large numbers to apply and we mention just, in passing, some of its counterparts in finance. In capital asset pricing model (CAPM; *see* **Capital Asset Pricing Model**) the hypothesis is that only systematic risk (aka systematic risk, market risk (*see* **Market Risk**), and nondiversifiable risk and is the risk that hits large numbers of assets in the market or the whole economy) has a price, in terms of expected return, while there is no price for taking unsystematic risk (aka idiosyncratic risk and diversifiable risk and is the risk that hits only one or a small number of assets in the market). The motivation is that unsystematic risk can be avoided without loss of expected return by choosing a sufficiently diversified portfolio. In arbitrage pricing theory (APT; *see* **Arbitrage Pricing Theory**), the CAPM is connected to arbitrage pricing by introducing the notion of asymptotic arbitrage. The conclusion is that to avoid asymptotic arbitrage (excess return for taking risk that can be cleared away

by diversification), the price of unsystematic risk cannot be systematically positive, but one cannot say much about the price of unsystematic risk in a particular asset.

In insurance terms, the explanation is as follows: to avoid systematic gains (and losses), the price for taking unsystematic risk must be zero, leading to the equivalence principle under the objective measure. An asymptotic arbitrage argument would give access to a positive price for taking on insurance risk from particular policy holders as long as this does not happen systematically across the policy holders in a large portfolio. However, competition among insurance companies and/or regulation will typically prevent that particular policy holders pay premiums that are larger than their expected benefits.

The Independence Assumption

This can be relaxed to some extent with some help from financial market considerations and actions. This is convenient since the independence assumption is difficult to justify in practice. The help from financial market considerations and actions depends on the source of dependence. Here, we distinguish between two different sources both coming from insufficiency in the model presented in the section Modeling of Payment Streams: first, in practice, the payment coefficients $b^j(t)$, $b^{jk}(t)$, and $\Delta B^j(t)$ are typically not only dependent on the policy holder state but also on the financial performance of some investment fund or index. Second, in practice, the Markov assumption is violated by uncertainties in demographic trends and movements coming from sociological, medicinal, and economic developments.

First, the payment coefficients are supposed to protect buying power upon insurance events rather than nominal values. Therefore, the contracts typically contain participation features in the sense that the policy holder participates in the capital gains of some investment fund. The participation is either directly spelled out in the contract (unit-linked life insurance) or indirectly enforced by the legislative environment in which the contract is underwritten (with-profit life insurance). Additionally, the payments may be linked to the policy holder's salary or some salary index. We take a closer look at these participation features in the section Design of Payment Streams.

Such participation features affect the valuation since they create dependence between payment processes. To the extent that the dependence is created

by financial market risk, a natural idea is to adopt other valuation techniques available for that part of the risk, for example, no arbitrage pricing. A example is to let the payment coefficients be contingent claims formalized as functions of an asset price. Consider, for example, the endowment insurance studied in Example 1, with the modification that a function f of the asset price at time m , $S(m)$, is paid out as benefit conditional on survival to time m . Formally $\Delta B^0(m) = f(S(m))$, and the (unknown) present value at time 0 of the payment process is

$$\int_{0-}^m e^{-\int_0^t r(s)ds} (-\pi I(t) dt + dN(t)) + e^{-\int_0^m r(\tau)d\tau} I(m) f(S(m)) \quad (26)$$

What is the financial value of this present value and what is the equivalence premium? Questions of this type were first approached with financial methods in [2] and since then by several other financial economists and insurance mathematicians [1]. To formalize the standard answer, we introduce two filtrations (see **Filtrations**) and corresponding measures generated by insurance risk and financial risk, respectively. Formally, $\mathcal{F}^Z(t) = \sigma(Z(s), 0 \leq s \leq t)$ and $\mathcal{F}^S(t) = \sigma(S(s), 0 \leq s \leq t)$ such that $(\Omega^Z, \mathcal{F}^Z, P^Z)$ and $(\Omega^S, \mathcal{F}^S, P^S)$ form two filtered probability spaces from which we can construct the filtered product probability space $(\Omega^Z \times \Omega^S, \mathcal{F}^Z \otimes \mathcal{F}^S, P^Z \otimes P^S)$. The standard answer to the valuation problem above is then to take a product measure combining the objective measure P^Z with respect to insurance risk in Z and some martingale measure (see **Equivalent Martingale Measures**) Q^S with respect to financial risk in S , that is,

$$\begin{aligned} E^{P^Z \otimes Q^S} & \left[\int_{0-}^m e^{-\int_0^t r(s)ds} (-\pi I(t) dt + dN(t)) \right. \\ & \left. + e^{-\int_0^m r(\tau)d\tau} I(m) f(S(m)) \right] \\ &= \int_{0-}^m e^{-\int_0^t (r(s) + \mu(x+s))ds} (-\pi + \mu(x+t)) dt \\ &+ e^{-\int_0^m (r(s) + \mu(x+s))ds} E^{Q^S} [f(S(m))] \quad (27) \end{aligned}$$

If the martingale measure is unique, so is the financial value of the payment process and the equivalence premium is obtained as in Example 1

by equating the financial value to zero and isolating π . It is a delicate issue to argue for equation (27) to be the “correct” value. It is clear that one somehow uses the diversification argument for the insurance risk and the no arbitrage argument for the financial risk. However, to make this precise, it takes arguments from incomplete markets pricing typically based on quadratic hedging or no asymptotic arbitrage approaches. Optimal investment is also an issue since, in general, no investment strategy hedges (see **Hedging**) the payment process for any outcome of the stochastic lifetime. In fact, parts of the development of quadratic hedging approaches (see [17] and references therein) are directly inspired by insurance applications in general and to life insurance payment processes in particular, [10, 11, 16].

The statewise reserves introduced above generalize to reserves for more general future payments. In the case of a Markov model for S , these are characterized by deterministic differential equations that generalize Thiele’s differential equation (16) [18].

Second, the transition intensities in the process Z move not only over age but also over calendar time in the sense that a mortality intensity for an x -year old is not necessarily the same in two different calendar years. If these movements are stochastic, this also creates dependence between policy holders and, thereby, between the payment processes written on their lives, and the Markov property of Z is lost. This is clear since the prospects of an insured’s life not only depends on his/her state of health but also on the historical demographic development that is common for all insured.

There are different approaches to the modeling of, for example, mortality intensities. These are typically either sociobiologically flavored or financially flavored. By sociobiological flavor we mean that the starting point for the modeling is the biological, medicinal, and sociological sources of uncertainties. If, say, the future invention of a particular vaccine is expected to have a significant impact on the mortality intensity, then one should try to predict the finding of such a vaccine. By financial flavor we mean that the main source of inspiration is modeling of the short rate of interest. The benefits of such an inspiration are clear when comparing the survival probability $e^{-\int_0^t \mu(s)ds}$ with a discount factor $e^{-\int_0^t r(s)ds}$. If we assume some kind of affine mortality intensity model, we achieve particularly simple expressions for, for

example, the expected survival probability,

$$E \left[e^{-\int_0^t \mu(s)ds} \right] \quad (28)$$

Such particularly simple expressions are appealing but suffer from mainly two disadvantages: it is difficult to motivate by sociobiological arguments that Vasicek or CIR (see **Cox–Ingersoll–Ross (CIR) Model**) model behavior should be present in mortality rates, and it is difficult to generalize the modeling advantages from the survival model to a multistate model.

The main challenge in stochastic modeling of the mortality intensity is the financial valuation of the demographic risk in a given model. Demographic risk is clearly nondiversifiable since the same mortality intensity hits a whole population or portfolio of insured. Is it priced by the financial market then? From the beginning of the millennium, there have been a few attempts by participants in the financial market to form a market for mortality derivatives. The idea is to involve the rest of the financial market, in addition to insurers and reinsurers, in taking part in mortality risk. The motivation for the investors is diversification since mortality intensities are not perfectly correlated with classical financial market risk. There are a lot of delicate questions concerning institutional and contractual design of products in such a market in addition to the modeling issues mentioned above. These questions are not so far answered sufficiently well by neither the market nor the academicians working in the field and, therefore, such a market is not flying yet. The prices in such a possible market place would give hints concerning the valuation of a survival probability, for example, under which probability measure (martingale measure) the expectation in equation (28) should be taken in order to form a market price. However, so far, the market is not sufficiently rich of assets and liquid to be really informative. See [3] for a review of the challenges of mortality risk.

The Deterministic Interest Rate Assumption

If this is not fulfilled we cannot, in the first place, say anything certain about the value at time t of one unit paid at time $u > t$. However, if there exists a sufficiently rich market of risk-free bonds, then we

can say something: if at time t we have access to a risk-free zero-coupon bond (see **Bond**) maturing at time u , then the value of that bond, denoted by $P(t, u)$, is the only value of the payment of one unit at time u that prevents arbitrage. From bond market theory, it is known that the zero-coupon price can be represented as the conditional expected value of the stochastic discount factor under an equivalent martingale measure Q^r derived from the price processes of the bond market. The question is, how can we use the bond market to update the valuation formula

$$E \left[\int_{0-}^n e^{-\int_0^t r(s) ds} dB(t) \right] \quad (29)$$

Here, the independence between payment processes does not carry over to the independence of present values in the following form:

$$\int_{0-}^n e^{-\int_0^t r(s) ds} dB(t) \quad (30)$$

since it is the same interest rate process that hits all policies. Again, we can take a product measure, now combining the objective measure P^Z with respect to insurance risk in Z and some martingale measure Q^r with respect to financial risk in r , and form the expectation

$$E^{P^Z \otimes Q^r} \left[\int_{0-}^n e^{-\int_0^t r(s) ds} dB(t) \right] \quad (31)$$

If the payments are in nominal terms, that is, independent of market risk including interest rate risk, this formula reduces to

$$E^{P^Z} \left[\int_{0-}^n P(0, t) dB(t) \right] \quad (32)$$

Again, the considerations apply: we cannot motivate equation (32) by either diversification or the no arbitrage principle but must rely on a combination of the two, introducing, for example, quadratic hedging in incomplete markets pricing and/or no asymptotic arbitrage arguments. No bond portfolio hedges the payment processes for all outcomes of Z , so the bond portfolio must reflect some optimization criterion.

What Can Finance Learn from Life Insurance: An Example

After a section about “what finance has done for life insurance”, we partly balance out things by a remark

on “what life insurance has done for finance.” Taking Figure 1 to represent the life history of a corporate rather than that of a human being, we have a simple credit risk model and in equation (2) a formalization of simple single-name credit risk derivatives. The default hazard rate μ is, however, usually then modeled as a stochastic process, depending on some underlying macroeconomic variables or processes creating cross-corporate dependence similar to the dependence between life histories of policy holders. Following this idea and replacing in Figure (3) the husband by Corporate A and his wife by Corporate B, we have a simple two-name model. In the payment process (9), we have formalized simple two-corporate credit risk derivatives. This model can from there be generalized to an arbitrary number of corporates. The cost of an exploding number of states and intensities as the number of corporates increases can be dealt by using appropriate cross-corporate homogeneity assumptions.

Figure (3) illustrates three sources of dependence playing distinct roles in the dependence modeling and in the calculation of expected values. One type of dependence between the two corporates exists, as mentioned in the previous paragraph, even if $\mu^{02} = \mu^{13}$ and $\mu^{01} = \mu^{23}$, but these intensities are functions of the same underlying economic processes. Then the corporates are dependent by being part of the same economy (or in the human life model by the two spouses being part of the same generation). Stronger dependence is present if, in addition, $\mu^{02} \neq \mu^{13}$ and/or $\mu^{01} \neq \mu^{23}$. The so-called positive contagion, for example, $\mu^{13} > \mu^{02}$, appears if, for example, there is some kind of contractual relation between the corporates (or in the human life model if, upon the death of his spouse, the widower tends to die from loneliness and isolation). The so-called negative contagion, for example, $\mu^{23} < \mu^{01}$, appears if, for example, the two corporates are competing and the surviving corporate wins market shares after the default of the other (or in the human life model if, upon the death of her spouse, the widow finds a young lover and lives out her second youth). Even stronger dependence is present if we, in addition, allow for a transition from state 0 to state 3. The so-called joint default appears if the corporates are so closely connected, for example, through some mother–daughter relationship, that one credit risk event draws down both corporates simultaneously (or in the human life model, if both spouses in

the married couple are sitting in the same lethally crashing car.)

Note that here the dependence is introduced in a dynamic way such that one clearly sees what is going on. The valuation of payment processes and fairness principles studied in this section carry over directly to dynamic valuation of involved multicorporate credit risk derivatives. Furthermore, nondiversifiable risk is probably less of a problem here, since much of the macroeconomic risk driving default intensities are actually priced by the market, for example, interest rate risk. The dynamic atomic modeling has some advantages over the usual copula approach to dependence modeling that aggregates static effects by the cost of opaqueness [7].

Design of Payment Streams

With-profit Life Insurance

On the one hand, the policy holder is interested in protecting buying power upon insurance events rather than nominal values. On the other hand, the insurance company is not necessarily interested in taking on nondiversifiable risk. Thus, the parties have delicate interests in sharing this risk, and this risk sharing takes many different institutional and contractual forms. In the traditional method for risk sharing, the payments finally paid are a consequence of the contract and the institutional management within certain regulatory constraints. The idea is, in the first place, to charge nominal premiums that are sufficiently high to cover nominal benefits under all possible outcomes of the economic–demographic environment. The surplus created by such a prudent payment stream belongs to the policy holder and should be redistributed as dividend payments to policy holders. The investment strategy and the redistribution are chosen at the discretion of the insurance company within a set of regulatory requirements.

The traditional way of forming the prudent nominal payment stream is to put up a set of artificial assumptions on the transition intensities μ^{jk*} and the interest rate r^* , thereby forming the so-called technical basis, representing some kind of worst-case scenario. Now the nominal payment stream B is settled in accordance with the equivalence principle under these artificial assumptions, that is, fulfilling

$$E^* \left[\int_{0-}^n e^{-\int_0^t r^*(\tau) d\tau} dB(t) \right] = 0 \quad (33)$$

where E^* denotes expectation with respect to the measure under which the transition intensities are given by μ^{jk*} . Under the technical basis, the independence assumption is fulfilled by construction, since the payments are nominal and all transitions intensities are deterministic such that Z is Markov under the technical measure. In addition, under the technical basis, the deterministic interest rate assumption is fulfilled by construction. Thus, since Condition 1 is fulfilled, the expectation on the left-hand side of equation (33) expresses a financial value, however, under the technical assumptions.

The surplus is created when capital gains different from r^* are realized and real transitions different from the ones expected through μ^{jk*} are experienced. This surplus is redistributed as dividend payments. We introduce a dividend payment process D with a structure similar to the one presented for B in equation (9). However, here we emphasize that the dividend payment process is not adapted to the filtration $\mathcal{F}^Z$ generated by Z , as is B in this case, but adapted to a larger filtration $\mathcal{F}^Z \vee \mathcal{F}^Y$, where $\mathcal{F}^Y$ is generated by realized capital gains and realized intensities. Now, what can we require from such a dividend payment process for the policy to be fair?

The classical approach is first to note that if the capital gains process and the transition intensities were known up to time n , this would be as working under Condition 1, and we could put up the usual equivalence principle. This observation inspires to an equivalence formula where the equivalence is required given the economic–demographic development, that is,

$$E \left[\int_{0-}^n \frac{1}{G(t)} d(B + D)(t) \middle| \mathcal{F}^Y(n) \right] = 0 \quad (34)$$

where G is the portfolio value corresponding to a self-financing investment strategy chosen by the insurance company. This means that all economic–demographic risk coming from the payment process B is returned to the policy holder through the payment process D . One can view this approach as careful risk management by design of the contract. This approach is discussed in [14].

The modern approach, inspired from mathematical finance, is actually to allow for an exchange of economic–demographic risk as long as this risk is priced in accordance with the market. This gives rise

to a substantial relaxation of equation (34) into

$$E^{P^Z \otimes Q^Y} \left[\int_{0-}^n e^{-\int_0^t r(s)} d(B + D)(t) \right] = 0 \quad (35)$$

This method does not necessarily return all economic–demographic risk to the policy holders but leaves parts of it at the insurer who is, in return, appropriately paid for this risk through the valuation measure Q^Y . There are, however, two main challenges in replacing equation (34) by equation (35): first, there is no (sufficiently rich and liquid) market to make the Q^Y measure unique with respect to demographic risk. Second, in the classical approach (34), all the risk management has been taken care of by the design of the contract. Introducing equation (35), risk management also means optimal decision making during the terms of the contract, for example, through optimal investment in incomplete markets. This approach is taken in many publications in the area, see the textbook [11] and references therein.

The insurance companies decide on the dividend process and the investment strategy to which it is adapted within certain constraints given by the contractual, institutional, and regulatory environment. There are, however, substantial regional and institutional differences in how this decision is constrained and made around the world. In addition, given the payment process D , there are differences in how this payment process materializes into payments experienced by the policy holders. In other words, these dividends are not necessarily paid out in cash. More often they are kept within in the insurance company but assigned to the policy holder who will then, in return, receive the so-called bonus payments in connection with either future premium payments and/or benefit payments in the payment process B . Here, we describe two of the more important arrangements. The first arrangement is important due to its major role in the UK/US tradition in the past, but during these decades it is partly superseded by the second arrangement that has a long tradition in the continental Europe.

In pension funding, the contract is part of an employment contract: the employer pays the premiums and the employee receives the benefits. The benefits are typically not nominal but rather linked to the policy holder's salary or some salary index. In general, prudent technical assumptions give rise to a positive and increasing dividend payment stream

that is experienced by the employer as a current or future discount in the premium level. However, if at all, one is not so concerned about prudence because a negative surplus can be filled by paying out negative dividends in terms of increasing current or future premiums paid by the employer. This arrangement splits the financial–demographic risk between the insurance company (pension fund) and the employer, whereas the employee goes free (apart from salary risk). This class of arrangements is called *defined benefits (DBs)* since the benefits are determined initially (possibly in terms of the salary), whereas the future premiums are unknown to the employer and will go up or down, respectively, when a negative or positive surplus is returned.

In with-profit life insurance, the contract is either a private contract or part of an employment contract. The employee or the employer pays the premium and the employee receives the benefits. Here, in contrast, the premiums are nominal or linked to some salary index. Sufficiently prudent technical assumptions give rise to a positive and increasing dividend payment stream that is experienced by the employee as a current or future bonus benefit level. One is here, in contrast, indeed concerned about prudence because a negative surplus can typically not be returned to the policy holder who is contractually protected from decreasing benefits: these are considered as binding minimum guarantees. The financial–demographic risk is shared between the policy holder who receives unknown benefits and the insurance company who bears the risk that the surplus becomes negative, whereas the employer goes free. This class of arrangements is called *defined contributions (DCs)* since the contributions (premiums) are determined initially (possibly in terms of the salary), whereas the future benefits are unknown to the employee (but often constrained downward by minimum guarantees).

It is clear that understanding the mathematics of finance in general and no arbitrage valuation in particular has played a crucial role in replacing the fairness constraint (34) by the fairness constraint (35). This replacement has given the involved parties, that is, policy holders typically represented by regulators, investors, and insurance companies, access to measuring and appropriately evaluating the exchanged risk. Note, however, that part of this risk is neither diversifiable nor hedgeable and therefore there is still some ambiguity left in the fairness criterion (35),

which is not present in equation (34). Also note that valuation is closely connected to existence of efficient hedging strategies—strategies that may be difficult to follow for leading insurance companies and pension funds due to regulatory restrictions and, more importantly, from a real economic point of view—these funds are often large investors in the sense that their actions in the market have adverse feedback effects in market prices.

Unit-linked Life Insurance

Unit-linked life insurance has partly replaced the traditional contracts described above, in particular in the private market. The idea comes from the US market and is making up an increasing part of the private market also in Europe with considerable regional differences, though. The idea is to formalize the adaptation to the economic–demographic environment explicitly in the contract in such a way that the investment strategy can be chosen at the discretion of the policy holder rather than the insurance company.

In pure unit-link insurance, that is, without any form of guarantee, all risks from the investment are eventually returned to the policy holder in one way or another. Then we can naturally take the counterpart to the equivalence criterion (34) and require equivalence of the payment process B ,

$$E \left[\int_{0-}^n \frac{1}{G(t)} dB(t) \middle| \mathcal{F}^Y(n) \right] = 0 \quad (36)$$

where G is the portfolio value corresponding to a self-financing investment strategy chosen partly by the policy holder. By multiplication of $G(n)$, which is $\mathcal{F}^Y(n)$ -measurable, and splitting up the integral, one sees that this is equivalent to the following settlement of the final lump sum payment:

$$\begin{aligned} E \left[\Delta B^{Z(n)}(n) \middle| \mathcal{F}^Y(n) \right] \\ = -E \left[\int_{0-}^{n-} \frac{G(n)}{G(t)} dB(t) \middle| \mathcal{F}^Y(n) \right] \end{aligned} \quad (37)$$

This relation restricts the terminal lump sum to finally make up for all movements on capital gains and transition intensities in the past. For this relation to hold, it is sufficient to formalize the payment process by

$$dB(t) = G(t) dB^*(t) \quad (38)$$

where $B^*(t)$ is a nominal payment process fulfilling the relation

$$E \left[\int_{0-}^n dB^*(t) \right] = 0 \quad (39)$$

In pure unit-link insurance, the (risk-minimizing) investment strategy implemented by the insurance company coincides essentially with the investment strategy stipulated in the contract.

Such an insurance contract is not provided in practice due to legislative constraints and the policy holder's demand for specification of guaranteed levels of premiums and benefits, typically nominal, that protect the policy holder from large capital losses. Such features introduce a genuine sharing of risk between the policy holder and the insurance company, and the criterion (36) is too tight. We can relax equation (36) by introducing the counterpart to equation (35) and require that the payment process B fulfills

$$E^{P^Z \otimes Q^Y} \left[\int_{0-}^n e^{-\int_0^t r(s)} dB(t) \right] = 0 \quad (40)$$

For example, let us say that the terminal lump sum payment given in equation (37) is floored off at some guarantee level K . This guarantee must be compensated, for example, by specifying that only a proportion a of the sum is paid out if it is larger than K , that is,

$$\begin{aligned} E \left[\Delta B^{Z(n)}(n) \middle| \mathcal{F}^Y(n) \right] \\ = \max \left(-a E \left[\int_{0-}^{n-} \frac{G(n)}{G(t)} dB(t) \middle| \mathcal{F}^Y(n) \right], K \right) \end{aligned} \quad (41)$$

Looking closer at this, the terminal payment is some kind of delicate Asian option on G . Now, financial mathematics helps us in calculating fair pairs of a and K such that equation (40) is fulfilled. Such pairs will depend on the riskiness in the portfolio underlying G . The pair $(1, -\infty)$ is one example, corresponding to equation (37), that does not depend on the investment portfolio, though. In such more advanced types of unit-link insurance, the (risk-minimizing) investment strategy implemented by the insurance company does not essentially coincide with the investment strategy chosen by the policy holder but contains some investment in or hedging of the embedded derivatives.

Management of Payment Streams

In the section Design of Payment Streams, we set the framework for the design of the payment streams. Furthermore, we discussed fairness constraints. Within the fairness constraints, there are a series of decisions to be made by the insurance company and/or the policy holder. In this section, we focus on these decisions. In with-profit life insurance and pension funding, there are two decision processes of paramount interest: the investment decisions underlying G and the dividend payment stream D .

We form the surplus X at time t as the assets at time t minus the liabilities at time t with respect to the contract or portfolio of contracts in focus. The assets are the payments to the insurance company (premiums less benefits) in the past accumulated with capital gains. The liabilities are some measure of the future obligations in the payment process B , here denoted by $V(t)$. Thus,

$$X(t) = - \int_{0-}^t \frac{G(\tau)}{G(s)} d(B + D)(s) - V(t) \quad (42)$$

From our analysis in the section Valuation of Payment Streams, we have several candidates for the liabilities to be plugged in: the classical liability is the technical reserve

$$E^* \left[\int_t^n e^{-\int_t^s r^*(\tau) d\tau} dB(s) \middle| Z(t) \right] \quad (43)$$

which is the present value of future payments under the same artificial assumptions as the ones used in the equivalence principle. This ensures that, by the equivalence principle, the surplus prior to inception is equal to 0, that is, $X(0-) = 0$. A more realistic estimate of the future obligations is a market consistent valuation along the lines of equation (32) of the future payments conditional on all information up to time t :

$$E^{PZ} \left[\int_t^n P(t, s) dB(s) \middle| Z(t) \right] \quad (44)$$

Anyway, the surplus will evolve according to the investment and dividend decisions made in the past. In fact, the surplus can be considered as the policy holder's extra wealth, in excess of V , which is managed by the insurance company. Since the investment strategy can be parameterized as an investment strategy for the surplus itself, this investment–dividend

problem is a delicate version of the classical investment–consumption problem known from finance. It is certainly a nonstandard version due to a series of particularities (related to financial topics mentioned in brackets): (i) the decision maker and the consumer do not coincide since the dividends go to the policy holder while the investment–dividend decisions are made by the insurance company (principal agent problem); (ii) usually, the dividends are not paid out as cash but transferred to a future payment stream (habit formation); (iii) obligations are nonhedgeable due to diversifiable death risk and nondiversifiable mortality risk (nonhedgeable income process); (iv) there are regional solvency constraints on the surplus—varying in strictness and underlying valuation principles (intermediate constraints on wealth); (v) there are regional constraints on investment and surplus redistribution strategies (intermediate constraints on investment and consumption); (vi) usually, the surplus is accounted for on (sub)portfolio level rather than individual level and the portfolio members have possibly inhomogeneous attitudes to risk; and (vii) short-fall risk due to nonhedgeable obligations or aggressive investment decisions is borne by the equity holders who should be appropriately paid for it.

The list of particularities is all parts of the subject matter of ALM, which is essential in modern mathematics of life insurance and its merger with parts of mathematical finance, including incomplete market valuation, liability-driven investments, constrained asset allocation, and strategic asset allocation. These areas of finance have gained momentum from their applications in life insurance and versions of these problems stemming from practical problems in the insurance market continue to challenge asset managers both on the theoretical and the practical side.

With the prevalence of unit-link insurance comes the importance for the policy holders to make good investment decisions and for the insurance companies to advise them in doing so. At first glance, this is a classical personal finance problem of optimal investment of pension savings, with or without uncertain lifetime. However, the appearance of genuine insurance elements in the contract, that is, financial protection against life history events, brings to the surface a more delicate process of decision making. This includes the financial protection bought throughout

life in terms of insurance payments: given the financial situation at time t , what is the best death sum, disability sum/annuity, and unemployment sum/annuity to buy? Optimal decisions are naturally influenced by the market price of risk and risk aversion of the policy holder, as is the investment decision in classical personal finance. However, the market price of risk is not necessarily easy to deduct from a market offering contracts with nonexplicit participation features and constraints in a heavily regulated environment. And should risk aversion with respect to life history risk be parameterized in the same way as risk aversion with respect to financial risk?

Clearly, the formulation of the optimization problem is much more involved in the area of personal finance and insurance than in the area of pure personal finance. Ambiguity in the way the optimization problem should be formulated at all and the opaque insurance market in which the decisions are to be taken contribute to a complication of the subject matter. However, it also turns out facilitating in some respects to introduce access to the insurance market: it is by now well-accepted that human wealth, that is, the financial value of future labor earnings, is an important asset in personal decision making. This explains partly the rationale of life cycle investment saying that younger people should hold a larger proportion of their financial assets in risky investments than should older people. However, uncertainties in future earnings in general and the risk of losing the human wealth in an insurance event—for example, disability or unemployment—make the human wealth difficult to assess. The introduction of an insurance market and insurance decisions has a completing effect on the market that makes the assessment of human wealth less ambiguous.

In the literature, the approach to personal finance and insurance go in two distinct directions: one direction is inspired from ruin considerations and formulates the optimization problem in terms of the probability of life time ruin in the sense of ever living below some defined level of poverty [9]. Another direction is inspired from personal finance and formulates the optimization problem in terms of expected utility of future consumption across life history events and uncertain lifetimes. This direction was initiated in [15] and recent contributions are [8, 12].

References

- [1] Aase, K.K. & Persson, S.-A. (1994). Pricing of unit-linked life insurance policies, *Scandinavian Actuarial Journal* **1994**, 26–52.
- [2] Brennan, M.J. & Schwartz, E.S. (1976). The pricing of equity-linked life insurance policies with an asset value guarantee, *Journal of Financial Economics* **3**, 195–213.
- [3] Cairns, A.J.G., Blake, D. & Dowd, K. (2008). Modelling and management of mortality risk: a review, *Scandinavian Actuarial Journal* **2008**(2–3), 79–113.
- [4] Gerber, H.U. (1995). *Life Insurance Mathematics*, 2nd edition, Springer-Verlag.
- [5] Hald, A. (1981). T. N. Thiele's contributions to statistics, *International Statistic Review* **49**, 1–20.
- [6] Hoem, J.M. (1969). Markov chain models in life insurance, *Blätter der Deutschen Gesellschaft für Versicherungsmathematik* **9**, 91–107.
- [7] Kraft, H. & Steffensen, M. (2007). Bankruptcy, counterparty risk, and contagion, *Review of Finance* **11**, 209–252.
- [8] Kraft, H. & Steffensen, M. (2008). Optimal consumption and insurance: a continuous-time Markov chain approach, to appear in *ASTIN Bulletin* **28**(1), 231–257.
- [9] Milevsky, M., Moore, K. & Young, V. (2006). Asset allocation and annuity-purchase strategies to minimize the probability of ruin, *Mathematical Finance* **16**(4), 647–671.
- [10] Møller, T. (2001). Risk-minimizing hedging strategies for insurance payment processes, *Finance and Stochastics* **5**(4), 419–446.
- [11] Møller, T. & Steffensen, M. (2007). *Market-Valuation Methods in Life and Pension Insurance*, Cambridge University Press.
- [12] Nielsen, P.H. & Steffensen, M. (2008). Optimal investment and life insurance strategies under minimum and maximum constraints, to appear in *Insurance: Mathematics and Economics*.
- [13] Norberg, R. (1991). Reserves in life and pension insurance, *Scandinavian Actuarial Journal* 3–24.
- [14] Norberg, R. (1999). A theory of bonus in life insurance, *Finance and Stochastics* **3**(4), 373–390.
- [15] Richard, S.F. (1975). Optimal consumption, portfolio and life insurance rules for an uncertain lived individual in a continuous time model, *Journal of Financial Economics* **2**, 187–203.
- [16] Schweizer, M. (2001). From actuarial to financial valuation principles, *Insurance: Mathematics and Economics* **28**, 31–47.
- [17] Schweizer, M. (2001). A guided tour through quadratic hedging approaches, in *Option Pricing, Interest Rates and Risk Management*. E. Jouini, J. Cvitanic & M. Musiela, eds, Cambridge University Press.

- [18] Steffensen, M. (2000). A no arbitrage approach to Thiele's differential equation, *Insurance: Mathematics and Economics* **27**, 201–214.
- [19] Thiele, T.N. (1880). *Sur la Compensation de Quelques Erreurs Quasisystematiques par la Methodes de Moindres Carres*, Reitzel, Copenhagen.

Related Articles

Credit Risk; Hedging; Market Risk; Risk-neutral Pricing.

MOGENS STEFFENSEN

Stochastic Control in Insurance

Classification of Models

The stochastic control problem in insurance can be classified according to the type of models used to describe the *surplus* or the *risk* process (see **Insurance Risk Models**), the particular controls that are allowed, and the nature of the objective function.

Modeling of the Surplus

In the classical *Cramér–Lundberg* model, the risk process is modeled by a compound Poisson process

$$R(t) = R(0) + pt - \sum_0^{A(t)} U_i \quad (1)$$

where $A(t)$ is the Poisson process of the incoming claims, whose intensity β without loss of generality can be set to 1, the i.i.d. sequence U_i represents the sizes of the successive claims and p is the premium rate with $p = (1 + \eta)EU$, with η being the *relative safety loading*. If one considers a sequence of such models with $R^n(0) = xn^{1/2}$ and $\eta^n = \mu n^{-1/2}$, then $R^n(nt)n^{-1/2}$ converges weakly to a Brownian motion with drift μ and diffusion coefficient $\sigma = [EU^2]^{1/2}$.

Reinsurance

The risk control in either of such models is done by way of *reinsurance* (see **Reinsurance**). This is a contract that the *cedent* (the original insurance company) makes with the reinsurance company so that the reinsurance company pays a part or all of each of the incoming claims in exchange for the portion of the premiums coming to the cedent. The cedent makes a decision on the *retention level* a , and the size of each claim is reduced from U_i to U_i^a , while the premium rate is simultaneously reduced from p to p^a ; the precise meaning of this is different for different types of the reinsurances involved. The following are the most common types of reinsurances considered in the literature:

- Proportional reinsurance. $U^a = aU$, $0 \leq a \leq 1$.
- Excess-of-loss reinsurance. $U^a = U \wedge a$, $0 \leq a < \infty$.

- *XL-reinsurance*. $U^a = U \wedge M + (U - M - L)^+$. Here $a = (M, L)$ is described by two parameters $0 \leq L, M < \infty$, with M being called the retention level and L being called the *limit*.

When the type of the reinsurance is fixed, the premium rate p^a depends on the premium calculation principle:

- Expected value principle. $p^a = p - (1 + \eta_1)E(U - U^a)$, where η_1 is the safety loading of the reinsurer.
 - *Cheap reinsurance*. $\eta_1 = \eta$. In this case, $p^a = (1 + \eta)E(U^a)$.
 - *Noncheap reinsurance*. $\eta_1 > \eta$.
- Variance principle. $p^a = p - [E(U - U^a) + p_1 \text{var}(U - U^a)]$

Given the reinsurance scheme, once the level a is chosen, the dynamics of the surplus process $R^a(t)$ will be governed by equation (1) with U replaced by U^a and p replaced by p^a . If one makes a diffusion approximation then the limiting process will be a linear diffusion with drift μ^a and diffusion coefficient σ^a , while $\sigma^a = E[(U^a)^2]^{1/2}$ is the same as before. Further, μ^a is equal to the difference between the premium rate per unit time and the expected losses per unit time.

- Cheap proportional reinsurance. $\mu^a = a\mu$, $\sigma^a = a\sigma$.
- Noncheap proportional reinsurance. $\mu^a = \mu - (1 - a)\lambda$, $\sigma^a = a\sigma$, $\lambda > \mu$.
- Excess-of-loss reinsurance. $\mu^a = \int_0^a (1 - F(x)) dx$, $\sigma^a = [\int_0^a 2x(1 - F(x)) dx]^{1/2}$.

In stochastic control models, the reinsurance is dynamically controlled, that is, $a = a(t)$, where $a(t)$ is a process adapted to the information filtration. Other controls used in the insurance optimization models are given below.

Dividend Control

In the problems where the objective is to maximize the present value of the dividends, the latter is controlled. The corresponding control is described by an increasing functional $L(t)$ which is the cumulative amount of dividends paid out up to time t . From the definition, it follows that $L(t)$ is an increasing

right-continuous process with $L(0-) = 0$. As with any control functional, it should be adapted to the information filtration, and in addition we require that for any t (deterministic or random) $L(t) - L(t-) \leq R(t)$, where $R(t)$ is the value of the surplus at time t .

Investment Control

In case the models allow the surplus to be invested in the financial market we have a third control $b(t)$ representing the amount invested in the risky asset at time t (for simplicity, we consider here only the Black–Scholes financial market (see **Black–Scholes Formula**) with one risky asset; the generalization to multiple assets and other markets is obvious).

Model Formulation

and the Hamilton–Jacobi–Bellman Equation

Any stochastic control model starts with a probability space $(\Omega, \mathcal{F}, P)$ endowed with a filtration (see **Filtrations**) $\mathcal{F}_t$, $t \geq 0$ and a process $\sum_{i=1}^{A(t)} U_i$ (in the case of Cramér–Lundberg modeling) or a standard Brownian motion $w(t)$ adapted to $\mathcal{F}_t$. A control π consists of a triple $(a(\cdot), b(\cdot), L(\cdot))$ of functionals adapted to the information filtration $\mathcal{F}_t$. If it satisfies the previously mentioned restrictions, then the control is *admissible*. Once the control π is chosen, then in the case of the Cramér–Lundberg model the dynamics of the controlled surplus process becomes

$$\begin{aligned} dR(t) &= p^{a(t)} dt - dN^{a(t)}(t) \\ &\quad + b(t) dS_1(t)/S_1(t) + (R(t) \\ &\quad - b(t)) dS_0(t)/S_0(t) - dL(t) \end{aligned} \quad (2)$$

$$R(0) = x \quad (3)$$

Here, $dN^a(t)(t) = U^{a(t)} 1_{A(t)-A(t-)=1}$. In the case of the diffusion approximation, the surplus is governed by

$$\begin{aligned} dX(t) &= \mu^{a(t)} dt + \sigma^{a(t)} dw(t) \\ &\quad + b(t) dS_1(t)/S_1(t) + (X(t) \\ &\quad - b(t)) dS_0(t)/S_0(t) - dL(t) \end{aligned} \quad (4)$$

$$X(0) = x \quad (5)$$

with $S_0(t)$ and $S_1(t)$ being the price process for risk-free and risky asset, respectively.

Objectives

We describe below the two most popular objectives that are studied in the optimization models in insurance.

- *Ruin probability minimization.* In this case, the third control related to dividends payouts is not present and $\pi = (a(\cdot), b(\cdot))$. The stopping time τ , which is the hitting time of $(-\infty, 0)$ by the surplus process (that is the first time when the surplus process becomes negative) is called the *ruin time* (see **Ruin Theory**). With each control π , a *performance index* $J_x(\pi) = P(\tau < \infty)$ is defined, where x stands for the initial surplus. The objective is to find the *value function* $V(x) = \inf_{\pi} J_x(\pi)$ and the associated *optimal control* π^* , such that $V(x) = J_x(\pi^*)$.
- *Dividend optimization.* The performance index in this case is related to the expected cumulative present value of the total dividend payouts until the time of bankruptcy. If there is no setup cost each time the dividend payments are initiated and the payments can be paid in a continuous manner, then $J_x(\pi) = E \left[\int_0^\tau e^{-ct} dL(t) \right]$, where c is the discount rate. The value function and the optimal control are defined as $V(x) = \sup_{\pi} J_x(\pi)$ and $J_x(\pi^*) = V(x)$.

In the models with a setup cost, $L(t) = \sum_{i=1}^\infty \xi_i 1_{\zeta_i \leq t}$ where ζ_i is an increasing sequence of stopping times, corresponding to the times when the dividends are paid out and ξ_i is a positive $\mathcal{F}_{\zeta_i}$ -measurable random variable representing the amount of dividends paid out at the times ζ_i . In this case, $J_x(\pi) = E \left[\sum_{i=1}^\infty e^{-c\zeta_i} g(\xi_i) \right]$, where $g(y)$ is a function showing the amount of dividends that reaches shareholders, if y of the surplus is used for the dividend distribution, for example, $g(x) = x - K$.

Optimality Equations and the Optimal Control

For any fixed pair a, b , let $R^{a,b}$ or $X^{a,b}$ be the process defined by equation (2) or (4), respectively, with $a(t) \equiv a$, $b(t) \equiv b$ and $L(t) \equiv 0$, and let $\mathcal{L}^{a,b}$ be the *infinitesimal generator* of this process. Then dynamic programming shows that if $V(x)$ is a C^2 function, then in the case of ruin probability minimization it satisfies the Hamilton–Jacobi–Bellman equation:

$$\min_{a,b} \mathcal{L}^{a,b} V(x) = 0 \quad (6)$$

In diffusion models, equation (6) is a fully nonlinear second-order differential equation, while in the Cramér–Lundberg case it is a second order integro-differential equation. One of the boundary conditions for equation (6) is $V(\infty) = 0$, and the second boundary condition is at the point 0. For the diffusion case, it is $V(0) = 1$, while in the Cramér–Lundberg case it is $\gamma V(0) + \delta V'(0) = 0$, with γ and δ determined from the exogenous parameters of the problem. The argmin $(A^*(x), B^*(x))$ of the left-hand side of equation (6) provides the *optimal feedback control functions*, which when substituted into equation (2) or (4) as $(A^*(R(t)), B^*(R(t)))$ or as $(A^*(X(t)), B^*(X(t)))$ in the place of $a(t)$ and $b(t)$ results in the stochastic differential equation whose solution yields the optimal surplus process and the optimal control as a function of time.

For the dividend optimization problem, the Hamilton–Jacobi–Bellman equation for the value function takes on the form

$$\max_{a,b} [\max \mathcal{L}^{a,b} V(x) - cV(x), \mathcal{M}V(x)] = 0 \quad (7)$$

The operator $\mathcal{M}$, in the case of continuous dividend payments, is the first-order differential operator $1 - V'(x)$, while in the case of discrete payments it is of a quasi-variational form $\max_{y < x} [V(x) - V(y) + (x - y) - K]$. The boundary conditions at ∞ are of a polynomial growth type. The boundary condition at 0 for the Cramér–Lundberg model is similar to the ruin minimization case, and for the diffusion model it is $V(0) = 0$. The optimal feedback risk and investment control functions are determined the same way as before, while the smallest point where the solution satisfies $\mathcal{M}V(x) = 0$ determines the level which the surplus under the optimal control should never exceed and reaching which triggers the dividends payments.

Bibliographical Remarks

A discrete time model of ruin probability minimization in which reinsurance and investment are present was considered in [16]. Ruin probability minimization in the Cramér–Lundberg case was studied [9–11]. For diffusion models, this problem was treated in [15, 17, 18, 20].

One of the first dividend maximization problems in the framework of the classical risk process was considered in [4] and [7]. There, the optimality of the

barrier strategy was stated. A dividend distribution model with the reinsurance control was studied in [3].

Dividend optimization in diffusion models with continuous dividend payouts was considered in [1, 2, 12–14]. An analysis of the model in which a setup cost leads to discrete optimal dividend payouts is done in [5].

Problems of optimal control of a pension fund were studied in [6] and [8]. The monograph [19] is a good source of the methods and techniques used in application of stochastic control in insurance. It also provides a comprehensive list of the literature.

References

- [1] Asmussen, S., Højgaard, B. & Taksar, M. (1999). Optimal risk control and dividend distribution policies. Example of Excess-of-Loss reinsurance, *Finance and Stochastics* **4**, 299–324.
- [2] Asmussen, S. & Taksar, M. (1997). Controlled diffusion models for optimal dividend pay-out, *Insurance: Mathematics and Economics* **20**, 1–15.
- [3] Azcue, P. & Muler, N. (2005). Optimal reinsurance and dividend distribution policies in Cramér–Lundberg model, *Mathematical Finance* **15**, 261–308.
- [4] Bühlmann, H. (1970). *Mathematical Methods in Risk Theory*, Springer, New York.
- [5] Cadenillas, A., Choulli, T., Taksar, M. & Zhang, L. (2006). Classical and impulse stochastic control for the optimization of the dividend and risk policies of an insurance firm, *Mathematical Finance* **16**, 181–202.
- [6] Cairns, A. (2000). Some notes on the dynamic and optimal control of stochastic pension fund models in continuous time, *ASTIN Bulletin* **30**, 19–55.
- [7] Gerber, H. (1969). Entscheidungskriterien für den zusammengesetzten poisson-prozess, *Mitteilungen der Vereinigung Schweizerischer Versicherungsmathematiker* **69**, 185–228.
- [8] Gerber, H. & Shiu, E. (2000). Investing for retirement: optimal capital growth and dynamic asset allocation, *North American Actuarial Journal* **4**, 42–62.
- [9] Hipp, C. & Plum, M. (2000). Optimal investment for insurers, *Insurance: Mathematics and Economics* **27**, 215–228.
- [10] Hipp, C. & Plum, M. (2003). Optimal investment for investors with state dependent income, and for insurers, *Finance and Stochastics* **27**, 215–228.
- [11] Hipp, C. & Vogt, M. (2003). Optimal dynamic XL-reinsurance, *ASTIN Bulletin*, **33**, 193–207.
- [12] Højgaard, B. & Taksar, M. (1998). Optimal proportional reinsurance policies for diffusion models, *Scandinavian Actuarial Journal* **2**, 166–168.

- [13] Højgaard, B. & Taksar, M. (1998). Optimal proportional reinsurance policies for diffusion models with transaction costs, *Insurance: Mathematics and Economics* **22**, 41–51.
- [14] Højgaard, B. & Taksar, M. (1999). Controlling risk exposure and dividend pay-out schemes: insurance company example, *Mathematical Finance* **2**, 153–182.
- [15] Luo, S., Taksar, M. & Tsoi, A. (2008). On reinsurance and investment for large insurance portfolios, *Insurance: Mathematics and Economics* **42**, 434–444.
- [16] Schäl, M. (2004). On discrete time dynamic programming in insurance: exponential utility and minimization of the ruin probability, *Scandinavian Actuarial Journal* **3**, 189–210.
- [17] Schmidli, H. (2002). On minimizing the ruin probability by investment and reinsurance, *The Annals of Applied Probability* **12**, 890–907.
- [18] Schmidli, H. (2004). Asymptotics of ruin probabilities for risk processes under optimal reinsurance policies, *QUESTA* **46**, 149–157.
- [19] Schmidli, H. (2008). *Stochastic Control in Insurance*, Springer-Verlag, London.
- [20] Taksar, M. & Markussen, C. (2003). Optimal dynamic reinsurance policies for large insurance portfolios, *Finance and Stochastics* **7**, 97–121.

MICHAEL TAKSAR

Esscher Transform

The *Esscher transform* has a long and remarkable history in actuarial science. Its origin can be traced back to the seminal work of a Swedish actuary, Esscher [7], whose motivation was to develop a flexible and convenient way to approximate the probability distribution of the aggregate claims of an insurance portfolio (see **Accumulated Claims**), an important problem in the classical risk theory. Bühlmann [1, 2] developed the Esscher premium calculation principle and showed that it is a particular case of an economic premium principle. Nowadays, the Esscher transform has applications in option valuation, asset allocation, and risk management, the so-called three pillars in quantitative finance.

Suppose $F(x)$ represents the probability distribution function of a nonnegative random variable X and θ denotes a real constant so that the moment generating function of X exists; that is,

$$M_X(\theta) := E[e^{\theta X}] < \infty \quad (1)$$

The Esscher transform $F(x, \theta)$ is defined as

$$F(x, \theta) := \frac{\int_0^x e^{\theta y} dF(y)}{M_X(\theta)} \quad (2)$$

Here, $F(x, \theta)$ is well defined since $M_X(\theta) < \infty$. $M_X(\theta)$ is a normalization constant to make $F(x, \theta)$ a probability distribution function.

The main idea of the Esscher transform is to define a new probability distribution function for X by adjusting the original distribution $F(x)$ of X using an exponential factor $e^{\theta y}$. In particular, for each given parameter θ , we associate it with a transformed probability distribution function $F(x, \theta)$ of X .

In this article, we focus on the applications of the Esscher transform in quantitative finance. For an excellent survey of the applications of the Esscher transform in actuarial science, see [14]. Here, we discuss the applications of the Esscher transform to option valuation and asset allocation (see **Merton Problem**). For the applications of the Esscher transform to risk management, see [12].

Application to Option Pricing

The pioneering work of Gerber and Shiu [8] used the Esscher transform for option valuation. Let $\{X(t); t \geq 0\}$ denote a stochastic process with stationary and independent increments and $X(0) = 0$. Assume that the force of interest or the risk-free interest rate r is a positive constant. For each $t \geq 0$, the random variable $X(t)$ has an infinitely divisible distribution. Let $S(t)$ denote the price of a non-dividend-paying risky asset at time t . Assume that

$$S(t) = S(0)e^{X(t)}, \quad t \geq 0$$

Here, $X(t)$ represents the continuously compounded return from the non-dividend-paying risky asset over the time horizon $[0, t]$.

Let $F(x, t)$ denote the probability distribution function of $X(t)$. Suppose $M_X(z, t)$ represents the moment generating function of $X(t)$; that is, $M_X(z, t) := E[e^{zX(t)}]$. The Esscher transform of the process $\{X(t); t \geq 0\}$ is also a process with stationary and independent increments. The probability distribution function of $X(t)$ defined by the Esscher transform with parameter θ is

$$F(x, t; \theta) := \frac{\int_0^x e^{\theta y} dF(y, t)}{M_X(\theta, t)} \quad (3)$$

The moment generating function of $X(t)$ evaluated under the probability distribution function $F(x, t; \theta)$ defined by the Esscher transform with parameter θ is

$$M_X(z, t; \theta) = \frac{M_X(\theta + z, t)}{M_X(\theta, t)} \quad (4)$$

By the fundamental theorem of asset pricing (see **Fundamental Theorem of Asset Pricing**) (see [10, 11]), to guarantee the absence of arbitrage (see **Arbitrage Strategy**) options should be priced by taking the expected discounted payoffs with respect to an equivalent measure under which the discounted stock price process $\{e^{-rt} S(t); t \geq 0\}$ is a martingale (see **Equivalent Martingale Measures**). Such a change of measure can be achieved using the Esscher transform in the following way. First, we look for a risk-neutral Esscher parameter θ^* such that

$$1 = e^{-rt} E^*[e^{X(t)}] \quad (5)$$

where $E^*[\cdot]$ represents expectation taken with respect to the probability distribution function $F(x, t; \theta^*)$. Equivalently, θ^* is the unique solution of the following equation (see [8]):

$$r = \ln[M_X(1, 1; \theta^*)] \quad (6)$$

Now, we consider a European-style option (see **Options: Basic Definitions**) with payoff $V(T)$ at maturity T . Then, the price of the option is

$$V(0) = E^*[e^{-rT} V(T)] \quad (7)$$

The Esscher transform provides a flexible and convenient way to price options under different parametric specifications of the return process of the underlying risky asset. It is also a convenient tool for option valuation under an incomplete market, in which there is more than one equivalent martingale measure and not every contingent claim can be hedged (see **Hedging**) perfectly. However, an important question arises. How does one justify an option price obtained from the Esscher transform in an incomplete market? Gerber and Shiu [8] showed that an option price arising from the Esscher transform can be obtained by solving a utility maximization problem of an economic agent with power utility. Bühlmann *et al.* [3, 4] provided justification for the use of the Esscher transform for asset pricing under general dynamics of the underlying risky assets. The novelty of the work of Gerber and Shiu is to highlight the interplay between actuarial and financial pricing, which is coined as an important research agenda in modern actuarial science (see [3]). Since the work of Gerber and Shiu [8], there has been much work on further exploring the use of the Esscher transform for option pricing.

Bühlmann *et al.* [3] introduced the conditional Esscher transform to price options under general semimartingales (see **Martingales**). Bühlmann *et al.* [4] adopted the Esscher transform in a discrete financial model and investigated the no-arbitrage theory of asset pricing. Chan [5] employed the Esscher transform to price options in a continuous-time financial model driven by Lévy processes (see **Exponential Lévy Models**). Yao [15] specified a forward-risk-adjusted measure and developed a consistent framework for pricing options on equities, interest rates and foreign exchange rates. Siu *et al.* [13] introduced the use of the Esscher transform for pricing options under a class of GARCH models with innovations having a general and flexible distribution. Elliott *et al.* [6]

considered the option pricing problem when the price dynamics of the underlying risky asset are governed by a Markov regime-switching model.

Asset Allocation

The Esscher transform can also be used to provide an elegant solution to the asset allocation problem [9]. Here, we outline the key results from Gerber and Shiu [9] under a complete securities market with a single primitive risky asset. Let $\{X(t); t \geq 0\}$ denote a Wiener process with “real-world” expected rate of return μ and diffusion coefficient σ ; that is, for each $t \geq 0$, $X(t)$ follows a normal distribution with mean μt and variance $\sigma^2 t$. Define a price kernel of the securities market by the Esscher transform as below:

$$\Lambda := \frac{e^{\theta^* X(T)}}{E[e^{\theta^* X(T)}]} \quad (8)$$

where, in this case, the risk-neutral Esscher parameter θ^* is given by

$$\theta^* = \frac{r - \sigma^2/2 - \mu}{\sigma^2} \quad (9)$$

Let $\beta := \theta^* \mu + (\theta^* \sigma)^2/2$. Then,

$$\Lambda = e^{\theta^* X(T) - \beta T} \quad (10)$$

Suppose Y denotes a random payment in the securities market at time T . Then, the price of the random payment at time 0 is $E[\Lambda Y]$. Consider an investor with initial wealth w and decision horizon T who buys the random payment Y . Then, the terminal wealth of the investor at time T is

$$W(T) = we^{rT} + Y - E[\Lambda Y] \quad (11)$$

This implies that

$$E[\Lambda W(T)] = we^{rT} \quad (12)$$

Now, we suppose that the investor has a risk-averse utility function (see **Utility Function**) $u(x)$ such that $u'(x) > 0$ and $u''(x) < 0$. The goal of the investor is to determine the terminal wealth $W(T)$, which maximizes the expectation $E[u(W(T))]$ subject to the budget constraint imposed by equation (12).

Gerber and Shiu showed that the necessary and sufficient condition for optimality is

$$u'(W(T)) = E[u'(W(T))]\Lambda \quad (13)$$

The necessary condition can be shown by a standard tool in actuarial science, namely, the variational argument (see [9]). The sufficient condition for optimality can be shown by exploiting the concavity of the utility function.

Let $m := E[u'(W(T))]$. Suppose I is an inverse of the marginal utility function u' . Since $u'' < 0$, I exists. Then, the optimal terminal wealth is

$$W(T) = I(m\Lambda) \quad (14)$$

By equations (12) and (14),

$$E[\Lambda I(m\Lambda)] = we^{rT} \quad (15)$$

Consider an exponential utility function as follows:

$$u(x) = -\frac{e^{-\alpha x}}{\alpha}, \quad x \in (-\infty, \infty) \quad (16)$$

where $\alpha > 0$ and $1/\alpha$ is the constant level of risk tolerance.

In this case, the optimal terminal wealth is given by

$$W(T) = we^{rT} + \frac{\theta^*[(r - \sigma^2/2)T - X(T)]}{\alpha} \quad (17)$$

Under the assumption of a frictionless and complete market, Gerber and Shiu [9] determined an optimal dynamic strategy by replicating a contingent claim related to the optimal terminal wealth $W(T)$. In particular, for the exponential utility function, Gerber and Shiu [9] obtained the following optimal amount invested in the risky asset at time $t \in [0, T]$ using the dynamic delta hedging argument:

$$\rho(t) = \left(\frac{\mu - r + \sigma^2/2}{\alpha\sigma^2} \right) e^{-r(T-t)} \quad (18)$$

References

- [1] Bühlmann, H. (1980). An economic premium principle, *ASTIN Bulletin* **11**, 52–60.
- [2] Bühlmann, H. (1984). The general economic premium, *ASTIN Bulletin* **14**, 13–21.
- [3] Bühlmann, H., Delbaen, F., Embrechts, P. & Shiryaev, A.N. (1996). No-arbitrage, change of measures and conditional Esscher transforms, *CWI Quarterly* **9**, 291–317.
- [4] Bühlmann, H., Delbaen, F., Embrechts, P. & Shiryaev, A.N. (1998). On Esscher transforms in discrete finance models, *ASTIN Bulletin* **28**, 171–186.
- [5] Chan, T. (1999). Pricing contingent claims on stocks driven by Lévy processes, *Annals of Applied Probability* **9**, 503–528.
- [6] Elliott, R.J., Chan, L.L. & Siu, T.K. (2005). Option pricing and Esscher transform under regime switching, *Annals of Finance* **1**(4), 423–432.
- [7] Esscher, F. (1932). On the probability function in the collective theory of risk, *Skandinavisk Aktuarietidskrift* **15**, 175–195.
- [8] Gerber, H.U. & Shiu, E.S.W. (1994). Option pricing by Esscher transforms (with discussions), *Transactions of the Society of Actuaries* **46**, 99–191.
- [9] Gerber, H.U. & Shiu, E.S.W. (2000). Investing for retirement: optimal capital growth and dynamic asset allocation, *North American Actuarial Journal* **4**(2), 42–62.
- [10] Harrison, J.M. & Pliska, S.R. (1981). Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and Their Applications* **11**, 215–280.
- [11] Harrison, J.M. & Pliska, S.R. (1983). A stochastic calculus model of continuous trading: complete markets, *Stochastic Processes and their Applications* **15**, 313–316.
- [12] Siu, T.K., Tong, H. & Yang, H. (2001). Bayesian risk measures for derivatives via random Esscher transform, *North American Actuarial Journal* **5**(3), 78–91.
- [13] Siu, T.K., Tong, H. & Yang, H. (2004). On pricing derivatives under GARCH models: a dynamic Gerber-Shiu approach, *North American Actuarial Journal* **8**(3), 17–31.
- [14] Yang, H. (2004). The Esscher transform, in *Encyclopedia of Actuarial Science*, J. Teugels & B. Sundt, eds, John Wiley & Sons, New York, Vol. 2, pp. 617–621.
- [15] Yao, Y. (2001). State price density, Esscher transforms, and pricing options on stocks, bonds, and foreign exchange rates, *North American Actuarial Journal* **5**, 104–117.

Related Articles

Equivalence of Probability Measures; Equivalent Martingale Measures; Exponential Lévy Models

TAK KUEN SIU

Large Deviations

Modern large deviations theory, pioneered by Donsker and Varadhan [11], concerns the study of rare events, and it has become a common tool for the analysis of stochastic systems in a variety of scientific disciplines and in engineering. The theory developed by Donsker and Varadhan is a generalization of Laplace's principle and Cramér's theorem. Here we concentrate on applications to finance and risk management.

A wealth of monographs discuss the large deviations theory in detail; see, for instance, [8–10]. Here, we shall discuss applications to mathematical finance, including option pricing (*see Option Pricing: General Principles*), risk management (*see Risk Exposures; Risk Management: Historical Perspectives*), and portfolio optimization (*see Risk–Return Analysis; Merton Problem*). The reader may also consult [21], which includes details behind many of the topics touched upon here. Before discussing these applications, however, we provide a brief introduction to some basic concepts underlying the theory of large deviations.

Definition 1 *A sequence of random objects $(Z_n : n \geq 0)$ taking values on some topological space S satisfies a large deviations principle with rate function $(J(z) : z \in S)$, if for each Borel measurable set $A \in \mathcal{S}$*

$$\inf_{z \in A^\circ} J(z) \leq \lim_{n \rightarrow \infty} -\frac{1}{n} \log P(Z_n \in A) \leq \overline{\lim}_{n \rightarrow \infty} -\frac{1}{n} \log P(Z_n \in A) \leq \inf_{z \in \bar{A}} J(z) \quad (1)$$

and $J(\cdot)$ is nonnegative and upper semicontinuous. Here A° and $\bar{A}$ stand for the interior and closure of A , respectively.

Throughout this article, we assume that S is a Polish space (i.e., a separable, completely metrizable topological space). One can show that if a large deviations principle holds, then there exists a deterministic element $\tilde{z} \in S$ such that $Z_n \Rightarrow \tilde{z}$ (where $\Rightarrow$ denotes weak convergence). In many applications, S is a function space, so Z_n is a stochastic process and $\tilde{z}$ is the asymptotically most likely path.

A large deviations principle is intuitively interpreted as having the formal approximation

$$P(Z_n \approx z) \approx \exp(-nJ(z)) \quad (2)$$

for z outside a neighborhood of $\tilde{z}$. The previous approximation does not carry any rigorous meaning. Nevertheless, the formal use of equation (2) often allows to infer large deviations properties of stochastic systems that can be later justified rigorously using large deviations theory. The rigorous meaning behind equation (2) is given by Varadhan's lemma (see [8], p. 137), which states that, in the presence of a large deviations principle, for every continuous function $(F(z) : z \in S)$ bounded from below we have that

$$\log \frac{1}{n} E \exp(-nF(Z_n)) \longrightarrow -\inf_{z \in S} (J(z) + F(z)) \quad (3)$$

as $n \nearrow \infty$. Varadhan's lemma is a generalized version of Laplace's principle. Indeed, note that in view of representation (2), Varadhan's lemma makes perfect intuitive sense after applying Laplace's principle since

$$\begin{aligned} E \exp(-nF(Z_n)) &\approx \int \exp(-n(J(z) + F(z))) dz \\ &\approx \exp\left(-n \inf_{z \in S} (J(z) + F(z))\right) \end{aligned} \quad (4)$$

The solution to the optimization problem $\inf_{z \in S} (J(z) + F(z))$ is the so-called minimum energy path or optimal path. Formally, if we put $F(z) = \infty \times I_{A^c}(z)$, then $E \exp(-nF(Y)) = P(Y \in A)$ and if $\tilde{z} \notin A$ and regularity assumptions on the set A hold, Varadhan's lemma yields

$$P(Z_n \in A) \approx \exp\left(-n \inf_{z \in A} J(z)\right) \quad (5)$$

A related class of asymptotic approximations arises in the setting of dynamical systems with a small random perturbation, for instance,

$$\begin{aligned} dZ_\varepsilon(t) &= \mu(Z_\varepsilon(t)) dt + \varepsilon \sigma(Z_\varepsilon(t)) dB(t); \\ Z_\varepsilon(0) &= z_\varepsilon(0) \end{aligned} \quad (6)$$

assuming the necessary regularity conditions for the existence of a solution to the previous stochastic

differential equations (SDEs) driven by a standard Brownian motion $B(\cdot)$. Note that, formally, if we send $\varepsilon \rightarrow 0$ and assume that $z_\varepsilon(0) \rightarrow \tilde{z}(0)$, we obtain convergence of Z_ε to a deterministic dynamical system satisfying $\tilde{z}'(t) = \mu(\tilde{z}(t))$ given $\tilde{z}(0)$. Therefore, under appropriate conditions, one would expect the existence of a theory backing up approximations such as equation (2) given by an appropriate rate function $I(\cdot)$. Such a theory has indeed been developed, and it is known as the Freidlin–Wentzel theory (see, [8], p. 212 or [22], p. 129). The rest of this article concentrates on the application of the ideas of large deviations in finance.

Option Pricing

Direct Use of Large Deviations Approximations

Large deviations principles are applied in finance to develop approximations for option prices (see **Option Pricing: General Principles**). It is easier to explain the techniques with a simple example. Consider the problem of pricing a digital knock-in call option (see **Barrier Options**) with maturity time T under a *Black–Scholes* economy (see **Black–Scholes Formula**), in particular,

$$\alpha_T = P\left(\min_{0 \leq s \leq T} B(s) \leq -a, B(T) > b\right) \quad (7)$$

for some barrier values $a, b > 0$, where $B(\cdot)$ is a standard Brownian motion. This probability can be, of course, evaluated in closed form ([17, 18], or see equation 12 below), but we illustrate the use of large deviations theory here. We develop approximations that can be asymptotically validated when the time to maturity and the barriers are large. First, we embed the problem in a large deviations setting by introducing an appropriate scaling

$$\alpha_T = P\left(\min_{0 \leq u \leq 1} Z_T(u) \leq -c, Z_T(1) \geq d\right) \quad (8)$$

where $a = -Tc$ and $b = Td$, and $Z_T(u) = B(uT)/T$, $u \in [0, 1]$. The rate function of $Z_T(\cdot)$ is defined on the space $C_1 := C_1[0, 1]$ of absolutely continuous functions and takes the form

$$J(z) = \frac{1}{2} \int_0^1 \dot{z}(u)^2 du \quad (9)$$

Note that $Z_T \rightarrow \tilde{z} = 0$ uniformly in probability. A formal application of equation (2) combined with Laplace’s principle then yields

$$\alpha_T \approx \exp\left(-T \inf_{z \in A} J(z) + o(T)\right) \quad (10)$$

where

$$A = \{z \in C_1 : z(0) = 0, \min_{0 \leq u \leq 1} z(u) \leq -c, z(1) \geq d\} \quad (11)$$

One can then show directly using standard techniques from calculus of variations (applying Euler–Lagrange’s principle) that the optimal path is a piecewise linear function. Note that equation (2) also suggests that the optimal path is the most likely way in which the particular large deviations event, given by A , can occur. In our simple example one can directly evaluate such an optimal path by using the reflection principle. In particular, note that

$$\begin{aligned} \alpha_T &= P(B(T) > (2a + b)T) \\ &= P(Z_T(1) > 2c + d) \end{aligned} \quad (12)$$

which allows to conclude (again by reflection) that

$$z^*(t) = I(t \leq t^*)\theta_- + (\theta_+ - c)I(1 \leq t \leq t^*) \quad (13)$$

where $\theta_- = -(2c + d)$, $\theta_+ = -\theta_-$ and $t^* = c/(2c + d)$. Moreover, we have that $J(z^*) = (2c + d)^2/2$. It is important to note using equation (12) and elementary properties of the Gaussian density that

$$\alpha_T = \frac{\exp(-TJ(z^*))}{(2\pi)^{1/2} T^{1/2}} \quad (14)$$

as $T \nearrow \infty$. In particular, it is worth noting that the large deviations results such as equation (11) provide only rough approximations because no information is given about premultiplying terms like the factor $T^{-1/2}$, which appears in equation (14).

Enhancing Monte Carlo Simulations

As indicated above, large deviations approximations typically provide only logarithmic asymptotics (i.e.,

only the exponential rate of decay is identified without any additional information). Sharp asymptotics (i.e., asymptotics with information about premultiplying factors such as equation 14) can only be developed under an additional problem structure. We continue with our discussion on pricing issues in the context of **barrier options**; the option pricing described in the previous paragraph is just an example of one of many types that are possible. In evaluating barrier options, an important aspect relates to the calculation of exit probabilities. For instance, in [5] such calculations are treated combining both Monte Carlo methods and large deviation principles. The model in use is a geometric Brownian motion under the risk-neutral measure (*see Fundamental Theorem of Asset Pricing; Risk-neutral Pricing*)

$$dS_t = rS_t dt + \sigma S_t dB_t, \quad S_0 = x \quad (15)$$

Using this model, we consider computing the probability that the process S solving equation (15) does not hit any of two barriers, a lower or an upper barrier, which are suitable, positive, twice continuously differentiable functions $(l(t), t \geq 0)$ and $(u(t), t \geq 0)$, respectively.

The approach suggested by Baldi *et al.* [5] consists in using sharp large deviations to reduce bias in **Monte Carlo simulation**. The simulation of S is done by making an equidistant partition of the time interval $[0, T]$, where T indicates the expiration of the contract. Then, the sample path is the collection

$$S_{t_{i+1}} = S_{t_i} \exp \left(\left(r - \frac{\sigma^2}{2} \right) \epsilon + \sigma (B_{t_{i+1}} - B_{t_i}) \right) \\ i = 0, 1, \dots, m \quad (16)$$

where $t_0 = 0 < t_1 < \dots < t_m = T$ and $t_{i+1} - t_i = \epsilon > 0$.

Using sharp asymptotics, it is possible to find approximations of the exit probabilities over small intervals, that is, when ϵ becomes small. Let $\varsigma \in (l(t_i), u(t_i))$ and $y \in (l(t_{i+1}), u(t_{i+1}))$. Then the approximation appears as follows:

$$p_i^\epsilon := \\ P(\exists t \in [t_i, t_{i+1}] : S_t \notin (l(t), u(t)) \mid S_{t_i} = \varsigma, S_{t_{i+1}} = y) \\ = f(t_i, \varsigma, y, \epsilon)(1 + O(\epsilon)) \quad (17)$$

where the function f is indeed known explicitly:

$$f(t, \varsigma, y, \epsilon) = \exp \left\{ -\frac{Q(t, \varsigma, y)}{\epsilon} - R(t, \varsigma, y) \right\} \quad (18)$$

with

$$Q(t, \varsigma, y) = \begin{cases} \frac{2(u(t) - \varsigma)(u(t) - y)}{\sigma^2} & \text{if } \varsigma + y > u(t) + l(t) \\ \frac{2(\varsigma - l(t))(y - l(t))}{\sigma^2} & \text{if } \varsigma + y < u(t) + l(t) \end{cases} \quad (19)$$

and

$$R(t, \varsigma, y) = \begin{cases} \frac{2(u(t) - \varsigma)u'(t)}{\sigma^2} & \text{if } \varsigma + y > u(t) + l(t) \\ \frac{2(\varsigma - l(t))l'(t)}{\sigma^2} & \text{if } \varsigma + y < u(t) + l(t) \end{cases} \quad (20)$$

The final estimator, with reduced bias, for the probability that S does not hit the barriers in the interval $[0, T]$ is constructed by simulating N independent identically distributed (i.i.d.) replications $(S^{(j)} : j \leq N)$ of the process S and obtaining

$$\frac{1}{N} \sum_{j=1}^N \prod_{i=0}^{m-1} I(S_t^{(j)} \in (l(t_i), u(t_i))) p(t_i, S_{t_i}^{(j)}, S_{t_{i+1}}^{(j)}) \quad (21)$$

where

$$p(t_i, S_{t_i}^{(j)}, S_{t_{i+1}}^{(j)}) = 1 - f(t_i, S_{t_i}^{(j)}, S_{t_{i+1}}^{(j)}, \epsilon) \quad (22)$$

Large deviations analysis can also aid the application of variance reduction techniques for Monte Carlo simulation *via* importance sampling, as is discussed next.

Importance Sampling

Once again, we concentrate on a concrete problem to illustrate the use of large deviations in the design of importance sampling (*see Variance Reduction*). Consider the problem of calculating equation (7). The optimal path equation (13) suggests a particular way in which one could bias the occurrence of the large deviations event; here we also consider time to maturity and barrier values to be large and thus we would be dealing with small probabilities (*see*

Rare-event Simulation). In particular, consider a probability measure, say $\mathbb{Q}$, under which a process $(Y(u) : 0 \leq u \leq T)$ follows a Brownian motion with drift θ_- up to time $\tau_{-a} = \inf\{s \geq 0 : Y(s) \leq -Ta\}$, and from time τ_{-a} up to time T (assuming $\tau_{-a} < T$) the drift of $Y(\cdot)$ is switched to the value θ_+ . Using the process Y , we obtain the representation

$$\alpha_T = E_{\mathbb{Q}} \left\{ I \left(\min_{0 \leq s \leq T} Y(s) \leq -aT, Y(T) > bT \right) \times h \right\} \quad (23)$$

where

$$h = e^{-\theta_- a + \theta_-^2 \tau_{-a} / 2 - \theta_+ (T - \tau_{-a}) (Y(T) + a) + \theta_+^2 (T - \tau_{-a}) / 2} \quad (24)$$

The expression above, involving the exponentials, is nothing but the likelihood ratio between the Wiener measure $\mathbb{P}$ (corresponding to the process $(B(s) : 0 \leq s \leq T)$ and the measure $\mathbb{Q}$ on the set $I(\tau_{-a} < T)$. More precisely,

$$\frac{d\mathbb{P}}{d\mathbb{Q}} I(\tau_{-a} < T) = I(\tau_{-a} < T) \times e^{-\theta_- a + \theta_-^2 \tau_{-a} / 2 - \theta_+ (T - \tau_{-a}) (Y(T) + a) + \theta_+^2 (T - \tau_{-a}) / 2} \quad (25)$$

The use of importance sampling proceeds by simulating, say, N i.i.d. replications $(Y_i : i \leq N)$ of the process $(Y(s) : 0 \leq s \leq T)$ and estimating $P(\min_{0 \leq s \leq T} B(s) \leq -a, B(T) > b)$ via

$$\alpha_T^{(N)} = \frac{1}{N} \sum_{i=1}^N I(\tau_{-a}^{(i)} < T, Y_i(T) > bT) \times h^{(i)} \quad (26)$$

where

$$h^{(i)} = e^{-\theta_- a + \theta_-^2 \tau_{-a}^{(i)} / 2 - \theta_+ (T - \tau_{-a}^{(i)}) (Y_i(T) + a) + \theta_+^2 (T - \tau_{-a}^{(i)}) / 2} \quad (27)$$

Here, $\tau_{-a}^{(i)}, i = 1, \dots, N$ are the i.i.d. replications of τ_{-a} obtained from the corresponding i.i.d. processes Y_i . Moreover, large deviations theory allows to conclude that the previous importance sampling estimator satisfies certain optimality properties as $T \nearrow \infty$ (see, for instance, [3], Ch. 7). Additional extensions of this methodology are applicable to

other types of processes. However, to design optimal importance sampling estimators, it is required to compute the associated optimal path under the corresponding rate function of the underlying process. Research in this direction has been developed recently [3, 16].

We now give a broader description of the importance sampling technique in option pricing. When pricing an option, the question is to find the expectation

$$I_g = E(g(S_t : 0 \leq t \leq T)) \quad (28)$$

where g is a payoff function (possibly path dependent) (see **Options: Basic Definitions**), and S is governed by a suitable SDE. Let us concentrate here on the geometric Brownian motion given by equation (15). The naive Monte Carlo estimator based on N replications is

$$\hat{I}_g^N = \frac{1}{N} \sum_{i=1}^N g(S_t^{(i)} : 0 \leq t \leq T) \quad (29)$$

where the $S^{(i)}$'s are i.i.d. replications ([3], a general reference on stochastic simulation) of the process S following equation (15).

The importance sampling procedure turns (28) into

$$I_g = E_{\mathbb{Q}}(g(S_t^{\phi}, 0 \leq t \leq T) \times L_T^{-1}) \quad (30)$$

where

$$\begin{aligned} dS_t^{\phi} &= (r + \sigma \phi_t) S_t^{\phi} dt + \sigma S_t^{\phi} dB_t^{\mathbb{Q}} \\ &= r S_t^{\phi} dt + \sigma S_t^{\phi} dB_t \end{aligned} \quad (31)$$

and the process $(B_t^{\mathbb{Q}} : 0 \leq t \leq T)$ is a standard Brownian motion under $\mathbb{Q}$. The relation between $B_t^{\mathbb{Q}}$ and B_t is given by

$$dB_t = \phi_t dt + dB_t^{\mathbb{Q}} \quad (32)$$

The likelihood ratio $L_T^{-1} = d\mathbb{P}/d\mathbb{Q}$ satisfies

$$\begin{aligned} L_T^{-1} &= \exp \left(- \int_0^T \phi_s dB_s + \frac{1}{2} \int_0^T \phi_s^2 ds \right) \\ &= \exp \left(- \int_0^T \phi_s dB_s^{\mathbb{Q}} - \frac{1}{2} \int_0^T \phi_s^2 ds \right) \end{aligned} \quad (33)$$

Under this specification, the importance sampling estimator takes the form

$$I_{g,IS}^N = \frac{1}{N} \sum_{i=1}^N g(S_{t,i}^\phi, 0 \leq t \leq T) L_T^{-1}(i) \quad (34)$$

where $S_{t,i}^\phi$ and $L_T^{-1}(i)$ are obtained from the corresponding i.i.d. copies of the process S^ϕ , which follows the evolution equations (31). Therefore, the ultimate problem is to find a process ϕ (which might or might not be deterministic) that guaranties an efficient estimation of I_g in terms of variance reduction. As we indicated in the simple barrier example, large deviation techniques provide tools to find such a process ϕ . In the next subsection we discuss another example.

Freidlin–Wentzel Theory

Let us suppose that we are dealing with a European option (see **Options: Basic Definitions**). Thus equation (28) becomes

$$I_g(x) = E(g(S_T) | S_0 = x) \quad (35)$$

and we write $I_g(x)$ to emphasize the dependence on the initial position.

Using **Itô's formula**, under measure $\mathbb{Q}$, the variance of $I_{g,IS}^N$ is given by

$$\begin{aligned} \text{Var}_{\mathbb{Q}}(I_{g,IS}^N) \\ = \frac{1}{N} E_{\mathbb{Q}} \left(\int_0^T L_t (\sigma I'_g(S_t^\phi) + \phi_t I_g(S_t^\phi))^2 dt \right) \end{aligned} \quad (36)$$

where I'_g is the derivative with respect to x . One can readily see that if $\phi_s = -\sigma I'_g(S_s^\phi) / I_g(S_s^\phi)$ the variance of the estimator would vanish. Unfortunately, the function $I_g(\cdot)$ is precisely what we want to find, but the idea is to find an approximation to $I_g(\cdot)$, say $\tilde{I}_g(\cdot)$, and consider $\tilde{\phi}_s = -\sigma \tilde{I}'_g(S_s^\phi) / \tilde{I}_g(S_s^\phi)$ which can be used to generate a change of measure that reduces the variance of the associated importance sampling estimator. For instance, Fourniè *et al.* [13] suggest developing a parametric expansion for $I_g(x)$ as a function of σ as $\sigma \searrow 0$. In turn, such an expansion is based, for options out of the money, on the Freidlin–Wentzel theory.

Varadhan–Laplace Principle

Recall that under the geometric Brownian motion, model (15), the path simulation at discrete-time points satisfies

$$S_{t_{i+1}} = S_{t_i} \exp \left(\left(r - \frac{\sigma^2}{2} \right) \epsilon + \sigma \sqrt{\epsilon} Z_i \right) \quad (37)$$

where $Z_i \sim N(0, 1)$ for $i = 1, \dots, n$. Let $\mathbf{Z} = (Z_1, \dots, Z_n)^\top$ ($\top$ stands for the transpose), then we denote by $G(\mathbf{Z})$ the payoff of the option under a sample $\mathbf{Z}$, for instance, for the Asian option we have $G(\mathbf{Z}) = \max \left(0, \frac{1}{n} \sum_{i=1}^n S_{t_i} - K \right)$. In this case, the partition indicates the times when the process is monitored to take the average.

The procedure developed in [14] is to change the mean of $\mathbf{Z}$ from 0 to a vector $\boldsymbol{\mu}$ to obtain an estimator of I_g . Multiplying each sample by the corresponding importance sampling weight or likelihood ratio the estimator is then

$$I_g^N = \frac{1}{N} \sum_{i=1}^N G(\mathbf{Z}^{(i)}) \exp \left(-\boldsymbol{\mu}^\top \mathbf{Z}^{(i)} + \frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\mu} \right) \quad (38)$$

where $\mathbf{Z}^{(i)}$ is the vector $(Z_1^{(i)}, \dots, Z_n^{(i)})^\top$ of independent random variables such that $Z_j^{(i)} \sim N(\mu_j, 1)$, $j = 1, \dots, n$. The $\mathbf{Z}^{(i)}$'s, $i = 1, \dots, N$ are i.i.d. replications. To choose $\boldsymbol{\mu}$ it suffices to minimize the variance of I_g^N , which turns out to be equivalent to minimizing

$$E \left(G(\mathbf{Z})^2 \exp \left(-\boldsymbol{\mu}^\top \mathbf{Z} + \frac{1}{2} \boldsymbol{\mu}^\top \boldsymbol{\mu} \right) \right) \quad (39)$$

where $\mathbf{Z}$ here is an n -dimensional vector of i.i.d. standard Gaussian random variables. It is not an easy task to solve this optimization problem, but it can be tackled using Varadhan–Laplace asymptotics. The idea is to scale $\mathbf{Z}$ by $\sqrt{\epsilon}$ and asymptotically estimate the expectation in equation (39) as $\epsilon \rightarrow 0$. This approach makes it simpler to find the optimal $\boldsymbol{\mu}$, at least as an approximation, which is asymptotically correct in an environment of small volatility, that is, σ close to 0 in equation (15).

Another interesting application of large deviations principles is in the context of index option pricing. Consider an index H composed of m stocks $(S_1, \dots, S_m)$. The index at time t is computed as $H(t) = \sum_{i=1}^m w_i S_i$, where $w_1, \dots, w_m$ are constants.

Pricing, for instance, a European call on the index involves knowing the so-called local volatility function (see **Local Volatility Model**) of the index. What is proposed in [4] is to approximate such a function using Varadhan's principle to handle it in a simpler way, and therefore, pricing the option.

Risk Management

The use of large deviations theory for computational purposes also arises in the context of risk management. For instance, Dembo *et al.* [7] developed approximations based on large deviations theory for the tail of a loss distribution; a relevant assumption is that the individual losses are conditionally i.i.d., given the state of the economy and their identifying class (these parameters can be given by the specific industry or business line).

Credit Risk

Typically, an important task in credit risk theory (see **Credit Risk**) involves computing the distribution of losses in a portfolio composed of several debt contracts. More precisely, in a portfolio composed of n obligors, the question is to calculate the tail probability of $L_n = c_1 Y_1 + \dots + c_n Y_n$,

$$P(L_n > l) \quad (40)$$

where the Y_i 's are Bernoulli random variables such that $P(Y_i = 1) = p_i = 1 - P(Y_i = 0)$, and indicating that the i obligor (for $i = 1, \dots, n$) may or may not default. The c_i 's represent the loss resulting from the default, and l is a threshold. Generally, the number of obligors is large, and surpassing threshold l may be a rare event, which represents significant losses, in which case one may use large deviations theory to approximate the probability in equation (40). In [15] such a probability is approximated under a multifactor Gaussian copula model (see **Gaussian Copula Model**) and using large deviation theory; here we present some results that can be found in [21].

Suppose that the variables $Y_i, i = 1, \dots, n$ are triggered by other variables $X_i, i = 1, \dots, n$ that might or not be related; this is done in the following way: $Y_i = 1_{\{X_i > x_i\}}, i = 1, \dots, n$ and the vector $(X_1, \dots, X_n)$ is normally distributed. The parameters x_i are such that $P(X_i > x_i) = p_i$ for $i = 1, \dots, n$.

The correlations among variables $\{X_1, \dots, X_n\}$ are determined by the following single-factor model:

$$X_i = \rho Z + \sqrt{1 - \rho^2} Z_i, i = 1, \dots, n \quad (41)$$

where $\rho \in [0, 1)$ and $Z, Z_1, \dots, Z_n$ are independent standard normal random variables (r.v.'s).

When $\rho = 0$, there is no correlation, and using **Cramér's theorem** one can find an asymptotic formula. Suppose that $p_i = p, i = 1, \dots, n$. Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P(L_n > nq) = -q \ln(q/p) \ln \left(\frac{1-q}{1-p} \right) \quad (42)$$

where $l = nq$ and $q \in (p, 1)$.

When $\rho > 0$, that is, there is dependence among obligors, it is also possible to derive an asymptotic result. Indeed, if $q_n = 1 - n^{-a}$ with $a \in (0, 1)$, it is shown in [21] that

$$\lim_{n \rightarrow \infty} \frac{1}{\log n} \log P(L_n > nq_n) = -a \frac{1 - \rho^2}{\rho^2} \quad (43)$$

A more elaborate model of this type is treated in [15].

We have discussed the use of large deviations theory in several computational settings, both in pricing and in risk assessment. Another application of large deviations, also in the context of risk management, arises in the theoretical analysis of risk measures.

Risk Measures

Financial institutions are constantly worried about the quantification of the risk in their portfolios of assets: stocks, bonds, credits, and options (see **Risk Exposures**). A financial asset is generally represented by a random variable, say X , and characterized by the probability measure μ that governs the outcome of X , that is, we write $\mathcal{L}(X) = \mu$ whenever $P(X \in A) = \mu(A)$. The risk associated with X is measured by means of a so-called *risk measure* ρ , which is a real-valued function on the space of random variables that satisfies certain properties (such as monotonicity, subadditivity, positive homogeneity, and translation invariance); see [2].

Assume that it is not possible to deal with the probability measure μ directly, but instead we have independent samples $X_1, \dots, X_n$ of X . Then,

one may want to approximate $\rho(X)$ by calculating $\rho(X^{(n)})$; in this case, $X^{(n)}$ is a random variable and its law $\mathcal{L}(X^{(n)})$ is the empirical measure

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i} \quad (44)$$

Here δ_X stands for the Dirac measure at X .

This approach generates an error, and the question suggested in [25] is to quantify the error by considering the asymptotics, as $n \rightarrow \infty$, of

$$P(|\rho(X) - \rho(X^{(n)})| > \epsilon), \text{ for } \epsilon > 0 \quad (45)$$

Then, for positive ϵ and when $n \rightarrow \infty$, the previous quantity becomes small, and large deviation principles can be used to deal with this probability. An important result used for this is Sanov's theorem [8].

Optimal Investment

We now move to investment. Maximizing revenue in investments under a certain risk profile (which could be induced by a specific set of assets types) has been a common question for financial investors. To tackle the problem, the usual approaches are the mean–variance analysis of Markowitz (*see Risk–Return Analysis*) or the use of utility functions (*see Utility Function; Expected Utility Maximization*). Alternatively, a criterion can be set by considering asymptotic performance of the portfolio over long time horizons [1]. One has to decide how to measure such a performance, and one way to do this is by estimating, given a fixed investment policy, the probability of performing well [6, 12]. The optimization problem can be set as selecting the investment policy that minimizes the probability of ending up below a threshold or the one that maximizes the probability of surpassing it. Since the problem involves calculating an asymptotic probability, it is appealing to use large deviations techniques (one can find this approach in [19, 20, 23, 24]).

As an illustration, we describe below the method described by Pham in [19, 20]. Given an investment policy $\alpha \in \mathcal{A}$ (where $\mathcal{A}$ is the set of admissible investment policies), we consider the rate of return of the associated wealth process (i.e., the logarithm of the value of the portfolio obtained by applying the policy α). We denote such a rate of return process

by $X(\alpha) = (X_t(\alpha), t \geq 0)$. The aim is to maximize over $\mathcal{A}$ the probability $P(X_t(\alpha)/t \geq x)$ in the long term, which is the probability of a rare event as $t \rightarrow \infty$. Here, level $x \in \mathbb{R}$ represents a benchmark that an investor wants to achieve; in [6] it is considered as a stochastic benchmark, such as an index.

It is natural to attempt using large deviations theory to find a (static) policy α^* from

$$v(x, \alpha^*) := \sup_{\alpha \in \mathcal{A}} v(x, \alpha) \quad (46)$$

where

$$v(x, \alpha) = - \lim_{t \rightarrow \infty} \frac{1}{t} \log P(X_t(\alpha)/t > x) \quad (47)$$

Under regularity conditions on the drift and the volatility of the process $X_t(\alpha)$ (for instance, if the drift and diffusion coefficients are bounded and continuously differentiable), one can evaluate $v(x)$ via

$$v(x, \alpha) = \sup_{\theta} (\theta x - \Gamma(\theta, \alpha)) \quad (48)$$

where

$$\Gamma(\theta, \alpha) := \limsup_{t \rightarrow \infty} \frac{1}{t} \ln E(e^{\theta X_t(\alpha)}) \quad (49)$$

This approach is considered in [19]. For instance, if $X_t(\alpha) = \alpha Y(t)$ where $\alpha \in \mathbb{R}$ and $Y(t) = rt + \sigma B(t)$ with $r > 0$ and $B(\cdot)$ is a standard Brownian motion, we have that

$$\Gamma(\theta, \alpha) = \theta \alpha r + \frac{(\theta \alpha \sigma)^2}{2} \quad (50)$$

and then the problem is solved when $\alpha = x/r$ [21]. More sophisticated and also explicit calculations can be found in [19].

References

- [1] Algoet, P.H. & Cover, T.M. (1988). Asymptotic optimality and asymptotic equipartition properties of log-optimum investment, *The Annals of Probability* **16**(2), 876–898.
- [2] Artzner, P., Delbaen, F., Eber, J.M. & Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**(3), 203–228.
- [3] Asmussen, S. & Glynn, P.W. (2007). *Stochastic Simulation: Algorithms and Analysis*, Springer.
- [4] Avellaneda, M., Boyer-Olson, D., Busca, J. & Friz, P. (2003). *Méthodes de grandes déviations et pricing*

- d'options sur indices, *Comptes rendus de l'Academie des sciences. Paris* **336**, 263–266.
- [5] Baldi, P., Caramellino, L. & Iovino, M.G. (1999). Pricing general barrier options: a numerical approach using sharp large deviations, *Mathematical Finance* **9**(4), 293–322.
- [6] Browne, S. (1999). Beating a moving target: Optimal portfolio strategies for outperforming stochastic benchmark, *Stochastic and Finance* **3**, 275–294.
- [7] Dembo, A., Deuschel, J.D. & Duffie, D. (2004). Large portfolio losses, *Stochastic and Finance* **8**, 3–16.
- [8] Dembo, A. & Zeitouni, O. (1999). *Large Deviation Techniques and Applications*, Springer.
- [9] Deuschel, J.-D. & Strook, D.W. (1989). *Large Deviation*, AMS Chelsea Publishing.
- [10] Dupuis, P. & Ellis, R.S. (1997). *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley & Sons, Inc.
- [11] Donsker, M.D. & Varadhan, S.R.S. (1983). Asymptotic evaluation of certain Markov process expectations for large time, *Communications on Pure and Applied Mathematics*; (I (1975) **28**, 1–47); (II (1975) **28**, 279–301); (III (1976) **29**, 389–461); (IV (1983) **36**, 183–212).
- [12] Föllmer, H. & Leukert, P. (1999). Quantile Hedging, *Stochastics and Finance* **3**, 251–273.
- [13] Fourniè, E., Lasry, J.M. & Touzi, N. (1997). Monte Carlo methods for stochastic volatility models, in *Numerical Methods in Finance*, L.C.G. Rogers & D. Talay, eds, Cambridge University Press.
- [14] Glasserman, P., Heidelberg, P. & Shahabuddin, P. (1999). Asymptotically optimal importance sampling and stratification for pricing path-dependent option, *Mathematical Finance* **9**(2), 117–152.
- [15] Glasserman, P., Kang, W. & Shahabuddin, P. (2007). Large deviations in multifactor portfolio credit risk, *Mathematical Finance* **17**(3), 345–379.
- [16] Guasoni, P. & Robertson, S. (2008). Optimal importance sampling with explicit formulas in continuous time, *Finance and Stochastics* **12**, 1–19.
- [17] Karatzas, I. & Shreve, S.E. (1998). *Brownian Motion and Stochastic Calculus*, Springer.
- [18] Klebaner, F. (2005). *Introduction to Stochastic Calculus with Applications*, 2nd Edition, Imperial College Press.
- [19] Pham, H. (2003). A large deviations approach to optimal long term investment, *Finance and Stochastics* **7**, 169–195.
- [20] Pham, H. (2003). A risk-sensitive control dual approach to a large deviations control problem, *Systems and Control Letters* **49**, 295–309.
- [21] Pham, H. (2008). *Some Applications and Methods of Large Deviations in Finance and Insurance*, Lectures from a Bachelier course at Institute Henri Poincaré, Paris.
- [22] Shwartz, A. & Weiss, A. (1995). *Large Deviation for Performance Analysis*, Chapman & Hall.
- [23] Sornette, D. (1998). Large deviations and portfolio optimization, *Physica A* **256**, 251–283.
- [24] Stutzer, M. (2003). Portfolio choice with endogenous utility: a large deviations approach, *Journal of Econometrics* **116**, 365–386.
- [25] Weber, S. (2007). Distribution-invariant risk measures entropy, and larger deviations, *Journal of Applied Probability* **44**, 16–40.

Further Reading

- Bares, P., Cont, R., Gardiol, L., Gibson, R. & Gyger, S. (2000). A large deviation approach to portfolio management, *International Journal of Theoretical and Applied Finance* **3**, 617–639.
- Huh, J. & Kolkiewicz, A. (2006). *Efficient Computation of Multivariate Barrier Crossing Probability and its Applications in Credit Risk Models*, manuscript.

Related Articles

Cramér's Theorem; Esscher Transform; Implied Volatility; Large Strike Asymptotics; Implied Volatility; Long Maturity Behavior; Monte Carlo Simulation; Rare-event Simulation; SABR Model; Saddlepoint Approximation; Variance Reduction.

JOSE H. BLANCHET & CARLOS G.
PACHECO-GONZÁLEZ